Eurasian Journal of Educational Research 88 (2020) 179-198

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

Latent Class Approach to Detect Differential Item Functioning: PISA 2015 Science Sample*

Seyma UYAR[1]

ARTICLE INFO

Article History:
Received: 15 Nov. 2019
Received in revised form: 20 May 2020
Accepted: 23 Jun. 2020
DOI: 10.14689/ejer.2020.88.8

Keywords
Latent class, mixture item response theory, differential item functioning, item bias

ABSTRACT

Purpose: This study aimed to compare the performance of latent class differential item functioning (DIF) approach and IRT based DIF methods using manifest grouping. With this study, it was thought to draw attention to carry out latent class DIF studies in Turkey. The purpose of this study was to examine DIF in PISA 2015 science data set. **Research Methods:** Only dichotomous items were considered in this study. Turkey and Singapore samples were used to examine DIF. There were 6115 students in Singapore data set and 5895 students in Turkey sample. To detect DIF among countries based on manifest grouping, Item Response Theory Likelihood Ratio (IRT-LR) and Lord's Chi-Square techniques were used. Besides, with Mixture Item Response Theory latent classes were defined and DIF items were detected with Mantel Haenszel method (MH) among latent classes. Number of DIF items were detected according to latent classes and the two countries were compared.
Findings: There were 8 items including DIF among latent classes. With Lord's Chi square method, four items were detected to include DIF at medium and high level among Turkey and Singapore. And IRT-LR method revealed that only two items included DIF among countries.
Implications for Research and Practice: According to the results, it was recommended to use latent class approach in the investigation of DIF items in cross-country studies.

* This study was presented as an oral presentation in 3rd International Congress on Education Sciences and Learning Technology in Atina on 15-19 November 2017.

[1] Mehmet Akif Ersoy University, TURKEY, e-mail: syuksel@mehmetakif.edu.tr,
ORCID: https://0000-0002-8315-2637

Introduction

A test item should be able to measure ability without involving characteristics of examinees that are in different subgroups. This is because examinees with equal abilities should have the same probability to answer an item correctly, even though they are in different subgroups. When an item has more advantages for one subgroup, then this item is considered biased (Camili & Shepard, 1994; Mellor, 1995; Zumbo, 1999). Biased items cause a systematic error, so they can affect the validity of scores. In addition, biased items prevent the comparability of scores across groups.

International large scale assessments such as Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) are assessments that are applied for different groups varying in culture, country, linguistic, socioeconomic status, school and gender. These demographic factors are irrelevant with test construct and not related to the characteristic measured by the test. But these factors may affect examinees' performance in different subgroups (Oliveri, Ercihan & Zumbo, 2013). PISA and TIMSS are applications, which have multiple language versions. Different language forms of a test may cause to occur biased items in tests, because bias can arise due to test administration, response procedures, or inappropriate translations (Asil & Gelbal, 2012; Hambleton, Merenda & Spielberger, 2007; Van de Vijyer & Tanzen, 2004; Wu & Ercikan, 2006). Results from international assesments may be helpful to policymakers to get educational decisions according to examinees' achievement, but before applications, test developers should examine the items in terms of bias to make the scores comparable across groups.

Item bias determination processes are carried out in two stages. The first stage is a statistical process called differential item functioning (DIF). In this process, item response distributions are examined in reference groups and focal groups, established by considering observed variables (gender, country etc.) under equal ability levels (Cohen & Bolt, 2005; Steinberg & Thissen, 2006). DIF, in the simplest sense, refers to the change of the statistical properties of an item between subgroups when the abilities of these groups are equivalent. (Angoff, 1993; Clauser & Mazor, 1998; Holland & Wainer, 1993). But the presence of DIF in an item is not enough to claim that this item is biased. In the second stage, these items should be examined qualitatively. The DIF items are examined by experts, whether they provide advantages to certain groups (Camilli & Shepard, 1994). DIF can occur in uniform and non-uniform forms (Mellenberg, 1982). In uniform DIF, item discrimination parameters do not vary across groups, but item difficulties vary across the reference and focal groups. An item favors only one group along the ability scale. If non-uniform DIF appeared, it means that this item varies in terms of item difficulties and item discrimination parameters across the reference and focal groups. And this item favours in an ability range one group and in another ability range it favours the other group (De Ayala et al., 2002; Zumbo, 1999).

It can be said that there are many methods to determine DIF items (Camilli & Shepard, 1994; Holland & Wainer, 1993; Millsap & Everson, 1993). To classify these methods, one approach seperates them based on the Classical Test Theory (CTT), such as Mantel Haenszel (Mantel & Haenszel, 1959) and Logistic Regression (Zumbo, 1999),

or methods based on the Item Response Theory (IRT) such as Lord's Chi square, Raju's Area or Item Response Theory Likelihood. However, each method may have disadvantages over the other. Test length, number of DIF items, DIF magnitude, or sample size can affect the performance of DIF methods (Clauser, Mazor & Hambleton, 1993; Gierl, Gotzmann & Boughton, 2004; Kabasakal, Gök, Kelecioglu & Arsan, 2012; Sunbul & Sunbul, 2016). But it is a common view that IRT methods are more effective than CTT based methods, for IRT based methods can estimate ability independently from items (Narayanan & Swaminathan, 1996). Methods based on item response theory (IRT) to detect DIF deal with the differences in the probability to answer the item correctly for two manifest groups. For this reason, IRT methods focus on comparing item characteristic curves (ICCs) (Raju, 1988) or item parameters of the groups (Lord, 1980; Thissen, Steinberg & Wainer, 1993). DIF studies, when used in the manifest grouping, assume that the groups, for example, males and females or ethnic groups, represent homogenous subgroups. Homogeneity means that the items function the same way for these subgroups, which means items do not include DIF within the subgroups (De Ayala, Kim, Stapleton & Dayton, 2002). In addition, these manifest variables are thought to be the source of the DIF. In reality, these manifest groups can be easily identified, but they often do not represent homogeneous populations in terms of the feature that is measured (Samuelsen, 2005). Therefore, it is a fact that an item may contain DIF within the same group. The individuals in a manifest group (e.g. all girls) can be divided into latent classes if all examinees (e.g. all girls) do not have homogeneous response patterns (De Ayala et al.,2002; Ercikan et al. 2013; Samuelsen, 2005). Samuelsen (2005) argued that it is considered a 100% overlap between latent class and manifest group if examinees of a manifest group are also clustered within a single latent class. However, the probability of overlapping manifest group and latent class is poor in real studies. In these cases, it is argued that DIF results obtained from manifest groups may be biased when the ratio of overlap is less than 70%. In this context, it is proposed that DIF studies should be examined among unknown groups/ latent classes (Bilir, 2009; Choi et al., 2015; Cho, 2007; Cohen & Bolt, 2005; De Ayala et al., 2002; De Mars & Lau, 2011; Finch & French, 2013; Karadavut, 2017; Maij-de Meij et al., 2010; Oliveri, Ercikan, & Zumbo, 2013; Samuelsen, 2005, Uyar, Kelecioglu & Dogan, 2017; Yalcin, 2018). Kelderman and McReady (1990) agree with the idea that a latent class approach to detect DIF can be productive. They argue that using latent classes allows DIF to be evaluated independently of any variable or set of variables. These efforts can be helpful for researchers to provide a more precise explanation of the presence or cause of DIF.

A Mixture Item Response Model (MixIRT) can be used to identify the unobservable groups that have similar response patterns and cluster these heterogeneous groups with the help of their response behaviours (Cho & Lee, 2016). MixIRT approach was proposed by Rost (1990) and Mislevy and Verhelst (1990) to have homogeneous subgroups from the tested data. MixIRT is a model that combines the Rasch model and latent class analysis, which allows to estimate item parameters differentially for each latent class. With this separation, examinee's responses in one latent class can be homogeneous, but it is heterogeneous between latent classes. MixIRT models can be

adopted to Rasch models, 2-PL, and 3-PL models (Bolt & Cohen, 2005; Finch & Finch, 2013).

In Mixture Rasch models, the probability of an item to answer it correctly is as follows (Cho, 2007):

$$\left(y_{ijg} = 1 \mid g, \theta_{jg}\right) = \frac{1}{1 + \exp[-(\theta_{jg} - \beta_{ig})]}$$

In this formula $g = 1, \ldots, G$ refers to index with latent class membership; j= $1, \ldots, J$ is responders; θ_{jg}: is examinee's latent ability in latent class g; β is the difficulty parameter of item i in class g. Ability has normal distribution with μ and σ parameters, where these parameters have class-specific features.

MixIRT models are important and valuable. They can establish hypotheses about individual characteristics, which are related with DIF (Sawatzky, Ratner, Kopec & Zumbo, 2012). This is because mixture modeling focuses on maximizing differences among latent classes. This procedure results in an existing large number of DIF items and high DIF effect sizes among latent classes (Samuelsen, 2005). Studies in this field revealed there was a weak correlation between gender and latent classes. This means that DIF analysis conducted with gender groups may produce misleading results (Cohen & Bolt, 2005; Yalcin, 2018). Some members in one group can have the advantage to respond to an item correctly, but other members in this group can be disadvantaged (Cohen & Bolt, 2005; De ayala et all, 2002). Therefore, according to previous studies, it can be said that the performance of manifest DIF analysis may be lower than latent DIF analysis. Studies, related to MixIRT DIF were carried with simulated and real data (Cohen & Bolt, 2005; Maij-de Meij, 2010), with simulated data (Bilir, 2009; Samuelsen, 2005, Uyar et al., 2017; Yuksel, 2012) or only with real data (Finch & Finch, 2013; Karadavut, 2017; Van Nijlen & Janssen, 2008; Yalcin, 2018). According to Cho & Lee (2016), it is required to examine the performance of manifest DIF detection methods for studying latent DIF approach in future studies. Thus, in this study it was aimed to compare the performance of latent class DIF approach and IRT based DIF methods using manifest grouping. With this study, it was thought to draw attention to carry out latent class DIF studies in Turkey. In this context, the following research questions were asked:

1. To which model does the data fit, consisting of Turkey and Singapore samples? And how is the distribution of members from different countries in latent classes?
2. How is the distribution of estimated item difficulties for latent classes?
3. How many items are detected, including DIF according to latent class approach?
4. How many items are detected including DIF among countries with Lord's Chi-square and Item Response Theory Likelihood procedures?

Method

Research Design

In this study, the number of DIF items in PISA science application was investigated with latent class and manifest group approaches. This study is a descriptive study for it tries to reveal the current situation (Büyüköztürk, 2019).

Research Sample

The countries, Turkey and Singapore, were chosen for this study. The countries were selected for some reason. First, this study focused on the countries which are different in terms of their culture and language. According to previous studies, greater DIF items were investigated across different language groups (Ercikan, Oliveri & Zumbo, 2013). On the other hand, Turkey and Singapore have different methods in terms of instruction, curriculum and education policies (Levent & Yazici,2014). These factors may affect the examinees' performance and response styles. Secondly, we focused on the achievement rank of countries. Singapore reached the first rank in science, reading and mathematics literacy, where Turkey's achievement was below the OECD means. To focus on potential DIF sources, it was thought that a comparison of these countries in terms of item response behaviours can provide an opportunity for DIF investigation.

There were 6115 examinees in the Singapore data set and 5895 examinees in the Turkey sample. But once the missing data were removed, and examinees who responded to the common items were selected, 614 examinees in Singapore and 498 examinees in Turkey sample were excluded from the study.

Research Instruments and Procedures

In PISA 2015 application, science was the major domain. For this study, dichotomous items from PISA 2015 science test were used. There were 17 dichotomous items common in Turkey and Singapore samples to test science literacy.

PISA 2015 Science Literacy: PISA is an ongoing program that can help policymakers to take decisions about education. Besides, with PISA applications, it is easy to follow examinees' knowledge and skills across countries although these examinees may be included in other subgroups in each country. PISA is implemented every 3 years. In each cycle, one domain is tested in detail (covering almost half of the test time). In 2006 and 2015 the major domain was science, in 2000 and 2009 reading was the major domain, and in 2003 and 2012 mathematics was the major domain. Since 2012, in each cycle an innovative domain has been tested together with the major domain. In the PISA 2015 assessment, science was the major domain, where collaborative problem solving and financial literacy were innovative domains. (OECD, 2018-PISA 2015 results in focus). Literacy is defined as examinees' adequacy to use their knowledge and skills, logical inferences, and effective communication in terms of interpreting and solving a problem they encountered. According to this, the 'science literacy' terminology expects the student to be a reflective citizen when dealing with science-related issues and ideas. This student can evaluate and design scientific

research, can interpret the data, can give reasoned answers about science and technology problems (OECD, 2016).

In PISA 2015 scientific literacy assessment contexts were health and disease, natural resources, environmental quality, hazards and frontiers of science, and technology. In addition, the questions were associated with personal, local/national, and global problems (OECD, 2018).

Data Analysis

First, data were checked for IRT assumptions. To analyze the dimensionality of the science items, confirmatory factor analysis (CFA) approach was applied in the Mplus 7 software (Muthén & Muthén, 2012). Despite the factor that indicated items were categorical, a robust weighted least square estimation method was preferred (Brown, 2006). To examine the model fit, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) were investigated. In the literature, it is a common opinion that RMSEA should be smaller than 0.08 and CFI and TLI should be greater than .90 for an acceptable fit (Browne & Cudeck, 1993; Hu & Bentler, 1995; Jöreskog & Sörbom, 1993). In this study, it was seen that the model fit to the data, because the fit indices were between the acceptable value ranges. Therefore, it was decided that the unidimensionality of the data was provided (RMSEA =. 02, CFI = 0.97, TLI = 0.97).

DIF detection:

By manifest DIF detection, items that function differently among Turkey and Singapore group members were examined. A Likelihood Ratio Test for DIF (IRT-LR) was used to detect DIF. This analysis was conducted using the computer program IRT-LR DIF (Thissen, 2001). The other procedure for manifest DIF detection in this study was Lord's Chi-square (χ^2) method from the IRT models. IRT-LR and Lord's χ^2 analysis was conducted with 2PL model, but only uniform DIF was reported in this study.

IRT-LR: This procedure is closely related to the IRT model and includes hypothesis testing of item response theory parameters, which are slope, guessing or difficulty parameters (Thissen, 2001). IRT-LR compares the results of the compact and augmented model. A compact model assumes that item parameters are equal for focal and reference groups. It means that items do not include DIF across groups (Thissen, 2001). On the other hand, the augmented model assumes that the parameters of item *i.* can differ for focal and reference groups, but other items supposed to be equal in terms of parameters across these groups (Cohen, Kim & Wollcak, 1996). IRT-LR is the difference between likelihood ratios, calculated from the compact model and augmented model. Distribution of IRT-LR is as a chi-square with the difference in the degree of freedom between the compact and the augmented models. This procedure is appropriate to polytomous and dichotomous data. Besides, with this procedure, it is possible to detect uniform and non-uniform DIF. According to Greer (2004), items will detect DIF with IRT-LR method, when G^2 values are between the following intervals (Greer, 2004):

3.84<G^2<9.4 negligible dif (A level)

9.4≤G^2< 41.9 middle level dif (B level)

G^2≥ 41.9 high level dif (C level)

Lord's Chi-Square (χ^2): Lord's χ^2 test is related to the differences in the variance-covariance matrix of difficulty and discrimination parameters. This method is based on the differences in the item parameters obtained for the reference and focal groups (Hambleton & Swaminathan, 1985). Lord's chi-square test is explained in Equation 1 (Kim, 2010):

$$\chi_i^2 = (a_{diff} b_{diff} c_{diff})' \Sigma^{-1} (a_{diff} b_{diff} c_{diff})$$

In this formula Σ^{-1} refers to the inverse variance-covariance matrix for differences in item parameter estimates; a_{diff}; refers to the difference of parameters obtained for reference and focal group; b_{diff}, is the difference between difficulty parameters obtained estimated for reference and focal group, and c_{diff} is the difference between pseudo guessing parameters among groups.

The obtained χ^2 statistic is distributed as chi-square at the degree of freedom "1' for 1PL model, with two degrees of freedom for 2PL model and with three degrees of freedom for 3PL model (Lord, 1980). When the χ^2 statistical value exceeds the critical value, the item is thought to contain DIF based on the relevant level of significance. Analyses related to this method were carried out in "difR" library in R 3.1.2 software. It is determined that an item contains DIF, when it is found to be significant at a 0.5 level. For DIF, item effect size was calculated, where the difference between item difficulties among reference and focal groups was -2.35 times. This effect size is similar to Mantel Haenszel's Δ_{MH}. To classify the effect sizes, ETS delta scale was used (Holland & Thayer, 1988; Magis, 2018; Penfield, 2007). Based on the size of this value, assessed items showed DIF at the A, B, or C level.

To detect DIF among latent classes, MRM was conducted in WINMIRA (von Davier, 2001) program. After deciding on the number of latent classes, DIF items were detected using the Mantel Haenszel method. ΔMH Coefficient suggested by Roussus, Scnipke and Pashley (1999) was;

$$\Delta MH = -2.35 \ln(\alpha) = -2.35 \ln\left[e^{-1.7a(b_R - b_F)}\right] = 4\alpha((b_R - b_F)$$

b_R is the item difficulty for focal group and b_F is the item difficulty for reference group. Based on ΔMH value intervals it is decided if an item contains DIF. The intervals are listed below:

If |ΔMH | < 1 DIF is negligible (A level)

1 ≤ |ΔMH | <1.5 middle level DIF (B level)

If |ΔMH | ≥1.5 high level DIF (C level)

Results

The examinees' responses were investigated with MixIRT. The fit of one class model was compared with the fit of two and three class models by comparing their Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Consistent Akaike Information Criterion (CAIC) statistics. Table 1 shows the information criteria for latent class models.

Table 1

Information Criteria for Latent Class Models

Number of Class	AIC	BIC	CAIC
1	22611.39	22701.64	22719.64
2	22409.28	**22594.79**	**22631.79**
3	22357.81	22638.59	22694.59

According to Table 1, the two-class model had the smallest BIC and CAIC values. For this reason, the model with two latent classes with sizes .56 and .44 was selected. Based on this model, we can interpret that the manifest group and latent class overlapping was poor. The distribution of examinees in latent classes according to the two latent class models by country is given in Table 2.

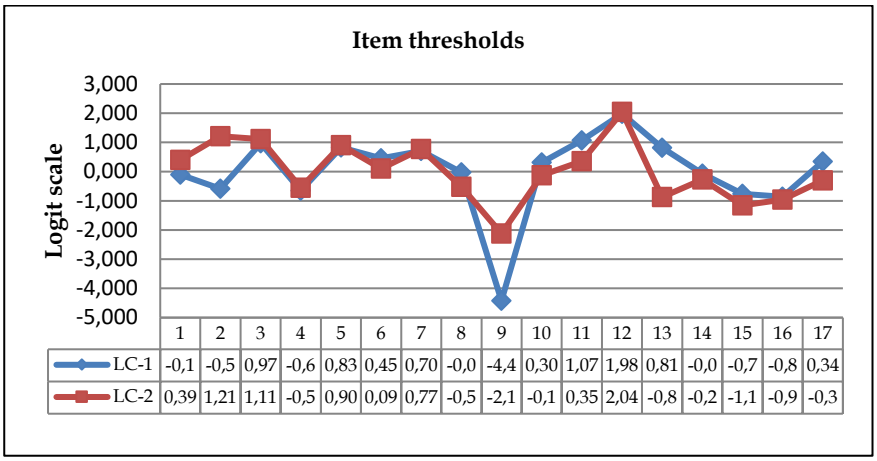
Table 2

Cross-tabulations of Country and Class Membership

Country		LC-1	LC-2	Total
Singapore	n	470	144	614
	%	76,55	23,45	100
Turkey	n	146	352	498
	%	29,32	70,68	100
Total	n	616	496	1112
	%	55,40	44,60	100

LC: Latent class

According to Table 2, there were 616 examinees in LC-1. In this class, 470 participants (76.55%) were from Singapore and 146 (29.32%) were from Turkey. In the second class, there were 496 examinees in total. Besides, 144 (23.45%) of examinees were from Singapore, and 352 (70.68%) were from the Turkey sample in the second class. Figure 1 displays the thresholds (difficulty) parameters for two classes.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
LC-1	-0,1	-0,5	0,97	-0,6	0,83	0,45	0,70	-0,0	-4,4	0,30	1,07	1,98	0,81	-0,0	-0,7	-0,8	0,34
LC-2	0,39	1,21	1,11	-0,5	0,90	0,09	0,77	-0,5	-2,1	-0,1	0,35	2,04	-0,8	-0,2	-1,1	-0,9	-0,3

Figure 1. *Item difficulty parameters obtained for each latent class*

According to Figure 1, every class had similar item difficulties except for items 1, 2, 9 11 and 13. In general, LC-1 found items easier than LC-2 in the first part of the test. But it is interesting that in the second part of the test, LC-2 found the test easier than LC-1.

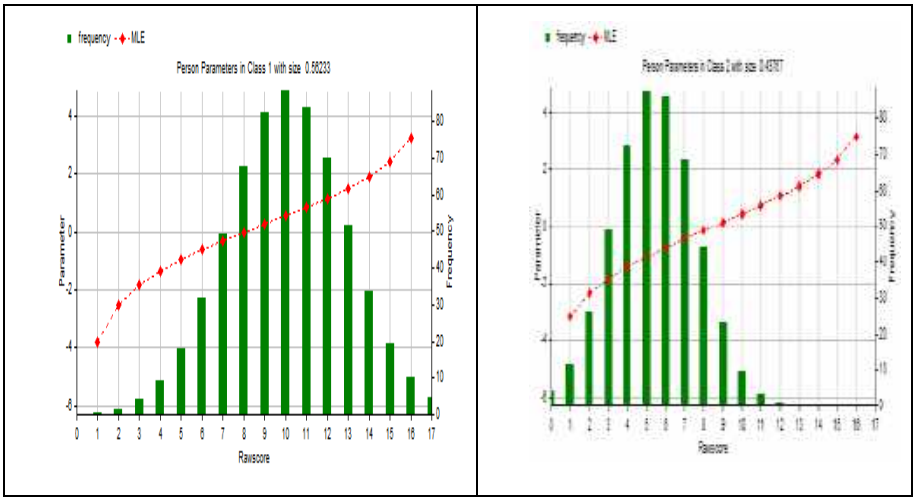


Figure 2. *Ability parameters obtained for each latent class*

As can be seen in Figure 2, ability parameters were higher in LC-1 than LC-2. In this context, it can be interpreted that the examinees in LC-1 achieved higher success than in LC-2. Membership in LC-1 included more examinees from Singapore (high-performing), and fewer from Turkey (low-performing). In addition, there were more examinees in LC-2 from Turkey and fewer from Singapore. According to these results,

it can be interpreted that the items were medium level for examinees in LC-1 and were at a difficult level for the examinees in LC-2.

Table 3

Mixture Rasch Difficulty Estimates for 2-Class Solution

	LC-1	LC-2			
Item	Est.	Est.	Diff.	-2.35*Diff.	DIF Level
1	-0.106	0.395	-0.501	1.177	**B**
2	-0.583	1.213	-1.796	4.221	**C**
3	0.974	1.110	-0.136	0.320	A
4	-0.654	-0.553	-0.01	0.237	A
5	0.831	0.906	-0.075	0.176	A
6	0.459	0.099	0.360	-0.847	A
7	0.705	0.773	-0.068	0.160	A
8	-0.027	-0.525	0.498	-1.170	**B**
9	-4.420	-2.119	-2.301	5,407	**C**
10	0.309	-0.125	0.434	-1.020	**B**
11	1.071	0.351	0.72	-1,691	**C**
12	1.980	2.043	-0.063	0,148	A
13	0.816	-0,.873	1.68	-3.969	**C**
14	-0.067	-0.267	0.200	-0.470	A
15	-0.76	-1.161	0.396	-0.942	A
16	-0.865	-0.963	0.098	-0.230	A
17	0.342	-0.303	0.646	-1.516	**C**

Est: Estimation, Diff: Difference

According to Table 3 it can be said that 8 of 17 items displayed DIF among latent classes. Items 1, 8 and 10 had DIF at B level. The items 2, 9, 11, 13 and 17 showed C level DIF.

Table 4

Lord's Chi Square DIF Solutions Among Turkey and Singapore

Item	Statistic	p-value	Δ_{χ^2}	DIF Level
1	0.27	0.60	-0.32	A
2	122.96	0.00***	4.19	**C**
3	3.22	0.07	0.55	A
4	2.49	0.11	0.41	A
5	0.92	0.33	-0.49	A
6	6.83	0.00***	0.78	A
7	6.16	0.01**	0.78	A
8	0.01	0.91	-0.18	A
9	0.45	0.50	-0.48	A
10	5.77	0.02*	0.69	A
11	19.90	0.00***	-1.74	**C**
12	6.85	0.00***	1.24	**B**
13	72.92	0.00***	-3.04	**C**
14	4.47	0.03*	-0.86	A
15	0.79	0.37	-0.46	A
16	0.00	0.94	-0.17	A
17	4.79	0.03*	-0.89	A

*Sig. codes: '***' 0.001 '**' 0.01 '*' 0.05*

Table 4 shows the χ^2 statistics, p significance and Δ_{χ^2} values obtained by Lord's χ^2 methods. The results indicated that items 6, 7, 10 and 14 were identified as DIF items at A level. These items can be considered, including DIF at negligible effect size. Only item 12 had B level DIF. The items 2, 11 and 13 were detected as DIF items at C level between Turkey and Singapore.

Table 5

The IRT-LR Solutions between Turkey and Singapore

Item	G^2	df	DIF Level
1	0,5	1	-
2	6,5	1	A
3	0,2	2	-
4	7,4	1	A
5	12,6	1	**B**
6	0,8	1	-
7	2,4	2	-
8	2,1	2	-
9	1,2	1	-
10	9,6	1	**B**
11	7	1	A
12	2,9	2	-
13	1,9	1	-
14	0,1	1	-
15	1,1	2	-
16	3	2	-
17	3,5	2	-

df: Degrees of Freedom

According to Table 5, it was determined using the IRT-LR technique that item 5 and item 10 included DIF. These items showed B level DIF between Turkey and Singapore samples. For items 5 and 10, G^2 test of the hypothesis that b parameters were equal for the reference and focal groups did not exceed 3.84 (the a = 0.05 critical value of the χ^2 distribution for one degree of freedom). To compare DIF detection methods, a summary of information was given in Table 6.

Table 6

Comparing Results of Latent Class and Manifest Groups Approaching

Method	Items with DIF (B and C Level)
MRM	1,2,8,9,10,11,13 and 17
Lord's χ^2	2,11,12 and 13
IRT-LR	5 and 10 (2,4,11)

According to Table 6, items 2,10,11 and 13 were determined including DIF on two different techniques results and DIF level of these items were not negligible. On the other hand, items 2 and 11 could be detected as DIF items with all techniques, where IRT-LR detected these items at a negligible level. Finally, it can be said that MRM could detect more items than manifest group methods. When we analyzed these items from PISA booklet, it was seen that item 2 is related to Earth's temperature, items 10 and 11 were related to Airbags, and item 13 and 12 were related to the subject extinction of the dinosaurs.

Discussion, Conclusion and Recommendations

A test item should be able to measure ability without involving characteristics of examinees who are indifferent subgroups. This is because examinees with equal abilities should have the same probability to answer an item correctly, even though they are in different subgroups. When this condition is not provided, this item is considered as a biased item. To investigate bias, one way is to examine this item in terms of differential item functioning (DIF). With DIF analysis, we can see whether an item differs in functioning among the reference and focal groups. When an item functions differentially, then we can infer with qualitative studies whether this item is biased.

This study aimed to examine DIF in PISA cognitive science items between Turkey and Singapore samples and among latent classes that emerged from these countries. In this study, it was seen that data were fit to two latent class models. This suggests a secondary nuisance dimension that is not measured by the item (Choi et al., 2015). The distribution of emerging latent classes showed that there were many members from Singapore in first class, where the second class consisted mostly of members from Turkey sample. It is a common idea that country can represent the class membership best to define reference and focal groups. However, this approach is not very accurate. For instance, approximately 23% of the examinees in the Singapore sample belonged to LC-2 at the same time. According to this, item-based interpretation for each latent class may give more insight into what constitutes the characteristics of each latent class (Nijlen & Janssen, 2008). Looking at item difficulties, items were at medium level for the members in first latent class, but they were difficult for members in the second class. This finding is consistent with the results of Yalcin (2018). So, we can say that conducting reference and focal groups in terms of the country may not be sufficient to represent equal ability level groups.

When DIF was investigated with MH among these classes, it was seen that three items showed B level and five items showed C level DIF among latent classes. According to manifest DIF results with Lord's χ^2 DIF method, it was observed that one item included B level DIF and three items showed C level DIF between Turkey and Singapore samples. According to another result of this study with the IRT-LR method, two items showed B level DIF between countries. Considering the results of the research, it is possible to state that the latent class approach can detect most DIF items than manifest group methods. Maij-de Meij et al. (2010) examined DIF among latent classes with Lord's χ^2 method. They found that DIF studies conducted with

latent classes were more effective than manifest group methods. In addition, if the correlation between the manifest group and latent class decreases, the effectiveness of the manifest group method decreases. Cohen and Bolt (2005) pointed out that ethnical features were related with latent classes. In addition, Asil (2012) specified that DIF in PISA items generally arises from translation and adaptation applications. Choi et al. (2015) applied 3PL MixIRT to TIMSS 2007 data among seven countries. They found that data fit to the two-class model, where the first latent class consisted of high achievement countries and the second class consisted of low achievement countries. Karadavut (2017) revealed that there appeared only one latent class in the PISA Turkey sample when groups were considered in terms of gender. According to Cohen and Bolt (2005) and Yalcin (2018), the gender variable is weakly correlated with latent class membership. According to the obtained results and literature, especially in cultural comparisons, more items can be detected, including DIF with a latent class approach. . It is also stated that at least two latent classes appeared in DIF studies based on culture. In this study, the appearance of two latent classes pointed the DIF in the items. So, latent class approach can be more effective to give ideas about the source of DIF if we examine the properties of latent classes.

According to the other finding of this study, Lord's χ^2 produced similar results with MRM method in 3 items and IRT-LR methods produced similar results with MRM only in 2 items. However, DIF magnitude obtained from these methods was different. This may occur due to the difference of DIF level intervals belonging to classifications (Arikan Akin, 2015). When three methods were compared, IRT-LR showed lower performance to detect DIF items. Gao (2019) compared Logistic Regression (lR), IRT-LR and Multiple Indicator and Multiple Causes (MIMIC) models performance in terms of detecting DIF with a simulation study and pointed that the LR and IRT-LR procedures were powerful to detect non-uniform DIF. On the other hand, the MIMIC model method was better than the IRT-LR under most conditions to identify DIF items. In the current study, it was aimed to detect uniform DIF, but not nonuniform DIF. This may explain why the IRT-LR procedure showed lower performance than the other methods in this study.

In summary, it can be concluded that DIF determination based on latent classes is a good alternative when compared with manifest DIF detection methods. On the other hand, to detect uniform DIF, it can be suggested using Lord's χ^2 method instead of IRT-LR. Items, which were detected to show DIF should be examined in terms of item bias. In the future, qualitative studies can be conducted to investigate items in terms of bias among Turkey and Singapore. These DIF items were related to subjects such as airbags, earth temperature, and extinction of dinosaurs. It may be appropriate to provide training in these areas in schools. This study had some limitations. First, it examined only uniform DIF. Therefore, future studies can focus on nonuniform DIF among latent classes. What is more, future studies might compare DIF results across many countries. Simulation studies may be effective to compare latent class and manifest group approaches based on IRT.

References

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In P. W. Holland &H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Erlbaum.

Arikan, A. C. (2015). Comparison of Likelihood Ratio Test (LRT), Poly-SIBTEST and Logistic Regression in Differential Item Functioning (DIF) Detection Procedures. *e-International Journal of Educational Research*, *6*(1), 1-16.

Asil, M. & Gelbal, S. (2012). Cross-cultural Equivalence of the PISA Student Questionnaire. *Education and Science, 37*, 236-249.

Bilir, M. K.(2009). *Mixture Item Response Theory-Mimic Model: Simultaneous Estimation of Differential Item Functioning For Manifest Groups and Latent Classes*. Doctoral Dissertation. Florida State University

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.). *Testing structural equation models* (pp. 136-162). Newsbury Park, CA: Sage.

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.

Cho, S. J. (2007). *A multilevel mixture IRT model for DIF analysis.* Unpublished doctoral dissertation, University of Georgia: Athens.

Cho, S. J., Suh, Y., & Lee, W. Y. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, *35*(1), 48-61. https://doi.org/10.1111/emip.12093

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.

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 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.

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.

Gierl, M. J., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17*(3), 241-264. doi: https://doi.org/10.1207/s15324818ame1703_2

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 & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Hu, L. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

Jöreskog, K.G. ve Sörbom, D. (1993). *Lisrel 8: Structural equation modeling with the SIMPLES comman language*. Lincolnwood, IL: Scientific Software International.

Kabasakal, K. A., Gök, B., Kelecioğlu, H., & Arsan, N. (2012). Degisen madde fonksiyonunun belirlemesinde kullanilan farkli yontemlerin karsilastirilmasi: bir simulasyon calismasi. *Hacettepe University Journal of Education*, *43*(43), 270-281. Retrieved from https://dergipark.org.tr/tr/pub/hunefd/issue/7795/102030

Karadavut, T. (2017). DIF analysis with manifest and latent groups: Analysis of PISA 2012 mathematics data from Turkey. *The Eurasia Proceedings of Educational & Social Sciences*, *8*, 103-106. Retrieved from: http://static.dergipark.org.tr/article download/cf1d/c744/dd82/5a341ef0af7f6.pdf?

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. https://doi.org/10.1111/j.1745-3984.1990.tb00751.x

Levent, F. & Yazici, E. (2014). Singapur egitim sisteminin basarisina etki eden faktorlerin incelenmesi. *Journal of Educational Sciences*, *39*, 121-143. doi: 10.15285/EBD.2014397401

Magis, D. (2018). Collection of methods to detect dichotomous differential item functioning (DIF). *Package 'difR'*.

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.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7(2), 105-118.

Millsap, R.E. & Everson, H.T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement*, 17(4), 297–334.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (Eighth Edition). Los Angeles, CA: Muthén & Muthén.

Oliveri, M. E., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing*, *13*(3), 272-293.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education*, *20*(3), 335-355.

Rost, J. (1990). *Rasch models in latent classes: An integration of two approaches to item analysis*. Applied Psychological Measurement, *14*, 271–282.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics*, *24*(3), 293-322.

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective*. Unpublished doctoral dissertation, University of Maryland, College Park.

Sawatzky, R., Ratner, P. A., Kopec, J. A., & Zumbo, B. D. (2012). Latent variable mixture models: A promising approach for the validation of patient reported outcomes. *Quality of Life Research*, *21*(4), 637-650.

Sunbul, S. O. & Sunbul, O. (2016). Değişen madde fonksiyonunun belirlenmesinde kullanılan yöntemlerde I. Tip hata ve güç çalışması. *Elementary Education Online, 15*(3), 882-897.

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.

Van Nijlen, D., & Janssen, R. (2008). Mixture IRT-models as a means of DIF-detection: Modelling spelling in different grades of primary school. In *Annual Meeting of the National Council on Measurement in Education, Date: 2008/01/01-2008/01/01, Location: New York*.

Von davier, M. (2001). *WINMIRA 2001: Software for estimating Rasch models, mixed and hybrid Rasch models and latent class analysis* [Computer software]. Retrieved from: http://www.von-davier.com/

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing*, *6*(3), 287-300.

Yalcin, S. (2018). Determining differentisl item functioning with the mixture item response theory. *Eurasian Journal of Educational Research, 74,* 187-206.

Yuksel, S. (2012). *Analyzing differential item functioning by mixed rasch models which stated in scales*. Yayımlanmamış Doktora tezi. Ankara University Graduate School of Health Sciences, Ankara.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic Regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottowa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Değişen Madde Fonksiyonunun Belirlenmesine Örtük Sınıf Yaklaşımı: PISA 2015 Fen Örneklemi

Atıf:

Uyar, S. (2020). Latent class approach to detect differential item functioning: PISA 2015 science sample. *Eurasian Journal of Educational Research 88*, 179-198. DOI: 10.14689/ejer.2020.88.8

Özet

Problem Durumu: Aynı yetenek düzeyinde farklı gruplarda yer alan bireylerin bir test maddesini doğru yanıtlama olasılıkları eşit olmalıdır. Eğer madde, gruplardan birine daha fazla avantaj sağlıyorsa maddenin yanlı olduğu düşünülür. Yanlı maddeler sistematik hata içerir, bu nedenle puanların geçerliğini düşürür. Aynı zamanda puanların gruplar arasında doğru bir şekilde karşılaştırılmasına tehdit oluşturur. PISA ve TIMSS gibi uluslararası sınavlar kültür, dil, sosyoekonomik düzey ya da cinsiyet gibi farklı gruplarda yer alabilen bireylere uygulanmaktadır. Bu demografik özellikler her ne kadar testle ölçülmek istenmese de bireyin performansına etki edebilir. Bu nedenle testler uygulanmadan önce madde yanlılığı açısından incelenmelidir. Yanlılığın ilk işareti maddenin aynı yetenek düzeyindeki iki grupta farklı fonksiyonlaşmasıdır. Değişen madde fonksiyonu (DMF), yanlı olabilecek maddelerin belirlenmesinde istatistiksel bir tekniktir. Bu yöntem cinsiyet ya da ülke gibi gözlenen gruplardan birini referans diğerini odak grup olarak belirlendikten sonra gruplar arasında madde parametrelerinin karşılaştırılmasına dayanır. Ancak gözlenen gruba dayalı yöntemlerde bazı sınırlılıklar bulunmaktadır. Bir gözlenen grubun (örneğin kızlar) içerisinde yer alan tüm bireyler aynı madde bakımından avantajlı ya da

dezavantajlı sayılmaktadır. Oysa madde aynı grup içerisinde yer alan farklı bireyler için avantajlı ya da dezavantajlı olabilir. Bu varsayımın sebebi gözlenen grupların homojen grup olma düşüncesinde yatmaktadır. Aynı zamanda bu gözlenen grup DMF'nin kaynağı olarak yansıtılır. Varsayımın sağlanmasının düşük olmasına yönelik eleştiriler örtük sınıflara göre DMF belirlemenin, DMF kaynağını bulmada daha etkili olduğunu belirtmişlerdir. Yapılan çalışmalar DMF incelemede örtük sınıf yaklaşımının avantaj sunabileceğini, DMF kaynağını herhangi bir değişken setinden bağımsız olarak incelemeye fırsat vereceğini belirtmektedir.

Araştırmanın Amacı: Bu çalışmanın amacı örtük sınıfa ve madde tepki kuramı çerçevesinde yöntemlerden gözlenen grup yaklaşımıyla belirlenen DMF sonuçlarının karşılaştırmaktır.

Araştırmanın Yöntemi: Araştırmada farklı kültürden biraraya gelen bireylerin örtük sınıfları yansıtma oranının yüksek olması nedeniyle PISA 2015 uygulamasına katılan Singapur ve Türkiye örneklemleri kullanılmıştır. Bu çalışmada PISA bilişsel fen maddelerinden yalnızca ikili (1-0) şeklinde puanlananlar dikkate alınmıştır. Çalışmaya maddeleri ortak olarak işaretleyen Türkiye örnekleminden 498, Singapur'dan 614 öğrenci dahil edilmiştir. Örtük sınıfların belirlenmesinde Karma Madde Tepki Kuramı (KTMK) modelinden yararlanılmıştır. Bu analiz Winmira (2001) programında yapılmıştır. Örtük sınıflar arasında DMF karşılaştırmak üzere Mantel-Haenszel tekniği kullanılmıştır. Gözlenen gruplara Dayalı DMF'yi belirlemek üzere Lord'un ki-kare (χ^2) yöntemi ve Madde Tepki Kuramı Olabilirlik Oranı (MTK-OO) yönteminden yararlanılmıştır. Bu analizler ise R programında 'difR' kütüphanesinde gerçekleştirilmiştir.

Araştırmanın Bulguları: KMTK modeline göre elde edilen bilgi kriterleri (AIC, BIC ve CAIC) bir sınıflı, iki ve üç sınıflı modellerde karşılaştırılmıştır. BIC ve CAIC istatistikleri indeksleri iki sınıflı modelde en küçük değeri aldığından iki sınıflı model kabul edilmiştir. Örtük sınıflarda ülkelerin dağılımı incelendiğinde birinci örtük sınıfta Singapur'dan daha çok öğrencinin, ikinci örtük sınıfta ise Türkiye'den daha çok öğrencinin olduğu görülmüşür. Madde güçlükleri incelendiğinde birinci örtük sınıfta yer alan öğrenciler için maddelerin orta güçlükte olduğu, ikinci örtük sınıfta yer alan bireyler için daha zor olduğu görülmüştür. Maddeler örtük sınıflar arasında DMF bakımından karşılaştırıldığında 3 maddenin B düzeyinde, 5 maddenin ise C seviyesinde DMF içerdiği görülmüştür. DMF analizi Türkiye ve Singapur ülkeleri arasında Lord'un χ^2 yöntemiyle yapıldığında 12. maddenin B düzeyinde, 2, 11 ve 13. maddeler olmak üzere üç maddenin C düzeyinde DMF gösterdiği görülmüştür. MTK-OO yöntemi ile yapılan DMF analizi sonucunda 5. ve 10. maddeler B düzeyinde DMF göstermiştir. Gözlenen gruba ve örtük sınıfa dayalı DMF yaklaşımları karşılaştırıldığında 2, 10, 11 ve 13. maddelerinin en az iki yöntemde DMF gösterdiği, DMF madde sayısının örtük sınıf yaklaşımıyla daha fazla olduğu görülmüştür.

Araştırmanın Sonuçları ve Öneriler: Bu araştırmada örtük sınıf ve gözlenen gruba dayalı DMF yaklaşımları DMF'li bulunan madde sayıları bakımından PISA 2015 fen testi üzerinde karşılaştırılmıştır. DMF'li bulunan madde sayısı örtük sınıf yaklaşımında daha fazladır. Maij-de Meij ve diğerleri (2010) Lord'un χ^2 yöntemiyle örtük sınıflar

arasında DMF karşılaştırdıklarında örtük sınıfa göre daha fazla DMF'li madde bulduklarını belirtmişlerdir. Ayrıca, gözlenen grup ve örtük sınıf arasındaki korelasyon düştükçe gözlenen grup yönteminin etkililiğinin azaldığını belirtmişlerdir. Cohen & Bolt (2005) kültürel özelliklerin örtük sınıflarla ilişkili olduğunu belirtmiştir. Asil (2012) ise PISA maddelerinin çeviri ve uyarlama uygulamalarında DMF içereceğini vurgulamıştır. Choi ve diğerleri (2015) maddelerin 7 ülke arasında DMF bakımından karşılaştırdıkları çalışmalarında iki örtük sınıfın ortaya çıktığını, birinci örtük sınıfın yüksek başarı gösteren ülkeler, ikinci örtük sınıfın başarısı düşük ülkelerden oluştuğunu ifade etmişlerdir. Karadavut (2017), Yalcin (2018) ve Cohen & Bolt (2005) cinsiyet değişkeni dikkate alınarak örtük sınıf oluşturduklarında tek bir sınıfın ortaya çıktığını be nedenle cinsiyetin örtük sınıfla düşük düzeyde ilişki gösterdiğini belirtmişledir. Elde edilen sonuçlar ve alanyazın birlikte değerlendirildiğinde özellikle kültürler arası karşılaştırmalarda örtük sınıf yaklaşımına göre DMF'li bulunan madde sayısı daha fazla olabilmektedir. Ayrıca kültüre göre DMF çalışmalarında en az iki örtük sınıfın ortaya çıktığı da belirtilmektedir. Bu çalışmada da iki örtük sınıfın ortaya çıkması maddelerde DMF'ye işaret etmekte ve örtük sınıfların özellikleri ayrıca incelenirse DMF'ye neden olan kaynağın bulunması konusunda da fikir verme bakımından daha etkili olabileceğini göstermektedir. Araştırmada ulaşılan bir diğer sonuç Lord'un χ^2 yönteminin KMTK ile 3 maddede, MTK-OO yönteminin 2 maddede benzer sonuçlar verdiğini göstermiştir. Ancak, DMF etki büyüklüğü farklıdır. Bunun nedeni ise DMF aralıklarını sınıflama yöntemlerinden kaynaklanabilir (Arikan Akin, 2015). MTK-OO yöntemi ise bu çalışmada en az sayıda DMF bulan yöntem olmuştur. Gao (2019)'a göre MTK-OO yöntemi tek biçimli olmayan DMF'yi bulmada etkilidir. Bu çalışmada yalnızca tek biçimli DMF incelendiğinden sonuç bu şekilde çıkmış olabilir. Sonuç olarak örtük sınıf yaklaşımının DMF bulmada alternatif bir yaklaşım olarak ele alınması, DMF kaynağını yalnızca bir alt gruba dayalı olarak değil örtük sınıf içerisinde oluşan alt grupları inceleyerek bulabilmeye olanak sunması bakımından kullanılması önerilmektedir. Gözlenen gruba dayalı yöntemlerden Lord'un χ^2 yöntemi tek biçimli DMF'yi inceleyen çalışmalarda kullanılabilir. DMF gösteren maddeler, madde yanlılığı açısından nitel araştırmalarla incelenebilir. Ülkeler arasında DMF gösteren maddelerin hava yastığı, küresel ısınma ve dinosorların neslinin tükenmesi ile ilgili olduğu görülmüştür. Bu nedenle okullarda benzer konularda eğitim ile destek verilmesi önerilebilir. İleriki araştırmalarda tek biçimli olmayan DMF bakımından örtük sınıf yaklaşımı incelenebilir. Farklı ülkelerde çalışmalar tekrar edilebilir.

Anahtar Sözcükler: Örtük sınıf, karma madde tepki kuramı, değişen madde fonksiyonu, madde yanlılığı.