Eurasian Journal of Educational Research 74 (2018) 187-206

Eurasian Journal of Educational Research
www.ejer.com.tr

Determining Differential Item Functioning with the Mixture Item Response Theory

Seher YALCIN[1]

ARTICLE INFO

Article History:
Received: 21 Nov. 2017
Received in revised form: 05 Feb. 2018
Accepted: 17 Mar. 2018
DOI: 10.14689/ejer.2018.74.10

Keywords
differential item functioning (DIF), causes of DIF, mixture item response theory, Mantel-Haenszel

ABSTRACT

Purpose: Studies in the literature have generally demonstrated that the causes of differential item functioning (DIF) are complex and not directly related to defined groups. The purpose of this study is to determine the DIF according to the mixture item response theory (MixIRT) model, based on the latent group approach, as well as the Mantel-Haenszel method, based on the observed group approach, compare the results, and determine the possible causes of the DIF. **Research Methods:** As this study is contributing to the production of information to develop the theory, it is considered basic research. In accordance with the purposive sampling method, the research sample consisted of 1166 fourth-grade level students from Singapore, Kuwait, and Turkey who participated in the Trends in International Mathematics and Science Study mathematics application and took the sixth booklet. During the data analysis, the model that adapted the data according to MixIRT was determined. Then, the status of the items displaying DIF was determined according to the adaptive model. **Findings:** According to the MixIRT, the two latent class models fit best to the data. No significant difference by gender was observed in either class or any country. This finding suggests that the gender variable, which is frequently used as the observed group in DIF studies, should not be dealt with alone. **Implications for Research and Practice**: Since it is difficult to state whether an item is advantageous for a subgroup when DIF is determined in accordance with known groups, it is recommended to employ the latent class approach to determine DIF.

[1] Ankara University, Faculty of Educational Sciences, TURKEY, e-mail: yalcins@ankara.edu.tr, ORCID: orcid.org/0000-0003-0177-6727

Introduction

In education and psychology, many measurements are done in order to make various decisions about individuals. The accuracy of the decisions that are constructed based on measurement results is closely related to the validity and reliability of the applications. One of the existing threats to the validity of the decisions can be expressed as item bias (Clauser & Mazor, 1998). Bias is defined as a systematic error in the measurement process (Osterlind & Everson, 2009). The state of items in a test containing systematic error decreases the validity of the measures. To investigate whether the items that compose a test are biased, it is necessary to determine whether differential item functioning is present. Differential item functioning (DIF) is the different probability of individuals in various groups responding correctly to an item after the individuals are matched at the same ability level or according to ability level (Clauser & Mazor, 1998; Embretson & Reise, 2000; Mellenberg, 1989). As the DIF is determined, two groups, which are referred to as the reference and focus groups, are compared. In related literature, the reference group is usually composed of the group considered to be favorable in terms of the features measured by the item, while the focus group represents the group considered to be disadvantageous in terms of the features measured by the item (De Ayala, 2009; Osterlind & Everson, 2009).

Many methods are recommended to find out the DIF, such as the Mantel-Haenszel (MH), SIBTEST, Logistic Regression, etc. (Camilli & Shepard, 1994; Holland & Wainer, 1993; Osterlind & Everson, 2009). The MH method is one of the most frequently used methods in literature. Developed by Mantel and Haenszel (1959), this method was first introduced by Holland and Thayer (1988) to determine DIF. A non-parametric method, MH is based on a comparison of groups matched according to matching criteria, with the help of 2x2 crosstabs that show the numbers of true and false responses separated by the focus and reference group indicators (Holland & Thayer, 1988). The MH methods are similar to other DIF methods and compare the state of functioning of an item between manifest or observed groups. It is assumed that the manifest/observed groups generally represent homogeneous subgroups, such as gender or ethnic groups, and are also associated with the origin of the DIF (Finch & French, 2012; Maij-de Meij, Kelderman, & van der Flier, 2010). However, the known/observed groups cannot always provide the assumption of group homogeneity (De Ayala, Kim, Stapleton, & Dayton, 2002; De Mars & Lau, 2013; Samuelsen, 2008). In addition, recent studies in the field of DIF have shown that the causes of DIF are usually complex and not directly associated with the defined groups (Cohen & Bolt, 2005; De Mars & Lau, 2013). In this context, it is emphasized that the DIF should be examined among latent or unknown groups (Cohen & Bolt, 2005; De Ayala et al., 2002; De Mars & Lau, 2013; Finch & French, 2012; Maij-de Meij et al., 2010; Samuelsen, 2008).

Latent variables are random variables hidden in the measurements that are made. The properties of the latent variables need be indirectly removed by using a statistical model that connects the latent variables to the observed variables (Skrondal & Rabe-Hesketh, 2007). It is seen in the literature that the latent variable models are classified according to the continuous and categorical states of the observed and latent variables.

The traditional latent variable models are presented in Table 1 (Skrondal & Rabe-Hesketh, 2007, p. 714).

Table 1

Traditional Latent Variable Models

Latent variables(s)	Observed variable(s)	
	Continuous	Categorical
Continuous	Common factor model - Structural equation model	Item response theory/ Latent trait model
Categorical	Latent profile model	Latent class model

As can be seen in Table 1, in traditional latent variable models, item response theory models are used when the observed variable is categorical and the latent variable is continuous. The item response theory (IRT) enables the prediction of an individual's abilities and parameters related to the items by associating his or her response to an item with the individual's level of ability and the properties of the item (Embretson & Reise, 2000). In other words, as traits or ability cannot be measured directly, the IRT determines the relationship between an individual's observed test performance and the unobserved traits that are assumed to underlie this performance (Hambleton & Swaminathan, 1985). While there is a continuous latent variable assumption in the IRT, it is assumed that the latent variable is categorical in latent class analysis (LCA) (De Ayala, 2009). As seen in Table 1, LCA is used when the observed variable is categorical and the latent variable is categorical. Latent class analysis is utilized to generate homogeneous subclasses from the heterogeneous latent traits that are sought to be measured. In LCA, it is accepted that all observed variables are the cause of a latent variable that cannot be observed (Vermunt & Magidson, 2002).

The combined use of IRT and LCA results in a powerful statistical method called the Mixture item response theory (MixIRT) (Cohen & Bolt, 2005). The MixIRT models (Kelderman & Macready, 1990; Maij-de Meij et al., 2010) do not have any assumptions about the type or cause of the qualitative differences in the responses of the participants. It only supposes that our sample comes from a community that is consisted of latent subgroups (De Ayala & Santiago, 2017). Latent classes (homogeneous subgroups) are defined in the MixIRT models. Different parameter estimates are calculated between the latent classes in which the same measurement model is present within each latent class. The MixIRT model assumes that a population consists of a limited number of latent individual classes, and that these classes can be differentiated based on item response patterns (von Davier & Rost, 2017). In contrast, these different response patterns are revealed as differences in the parameters of the item response model associated with each group. The formula for the MixIRT model with two parameters is as follows (Finch & French, 2012):

$$P(U = 1|g,\ \theta_{ig}) = \frac{e^{(a_{jg}(\ \theta_{ig} - b_{jg}))}}{1 + e^{(a_{jg}(\ \theta_{ig} - b_{jg}))}}$$

In the formula, "g: 1, 2, …, G" demonstrates latent class membership, " b_{jg}" shows intra-class difficulty for the item j, "a_{jg}" indicates the intra-class discrimination for the item j, and " θ_{ig}" shows the level of latent trait that is measured in the class for the individual referred as i. In the literature, MixIRT is used to find solutions to different research questions at different levels, like determining the DIF at item level (Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005; Samuelsen, 2005) in addition to a bundle level or a scale level (von Davier & Yamamoto, 2004). In this study, MixIRT is used to determine the item level DIF.

MixIRT models do not limit examination to specific variables, since they do not compose DIF analysis according to known variables to determine DIF. For this reason, it is stated that it is more appropriate to determine the cause of the DIF (Maij-de Meij et al., 2010). The determining of the DIF cause also allows the test to avoid the construct validity threat and leads to an increase in the accuracy of the ability parameter estimates (Ong, Williams & Lamprianou, 2011). According to MixIRT, the DIF determination process is generally as follows: The model that is adapted the best is determined with the MixIRT. For this determination, starting from the model with one latent class, the analyses are repeated by increasing the number of latent classes until the model fit statistics give the best value. After the model that adapts to the data the best is identified, the potential presence of DIFs between the determined latent classes is examined.

When examining studies in the field that were conducted to determine the DIF with MixIRT (Cho & Cohen, 2010; Choi, Alexeev & Cohen, 2015; Cohen & Bolt, 2005; Cohen at al., 2005; Finch & Finch, 2013; Kelderman & Macready, 1990; Maij-de Meij et al, 2010; Samuelsen, 2008; Uyar, Kelecioglu, & Dogan, 2017; Yuksel, 2012), it is seen that researchers have generally compared the approaches based on observed groups that are frequently employed in determining the DIF (MH and / or Lord's Chi-square) with the results of DIF based on latent classes (Mixture Rasch, MixIRT, or multilevel MixIRT). Results have shown that the DIF determined according to the latent classes was more effective, and the results based on the real data showed that the latent class and the observed group methods gave similar results (Maij-de Meij et al., 2010). In addition, Cohen and Bolt (2005) determined that known properties that may be associated with DIF, such as gender, are generally poorly associated with latent classes. Such analyses have been usually conducted on the simulated data in the studies (Uyar et al., 2017; Yuksel, 2012). However, there are studies that have been executed with both simulated and real data, as well (Cho & Cohen, 2010; Maij-de Meij et al, 2010). In addition, it has also been shown that MixIRT models both determine the DIF and allow for direct interpretation of the possible causes of the DIF. Although studies that were conducted to determine the DIF according to the MixIRT started to become widespread in the 2000s, it is thought that they are not known in the literature in detail. As for this study, it is aimed to determine the possible causes of the DIF by conducting analyses on only real data. In this context, the purpose of this study is to determine the DIF, compare the results, and determine the possible causes of the DIF according to the MH method based on the observed group approach and the MixIRT model based on the latent group approach. In this context, these are the questions sought to be answered:

1. Which model is adapted the best to the data, according to MixIRT? How is the distribution of characteristics related to gender, country, and item difficulty levels in the latent classes that emerge, according to the model that is adapted to the data?
2. What are the items that show DIF, according to the MixIRT, among the latent classes that emerge, according to the model that is adapted to the data?
3. What are the items indicating DIF among the latent classes, according to the MH method? Are the items that show DIF among the latent classes, according to the MixIRT and MH methods, consistent with each other?

Method

Research Design

This study is basic research because it aimed to determine the DIF, compare the results, and determine the possible causes of the DIF in accordance with the MixIRT and the MH methods, meaning that it will contribute to the production of information for developing the theory.

Research Sample

Purposive sampling method is used in this study. Since the original model (MixIRT) used in the DIF test is based on the item response theory (IRT), it is considered appropriate to use the items of the Trends in International Mathematics and Science Study (TIMSS), which is developed in accordance with IRT models. Items of the TIMSS 2015 fourth-grade mathematics subtest were examined, and analyses were executed only on the sixth booklet, which consists of dichotomous scored items. The reason the dichotomous scored items were considered is that they are appropriate for both the MixIRT and the MH methods. Moreover, since the MixIRT models identify the homogeneous latent classes in data, three countries were included in the study to create a heterogeneous data set. The TIMSS 2015 fourth-grade mathematics achievement averages were taken into consideration in the choosing of the countries. In the TIMSS 2015 fourth-grade mathematics application, the country with the highest achievement score (618) was Singapore, and the country with the lowest achievement score (353) was Kuwait. Turkey remained at the medium level with the average of 483 points (Mullis, Martin, Foy, & Hooper, 2016). The sampling of this study comprised 1166 students from these three countries who participated in the fourth-grade TIMSS mathematics application and took the sixth booklet. Demographic information is presented in Table 2.

Table 2

Demographic Information of Students Composing the Sampling by Country

Countries	Age		Gender (f*)		Total
	Mean	Standard deviation	Girl	Boy	
Kuwait	10.07	5.80	126	113	239
Singapore	10.38	.350	225	240	465
Turkey	9.85	.425	234	228	462
All groups	10.10	2.65	585	581	1166

*f: frequency

As shown in Table 2, 1166 students in total were included in the study. Approximately the same number of students from Singapore and Turkey participated in the application, while fewer students were from Kuwait. The number of students who participated in the TIMSS 2015 application from Kuwait was lower than other countries; therefore, the number of students who took this booklet was also lower (259 students). In addition, 20 students were excluded from the analysis by taking into account the missing data rates of students who participated in the application from Kuwait. Therefore, analyses were conducted on the responses of 239 students. When the average age of students was examined, it is seen that the lowest average age was in Turkey while the highest average age is in Singapore. When standard deviations were examined, a high standard deviation in Kuwait, compared to other countries, draws attention. This indicates that the students who participated in the application from Kuwait are more heterogeneous in age. It is seen that the gender proportions of the students who participated from the three countries is close to each other.

Research Instruments and Procedures

In the TIMSS application, the students' responses are obtained by using 14 different booklets. Within the scope of this study, the items in all booklets are examined; only the sixth booklet was chosen because its items consisted of dichotomous scored items. There is a total of 29 mathematical items in the booklet numbered six. Twelve of these items are from the subject field "Numbers," 11 of them are from "Geometric Shapes and Measures," and six are from "Data Display." When the questions are examined in terms of cognitive level, 15 of them are at knowledge level, eight of them are at applying level, and six are at reasoning level. In terms of item type, 16 of them are multiple choice questions and 13 (1-0 scoring) are open-ended questions.

Before analyzing the data, correlations between the items and the unidimensionality of the data were examined. Four items [M051061Z (item11-i11), M051236 (i13), M041276A (i28), M041276B (i29)] were excluded from the analysis, because of the high correlation between the items. Analyses were conducted on 25 items. Confirmatory factor analysis (CFA) was carried out in the Mplus 8 package program (Muthén & Muthén, 2017) to examine the unidimensional nature of the items in this booklet. As a result of the analysis, when the model fit statistics were evaluated, the items seemed to show a unidimensional construct ($\chi^2_{(275)}$: 757.895, p=0.00; RMSE: 0.039, CFI: 0.966, TLI: 0.963). In addition to, it is seen that the factor loadings of the items range from .403 (i24) to .865 (i4). As a result of the

CFA, when the model fit statistics were evaluated, the items showed a unidimensional construct. In this context, it can be said that the construct validity of the test is high. In addition, Cronbach's alpha reliability coefficients were computed for reliability and found to be .875. This value has shown that the internal consistency of the test is good.

Data Analysis

To analyze the first research question, a model that adapts the data in accordance with MixIRT was determined. The distribution of features such as gender and country, which are known in the emerging classes and are frequently used in the literature, was examined. Average and standard deviation information on item difficulty level were presented. The Bayesian Information Criterion (BIC) value, which is suggested in the literature (Li, Cohen, Kim, & Cho, 2009), was used to determine the appropriate model for parameter estimate based on MixIRT.

To analyze the second research question, since comparisons will be made between the latent groups, whether the same construct existed between the latent classes is tested at first. Following, the items that display DIF in accordance with the MixIRT are identified among the latent classes that emerged in accordance with the fitting model. To determine the appropriate model based on MixIRT and the DIF, the Mplus 8 package program was used (Muthén & Muthén, 2017). Mplus uses the maximum likelihood method in parameter predictions.

To analyze the third research question, the items displaying DIF among the latent classes were determined in accordance with the MH method, which is among the observed group approaches. In addition, the consistency nature of the items displaying DIF among latent classes in accordance with the MixIRT and MH methods is examined. To determine the DIF in accordance with the MH method, the "difR" package in the R software language was used (Magis, Béland, Tuerlinckx, & De Boeck, 2015). In the analyses, the iterative method is used to determine the DIF by the MH method; 1000 iterations were calculated. As a result of the analyses, the iterations with significant MH chi-square values according to the level of significance of .05 are evaluated as items with the DIF. In the MH method, the "deltaMH" value is interpreted to determine the size of the DIF. When this value is "0, " it means the DIF is "A: at a negligible level," when it is " 1.0," "B: at medium level;" when it is "1.5," "C: at large level" (Dorans & Holland, 1993). In the MH analysis, LC-2 was utilized as the focus group, since it mostly consisted of students in Kuwait and Turkey who were considered to be disadvantaged.

Results

Model Data Fit and Distribution of Characteristics Related to the Latent Classes According to MixIRT

The responses of the students to 25 mathematics items were analyzed according to the MixIRT, and the model with two latent classes (BIC: 30709.762) was found to fit the data the best. The model with one latent class (BIC: 30757.065) and the model with three latent classes (BIC: 30742.004) had a higher BIC value. As a result of the classification, the entropy value was found to be 0.815. Clark (2010) stated that an entropy value between .60 and .80

regarding the accuracy of the classification is moderate level and adequate for classification, and above .80 is considered to be high entropy. In this context, it can be interpreted that the classification quality of the latent class membership in this study is good. The distribution of students in latent classes according to the model with two latent classes by country is given in Table 3.

Table 3

Distribution of Students in Latent Classes by Country

	Latent Class (LC) – 1		Latent Class (LC) - 2		Total	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Kuwait	9	2	230	31.8	239	20.5
Singapore	396	89.4	69	9.5	465	39.9
Turkey	38	8.6	424	58.7	462	39.6
Total	443	100	723	100	1166	100

As seen in Table 3, there are a total of 443 students in LC-1. Of the 443 participants, 89.4% (396) were from Singapore, and 85% of the students who participated in the application from Singapore are in this class. In addition, 8.6% (38) of the students were from Turkey, and 2% were from Kuwait. Furthermore, 58.6% (424) of students in LC-2 were from Turkey, 31.8% (230) were from Kuwait, and 96% of students who participated in the application from Kuwait are in this class. The distribution of students in countries that are classified in latent classes by gender is given in Figure 1.

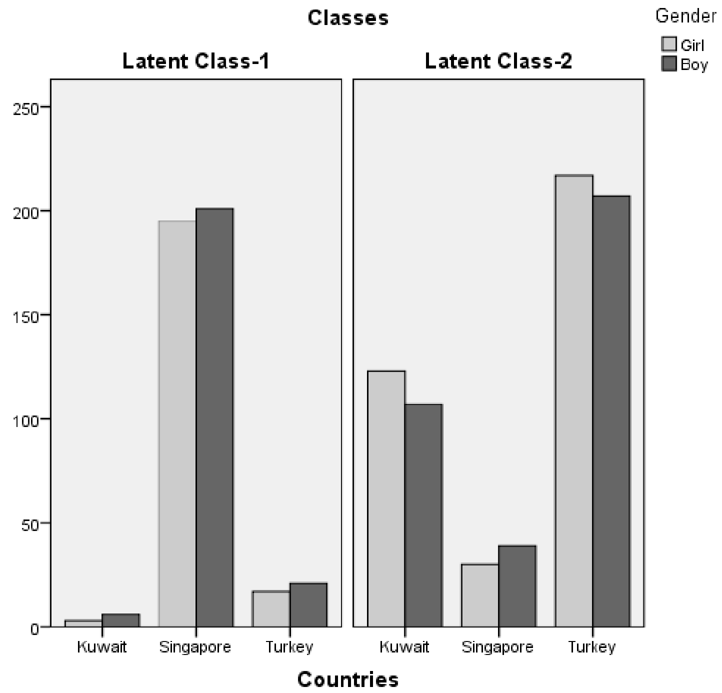

Figure 1. The number of students in countries classified in classes by gender

As can be seen in Figure 1, there is no important distinction in any country or class by gender, and the numbers according to gender are similar. However, the majority of the first latent class consisted of students participating in the application from Singapore, while the second latent class was composed of students from Kuwait and Turkey. The threshold values of the items according to the latent classes are presented in Figure 2.

Figure 2. Threshold values of the items by latent classes

As can be seen in Figure 2, the threshold value of the items was generally higher in LC-1 than LC-2. In this context, it can be interpreted that the individuals in LC-1 achieved higher success than those in LC-2. In addition, when the average difficulty values of the items in the latent classes were evaluated, the average difficulty of the items for LC-1 (mean: -2.67) was lower than LC-2 (mean: 0.88). The standard deviation (sd: 4.33) of the difficulty values of the items in LC-1 is greater than the standard deviation (sd: 1.44) of LC-2. According to these results, it can be interpreted that the items were easy for individuals in LC-1 and are at medium difficulty level for the individuals in LC-2. Only for two items (items 4 and 26) was the threshold value of the items higher in LC-2. Moreover, for items 1 and 21, the threshold values were quite close to each other.

Items Displaying DIF According to MixIRT among Latent Classes

The determination of whether the same construct existed between the latent classes was tested. As a result of the analyses conducted, five items [M051089 (i4), M051125A (i14), M051125B (i15), M041059 (i19) and M041177 (i26)] were excluded from the analyses because they did not measure the same construct on the basis of latent classes. Afterwards, analyses were carried out regarding the nature of the remaining 20 items to express DIF in accordance with the MixIRT among the latent classes. The results are presented in Table 4. Moreover, since the factor variance of item-1 (i1) was set to 1 during the analyses, the results of i1 are not present.

Table 4

DIF Results According to MixIRT

Items	Estimate	Standard error	Estimate/ Standard error
M051017 (i2)	-1.321	0.506	-2.612**
M051111 (i3)	0.261	0.579	0.450
M051094 (i5)	-0.435	0.515	-0.845
M051227 (i6)	-0.924	0.718	-1.287
M051060 (i7)	-0.203	0.542	-0.375
M051061A (i8)	0.334	0.497	0.672
M051061B (i9)	2.591	1.341	1.932
M051061C (i10)	2.877	1.632	1.762
M051129 (i12)	0.009	0.514	0.018
M041298 (i16)	-3.017	1.368	-2.205*
M041007 (i17)	-1.250	0.448	-2.787**
M041280 (i18)	-1.501	0.412	-3.646***
M041046 (i20)	0.056	0.604	0.092
M041048 (i21)	0.221	0.538	0.411
M041169 (i22)	-0.760	0.455	-1.668
M041333 (i23)	-0.506	0.544	-0.929
M041262 (i24)	-0.329	0.335	-0.984
M041267 (i25)	-0.558	0.506	-1.104
M041271 (i27)	-0.416	0.634	-0.655

Note: '***': 0.001, '**': 0.01, '*': 0.05: Indicates the level of significance.

As seen in Table 4, four items (i2, i16, i17, and i18) showed DIF at .05 level. These four items displayed DIF among the latent classes after the students' latent ability was checked. Four of these items are in the subject field of "Numbers." When the questions were examined in terms of cognitive level, all of them were at knowledge level. In terms of item type, all were multiple choice questions. All the DIF displaying items were in favor of LC-1.

Comparing MH Results with MixIRT and the Items Displaying DIF According to the MH Method among the Latent Classes

With the purpose of comparing DIF results, whether DIF exists among latent classes was examined with the MH method based on the observed approach. The DIF results according to latent classes with the MH method are given in Table 5.

Table 5

DIF Results According to Latent Classes with MH Method

Items	Chi-square	alphaMH	deltaMH	Effect size
M051140 (i1)	10.5607**	1.9228	-1.5363	C
M051017 (i2)	1.2162	0.7948	0.5398	A
M051111 (i3)	0.0715	0.9201	0.1956	A
M051094 (i5)	2.3869	1.3946	-0.7816	A
M051227 (i6)	0.1644	0.8992	0.2497	A
M051060 (i7)	3.5403	1.5007	-0.9539	A
M051061A (i8)	38.0975***	0.2447	3.3083	C
M051061B (i9)	0.0044	1.0430	-0.0990	A
M051061C (i10)	1.1145	0.7687	0.6180	A
M051129 (i12)	1.3541	0.7762	0.5954	A
M041298 (i16)	0.0767	0.8482	0.3868	A
M041007 (i17)	68.1786***	0.1573	4.3461	C
M041280 (i18)	19.8569***	0.3944	2.1867	C
M041046 (i20)	7.3242**	1.8559	-1.4532	B
M041048 (i21)	10.5624**	1.9078	-1.5180	C
M041169 (i22)	19.8727***	0.3904	2.2106	C
M041333 (i23)	7.7955**	0.5702	1.3202	B
M041262 (i24)	33.0372***	0.3226	2.6589	C
M041267 (i25)	0.6591	1.1925	-0.4137	A
M041271 (i27)	5.0007*	0.5794	1.2824	B

Note: '***': 0.001, '**': 0.01, '*': 0.05: Indicates the level of significance.

As can be seen in Table 5, 10 items (i1, i8, i17, i18, i20, i21, i22, i23, i24 and i27) showed DIF among the latent classes in accordance with the MH method. Seven of these items displayed DIF at C level, while three displayed DIF at B level. Five of these items were in the subject field of "Numbers," four were "Geometric Shapes and Measures," and one was "Data Display." When the questions were examined in terms of cognitive level, five of them were at knowledge level, three were at applying level, and two were at reasoning level. In terms of item type, nine of them were multiple choice questions, and one (scoring 1 to 0) was an open-ended question. In addition, seven of the 10 items (i8, i17, i18, i22, i23, i24 and i27) were in favor of LC-1, which is the reference group. In this group, two items (i17 and i18) that were in favor of the latent class 1 were consistent with the results obtained based on MixIRT. Three items (i1, i20 and i21), according to the MH method, were in favor of LC-2, which is the focus group.

Discussion, Conclusion and Recommendations

In this study, it is aimed to determine the causes of DIF in addition to DIF according to the MixIRT model based on the latent group approach on real data. It is also aimed to compare the results obtained with the results of the MH method, which determines the DIF based on the observed group approach and is frequently used in literature. In this context, firstly, a model that adapts to the data according to the MixIRT is determined. The distribution of properties such as gender and country, which are commonly known in the emerging latent classes and used frequently in the literature, are examined. Afterwards, the DIF display status of the items is determined according to the fitting model. In addition, items displaying DIF are determined among latent classes according to the MH method and compared with the results of the MixIRT.

According to the MixIRT, the two latent class models fit best to the data. When the individuals in the determined two latent classes were examined separately, there was no remarkable distinction in terms of gender in any country or either class. In their study, which was conducted using the Mixture Rasch model to define biased items in an achievement test, Cohen and Bolt (2005) determined that gender weakly correlates to latent classes, similar to the results of this study. In addition, Tay, Newman, and Vermunt (2011) found that the relationship between latent classes and gender was not significant. This finding, which is consistent with the literature in which the DIF is determined with the latent class approach, suggests that the gender variable, which is frequently used as the observed group in the DIF studies, should not be dealt with alone.

When the latent classes were analyzed by country, the first latent class mostly consisted of students who participated in the application from Singapore, while the second latent class mostly consists of students from Kuwait and Turkey. Cohen and Bolt (2005) also revealed that there was a relationship between ethnic origins and latent classes. Choi et al. (2015) analyzed responses of students from seven countries with different achievement levels to the TIMSS 2007 fourth-grade mathematics sub-test according to the 3PL logistic mixture item response model. As a result of the analysis, the model with two latent classes fit best to the data. Consistent with the findings of this study, it is seen that the first latent class consisted of individuals in countries that demonstrated high performance, such as Hong Kong and Singapore, while the other latent class consisted of individuals with low performances, such as Qatar and El Salvador.

When the student responses to the items were examined according to the two class models, it is seen that the items were quite easy for individuals in LC-1, and the items in LC-2 were at a medium difficulty level. In other words, individuals in LC-1 demonstrated higher achievement than those in LC-2. In their study, Choi et al. (2015) analyzed the data of the mathematics achievement test according to the MixIRT and determined that the model with two latent classes fit best. It is expressed that one of these latent classes consisted of individuals from high-performing countries, while the other latent class consisted of individuals with low performances. These findings are consistent with the findings of the study.

When the DIF was examined according to the MixIRT, four items showed the DIF among the latent classes after the students' latent ability had been checked. All the items showed DIF are in favor of LC-1, which is the group with high achievement. Items identified as DIF among latent classes were examined with regards to the subject area, cognitive level, or item type, and a pattern was revealed. All four items were in the subject field of "number," at the level of "knowing," and in "multiple choice" type. In their study, Cohen and Bolt (2005) found a relationship between subject areas (algebra, geometry, etc.) and latent classes similar to the findings of this study. In the literature, relations between subtopic subject areas and latent classes have been generally found. Finch and Finch (2013) identified three student levels and two school level latent classes with "multidimensional multilevel MixIRT" by considering students' responses to items in mathematics and language tests. The presence of DIF in the items of the latent classes was examined through MH or generalized MH techniques. Three latent classes at the individual level were expressed as follows: those who are successful in both mathematics and language; those who are unsuccessful in both; and those who are successful in mathematics, but unsuccessful in language. Some latent classes have been seen to be more successful according to mathematical subtopics. A similar finding was also found by Cohen et al. (2005).

In this study, the lack of any DIF item in favor of the focus group may be associated with the small number of items that were analyzed for DIF. Moreover, next to the highly successful Singapore, Kuwait's low and Turkey's moderate level of success is thought to be influential for items displaying DIF in favor of disadvantaged groups. In this context, it is suggested for researchers that, while creating a heterogeneous group, countries with moderate to upper and moderate to low levels of successes should be included in the studies conducted in this area, in addition to including countries with very high, very low, and moderate achievements.

As a result of the DIF analysis conducted according to the MH method, it is seen that the 10 items display the DIF among the latent classes. Seven of these items are in favor of the reference group, LC-1. Two items (i12 and i13) that are in favor of LC-1 in this group are consistent with the results obtained based on MixIRT. According to the MH method, three items are in favor of the focus group, LC-2. When the findings are generally evaluated, two out of 10 items that are determined as DIF according to the MH method also displayed DIF according to the MixIRT. In this context, it can be stated that the results of the MH method and the results of the MixIRT are consistent at low level. In their study, Maij-de Meij et al. (2010) used the Mixture Rasch model to determine DIF among latent classes, using Lord's chi-square statistics, which is among the observed group-based DIF determining methods. The results of their study demonstrate that the DIF determined according to the latent classes is more effective, while the results based on the real data showed that the latent class and the observed group methods gave similar results.

When the results of the analysis are broadly evaluated, the DIF determination approach based on the MixIRT is seen to be effective in determining DIF according to latent classes. In this context, it is suggested for all researchers who will conduct DIF examinations to also utilize the latent class approach in their analyses. In addition, it is suggested to use the latent class approach in determining the DIF, since it is difficult to say that an item is advantageous or disadvantageous for all individuals in a subgroup when DIF is determined according to known groups.

The study also has some limitations. One of these is the use of data from only three countries. Interested researchers can also compare the situation in other countries with different levels of achievements. Another limitation is that the MixIRT analyses were conducted with the use of the maximum likelihood method in the Mplus program. Interested researchers can make parameter estimates using the Bayesian approach and/or compare the results of the two methods. Furthermore, only the MH method was used from the methods based on the observed group approach. Interested researchers can compare results using different methods.

References

Camilli, G., & Shepard, L. A. (1994). *MMSS: Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cho, S. J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*(3), 336-370. doi: 10.3102/1076998609353111.

Choi, Y., Alexeev, N., & Cohen, A. S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing*, *15*(3), 239-253. doi: 10.1080/15305058.2015.1007241.

Clark, S. L. (2010). *Mixture modeling with behavioral data*. Unpublished doctoral dissertation. University of California, Los Angeles.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44. doi: 10.1111/j.1745-3992.1998.tb00619.x

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148. doi: 10.1111/j.1745-3984.2005.00007.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice*, *20*(4), 225–233. doi: 10.1111/j.1540-5826.2005.00138.x.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*(3&4), 243-276. doi: 10.1080/15305058.2002.9669495.

De Ayala, R. J. & Santiago, S. Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, 60, 25–40. doi: 10.1016/j.jsp.2016.01.002.

De Mars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: How accurately can we detect who is responding differentially? *Educational and Psychological Measurement*, *71*(4), 597–616. doi: 10.1177/0013164411404221.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel haenszel and standardization. In P. W. Holland, and H. Wainer, (Eds.), *Differential item functioning* (p. 35–66), New Jersey: USA.

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

Finch, W. H. & Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement*, *73*(6) 973–993. doi: 10.1177/0013164413494776.

Finch, W. H. & French, B. F. (2012). Parameter estimation with mixture item response theory models: A monte carlo comparison of maximum likelihood and bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*(1), 167-178. doi: 10.22237/jmasm/1335845580.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and application*. Boston, MA: Kluwer Academic Publishers Group.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the mantel-haenszel procedure. In H. Wainer, and H. I. Brown (Eds.), *Test validity* (p. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, *27*(4), 307–327.

Li, F., Cohen, A. S., Kim, S., & Cho, S. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*(5), 353–373. doi: 10.1177/0146621608326422.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2015). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847-862. doi:10.3758/BRM.42.3.847.

Maij-de Meij, A. M., Kelderman, H. & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, *45*(6), 975-999. doi:10.1080/00273171.2010.533047.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127-143. doi: 10.1016/0883-0355(89)90002-5.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Retrieved from Boston College, TIMSS & PIRLS International Study Center. Retrieved December 12, 2017, from http://timssandpirls.bc.edu/timss2015/international-results/

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (Eighth Edition). Los Angeles, CA: Muthén & Muthén.

Ong, Y. M., Williams, J., & Lamprianou, I. (2013). Exploring differential bundle functioning in mathematics by gender: The effect of hierarchical modelling. *International Journal of Research & Method in Education*, 36(1), 82-100. doi: 10.1080/1743727X.2012.675263.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Los Angeles: Sage.

Samuelsen, K. M. (2008). Examining differential item functioning from a latent mixture perspective. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (p. 177-197). Charlotte, NC: Information Age.

Skrondal, A. & Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, *34*(4), 712–745. doi: 10.1111/j.1467-9469.2007.00573.x.

Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods*, *14*(1), 147-176.

Uyar, S., Kelecioglu, H., & Dogan, N. (2017). Comparing differential item functioning based on manifest groups and latent classes. *Educational Sciences: Theory & Practice*, *17*(6), 1977–2000. doi: 10.12738/estp.2017.6.0526.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars, & A. L. McCutcheon, *Applied latent class analysis* (p. 89-107). New York: Cambridge University Press.

von Davier, M., & Rost, J. (2017). Logistic mixture-distribution response models. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (p. 393-406). Boca Raton: Chapman and Hall/CRC.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement*, *28*(6), 389-406.

Yalcin, S. (2017). The effect of latent classes formed according to the affective characteristics of students on differential item functioning based on sex.

Elementary Education Online, *16*(4), 1917-1931. doi: 10.17051/ilkonline.2017.364505.

Yuksel, S. (2012). *Olceklerde saptanan madde islev farklılıgının karma rasch modelleri ile incelenmesi (Analyzing differential item functioning by mixed rasch models which stated in scales)*. Yayımlanmamıs Doktora tezi. Ankara Universitesi Saglik Bilimleri Enstitusu, Ankara.

Karma Madde Tepki Kuramıyla Farklılaşan Madde Fonksiyonunun Belirlenmesi

Atıf:

Yalcin, S. (2018). Determining of the differential item functioning with the mixture item response theory. *Eurasian Journal of Educational Research*, 74, 187-206, DOI: 10.14689/ejer.2018.74.10

Özet

Problem Durumu: Farklılaşan madde fonksiyonu (FMF), aynı yetenek düzeyinde ya da yetenek düzeyine göre bireyler eşleştirildikten sonra farklı gruplardaki bireylerin bir maddeyi doğru yanıtlama olasılığının farklı olmasıdır. FMF'nin ortaya çıkarılmasında pek çok yöntem [Mantel-Haenszel (MH), Lojistik Regresyon vb.] önerilmektedir. Bu FMF yöntemleri, bir maddenin bilinen veya gözlenen gruplar arasındaki fonksiyonlaşma durumunu kıyaslamaktadır. Gözlenen grupların ise genellikle cinsiyet (kadın ve erkek) ya da etnik gruplar gibi homojen alt grupları temsil ettiği ve FMF'nin kaynağıyla da ilişkili olduğu varsayılmaktadır. Ancak bilinen/gözlenen gruplar, grup homojenliği varsayımını her zaman sağlayamamaktadır. Ayrıca, FMF alanında yapılan son çalışmalar, FMF'nin nedenlerinin genellikle karmaşık olduğunu ve tanımlanmış gruplarla doğrudan ilişkili olmadığını göstermiştir. Bu bağlamda, FMF'nin gizil (bilinmeyen) gruplar arasında incelenmesi gerektiği vurgulanmaktadır.

Araştırmanın Amacı: Bu çalışmanın amacı gizil grup yaklaşımına dayalı Karma Madde Tepki Kuramı (KMTK) modeline ve gözlenen grup yaklaşımına dayalı MH yöntemine göre FMF'nin belirlenmesi, sonuçların karşılaştırılması ve FMF'nin olası nedenlerini belirlemektir.

Araştırmanın Yöntemi: Bu çalışmada, KMTK modeline ve MH yöntemine göre FMF'nin belirlenmesi, sonuçların karşılaştırılması ve FMF'nin olası nedenlerinin belirlenmesi amaçlandığından, yani kuramı geliştirmeye yönelik bilgi üretimine katkıda bulunduğundan temel bir araştırmadır. Bu çalışmada, amaçlı örnekleme yöntemi kullanılmıştır. FMF testinden kullanılan asıl model (KMTK), Madde Tepki Kuramı'na (MTK) dayalı olduğundan MTK modellerine göre geliştirilen Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS) maddelerinin kullanılmasının uygun olduğu

düşünülmüştür. TIMSS 2015 dördüncü sınıf matematik alt testi maddeleri incelenmiş, sadece ikili (1-0) puanlanan maddelerden oluşan altıncı kitapçık üzerinden analizler gerçekleştirilmiştir. İkili puanlanan maddelerin seçilmesinin nedeni, hem KMTK hem de MH yöntemine uygun olmasıdır. Ayrıca KMTK modelleri, verilerdeki homojen gizil sınıfları belirlediğinden heterojen bir veri seti oluşturmak için üç ülke çalışmaya dâhil edilmiştir. Ülkelerin seçiminde TIMSS 2015 dördüncü sınıf matematik başarı ortalamaları dikkate alınmıştır. TIMSS 2015 dördüncü sınıf matematik uygulamasında, en yüksek başarı puanına (618) sahip olan ülke Singapur iken en düşük başarı puanına (353) sahip olan ülke Kuveyt'tir. Türkiye ise 483 ortalama puanıyla orta düzeyde kalmaktadır. Heterojen bir veri seti yaratmak amacıyla bu üç ülkeden dördüncü sınıf düzeyinde TIMSS matematik uygulamasına katılıp altıncı kitapçığı alan 1166 öğrenci bu araştırmanın çalışma grubunu oluşturmuştur. TIMSS uygulamasında, 14 farklı kitapçık kullanılarak öğrencilerin cevapları alınmaktadır. Verilerin analiz edilmeden önce maddeler arası korelasyonlar ve verilerin tek boyutlu olma durumu incelenmiştir. Dört madde, maddeler arası korelasyonu yüksek olduğu için analizden çıkarılmıştır. Analizler 25 madde üzerinden yapılmıştır. Tek boyutluluk analizi sonucu, model uyum istatistikleri değerlendirildiğinde, maddelerin tek boyutlu bir yapı gösterdiği görülmüştür. Veriler analiz edilirken öncelikle KMTK'na göre veriye uyum sağlayan model belirlenmiştir. Oluşan sınıflarda bilinen ve alan yazında sıkça kullanılan cinsiyet, ülke gibi özelliklerin dağılımı incelenmiştir. Ardından gizil sınıflara göre oluşan gruplarda maddelerin tek boyutlu bir yapı gösterme durumu incelenmiş, beş maddenin aynı yapıyı ölçmediği görülerek analizden çıkarılmıştır. Kalan 20 maddenin KMTK'ya göre gizil sınıflar arasında FMF gösterme durumu tespit edilmiştir. Ayrıca, gözlenen grup yaklaşımlarından sıklıkla kullanılan MH yöntemine göre gizil sınıflar arasında FMF gösteren maddeler belirlenmiştir. KMTK'na dayalı uygun modelin ve FMF'nin belirlenmesinde Mplus 8 paket programı kullanılmıştır (Muthén & Muthén, 2017). FMF'nin MH yöntemine göre belirlenmesinde R yazılım dilinde "difR" paketi kullanılmıştır.

Araştırmanın Bulguları, Sonuçları ve Öneriler: KMTK'na göre iki gizil sınıflı model veriye en iyi uyum sağlamıştır. Belirlenen iki gizil sınıftaki bireyler ayrı ayrı incelendiğinde, her iki sınıfta da tüm ülkelerde cinsiyete göre dikkat çeken bir ayrım söz konusu değildir. Cohen ve Bolt (2005), bir başarı testinde yanlı maddeleri tanımlamak için Karma Rasch modelini kullandığı çalışmada, bu çalışmanın sonuçlarına benzer olarak cinsiyetin gizil sınıflarla zayıf bir ilişki içinde olduğunu tespit etmişlerdir. Bu durum, FMF çalışmalarında gözlenen grup olarak sıklıkla kullanılan cinsiyet değişkeninin tek başına ele alınmaması gerektiğini göstermektedir. Gizil sınıflar, ülkelere göre incelendiğinde, ilk gizil sınıfın büyük çoğunluğu Singapur'dan uygulamaya katılan öğrencilerken ikinci gizil sınıf daha çok Kuveyt ve Türkiye'den katılan öğrencilerden oluşmaktadır. Cohen ve Bolt (2005) da yaptıkları çalışmada, etnik köken ile gizil sınıflar arasında ilişkiler olduğunu görmüşlerdir. Öğrencilerin maddelere verdikleri tepkiler incelendiğinde, Gizil sınıf-1'deki bireyler için maddelerin oldukça kolay, gizil sınıf-2 için de maddelerin orta güçlükte olduğu görülmüştür. Bir diğer deyişle, Sınıf-1'deki bireyler, Sınıf-2'dekilerden daha yüksek başarıya sahiptir. KMTK'na göre FMF incelendiğinde, dört madde öğrencilerin gizil yeteneği kontrol edildikten sonra gizil sınıflar arasında FMF göstermektedir. FMF olarak belirlenen maddeler; konu alanı, bilişsel düzey veya madde türü açısından incelenmiş ve bir örüntü olduğu görülmüştür. Cohen ve Bolt (2005) yaptıkları çalışmada, bu çalışmanın bulgularına paralel olarak konu alanları (cebir, geometri vb) ile gizil sınıflar arasında ilişkiler olduğunu tespit etmiştir. MH yöntemine göre yapılan FMF analizi sonucu, 10 madde gizil sınıflar arasında FMF göstermektedir. Bu maddelerden yedisi referans grup olan gizil sınıf-1'in lehinedir. Bu maddelerden ikisi, KMTK'na dayalı çıkan sonuçlar ile tutarlıdır. MH yöntemine göre üç madde ise odak grup olan gizil sınıf-2'nin lehinedir. Maij-de Meij ve diğerleri (2010) çalışmalarında, gözlenen gruba dayalı FMF belirleme yöntemlerinden Lord'un ki-kare istatistiğinden, gizil sınıflar arasında FMF'yi belirlemek için ise Karma Rasch modelini kullanmışlardır. Çalışma sonucunda, gizil sınıflara göre belirlenen FMF'nin daha etkili olduğu, gerçek veriye dayalı sonuçlar ise gizil sınıf ve gözlenen grup yöntemlerinin birbirine yakın sonuç verdiğini göstermiştir. Bu çalışmada ise MH yöntemiyle KMTK sonuçlarının düşük düzeyde tutarlı olduğu ifade edilebilir. Yapılan analiz sonuçları genel olarak değerlendirildiğinde, bilinen gruplara göre FMF tespit edildiğinde o alt gruptaki tüm bireyler için maddenin avantajlı ya da dezavantajlı olduğunu ifade etmek zor olduğundan gizil sınıf yaklaşımının, FMF belirlemede kullanılması önerilmektedir.

Anahtar Sözcükler: farklılaşan madde fonksiyonu (FMF), FMF'nin nedenleri, Karma madde tepki kuramı, Mantel-Haenszel