Int. J. Asst. Tools in Educ., Vol. 4, Issue 2, (2017) pp. 223-239

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Investigating Different Theory-Based Differential Item Functioning Methods Under Different Scoring Situations

Hüseyin Selvi [*1], **Devrim Özdemir Alıcı** [2]

[1]Mersin University, Medical Faculty, Medical Education Department, Turkey
[2]Mersin University, Faculty of Education, Department of Measurement and Evaluation in Education, Turkey

Abstract: This study intended to investigate the operations of various theory-based Differential Item Functioning (DIF) methods under different scoring situations. In this context, data obtained from conducting a Verbal Reasoning Skills Test (scored both in dichotomous and weighted scores) on 1593 individuals was examined via Mantel-Haenszel/Mantel, Standardization and Likelihood Ratio Test techniques and items that point to DIF and numbers of items with DIF for each technique and scoring situation were identified. Results show that Classical Test Theory (CTT) based DIF methods generally provided results that were consistent in themselves and that these results were not significantly different from each other. Agreement between techniques based on Item Response Theory (IRT) and Classical Test Theory (CTT) was found to be low under different scoring situations and results were significantly different from each other.

ARTICLE HISTORY

Received: 31 March 2017
Revised: 28 July 2017
Accepted: 31 July 2017

KEYWORDS

Item Scoring, Differential Item Functioning, DIF, Item Bias, Bias, Mantel-Haenszel, Likelihood Ratio Test

1. INTRODUCTION

In terms of psychological and educational sciences, reliability of a measurement tool is affected from variables such as test length, group homogenity and item difficulty. Validity is a broader term that includes reliability as well and is affected from many variables such as the structure of the variables that is to be measured, reason for use and bias (Aiken, 2000; Anastasi and Urbina, 1997; Crocker and Algina, 1986; Magnusson, 1967; Murphy and Davidshofer, 2005; Thorndike, 1982;). In other words, random error sources affect the reliability whereas systematic error sources affect validity.

Examination of variables that affect validity points to bias as a crucial variable. In this context, bias can be defined as a systematic source of variation that affects the validity of the scores (Osterlind, 1983). Bias that occurs as a systematic variation source and affects the validity is defined as "the difference between the possibilities of the individual within different subgroups

*Corresponding Author E-mail: hsyn_selvi@yahoo.com.tr

ISSN: 2148-7456 online /© 2017 DOI: 10.21449/ijate.334584

with same ability level to give the right answer to the related test item (Angoff, 1993). From this definition, in the studies regarding the determination of the bias initially, it is understood that it is necessary to match the individuals in different subgroups regarding the ability levels and to examine statistically the item answering conditions of these individuals. This situation is defined as the examination of whether there is Differential Item Function (DIF) in the items or not. To maintain the distinction between relative difficulty and bias, we refer to the raw or uninterpreted relative difficulty as differential item functioning or DIF (Camilli & Shepard, 1994).

It is required that the items with detected DIF (relative advantage for one group over the entire ability range defined as Uniform DIF; relative advantage at some ability levels for one group and relative advantage for the other group at other ability levels defined as Non-Uniform DIF) should be checked by the experts and whether the DIF is due to another source than the desired measured quality shall be investigated. In cases that the DIF is detected to be caused by another source than the desired measured quality, it can be convinced of that the related item(s) is/are biased (Camilli & Shepard, 1994; Selvi & Alıcı, 2016; Zumbo, 1999).

Therefore, studies to identify DIF which presents a preventable structure as a source of systematic variation are included in validity studies and it is known that methods and techniques used in this regard are getting more comprehensive along with developing computer infrastructure and theories. Methods and techniques based on Classical Test Theory (CTT) and Item Response Theory (IRT) are among the most comprehensive ones in this context.

Some of the methods and techniques to determine DIF developed in line with CTT and IRT and widely used today are briefly presented below:

Standardized Proportion Difference Measure: The technique based on calculations using the Contingency Table mainly calculates the weighted means between the differences of item difficulty values (*p*) separately assessed for each group by the number of individuals.

For the standardization technique, Equation 1 presents the formula used for Signed *p* Difference (SPD-X) DIF for dichotomous scored items, Equation 2 presents the formula used for Unsigned *p* Difference (SPD-X) and Equation 3 presents the formula used for Signed Mean Difference (SMD) for weighted items (Dorans & Holland, 1993; Gonzales et. al., 2010).

$$SPD-X=\frac{\sum_{j=1}^{s} n_{Fj}\Delta_{pj}}{\sum_{j=1}^{s} n_{Fj}} \quad (1)$$

$$UPD-X=\sqrt{\frac{\sum_{j=1}^{s} n_{Fj}(\Delta_{pj})^2}{\sum_{j=1}^{s} n_{Fj}}} \quad (2)$$

$$SMD=\sum_{j=1}^{K} N_{F.j}MD_j/N_{F.}\ MD_j=\sum_{c=1}^{c} N_{Fcj}R_c/N_{F.j}-\sum_{c=1}^{c} N_{Rcj}R_c/N_{F.j} \quad (3)$$

j : Paired score level for both groups,

n_{Fj} : Number of individuals in the focus group for paired j score level for both groups,

N_{Rj}: Number of individuals in the reference group for paired j score level for both groups,

K : Number of paired scores for both groups,

Δ_{pj} : Item difficulty difference of the related item for both groups at paired j score level

Focus/focal and reference groups: Focus/focal group term used here is usually refers to the minority (disadvantaged/interest labeled) group and the reference group term refers to the majority (advantageg/not interest labeled) group (Santelices & Wilson, 2012).

Values obtained from these equations are reaches from -1 to +1.values between -0.05 and 0.05 interval show negligible DIF existence whereas values between -1 and -0.05 and 0.05 and 1 intervals show not negligible DIF existence (Gonzales et. al, 2010).

Mantel-Haenszel Log Odds Ratio: The technique based on calculations using the Contingency Table and combines odds ratios obtained from all sub group members paired in terms of total scores by taking group weights into consideration. Related formula is presented in Equation 4.

$$\alpha_{MH} = \frac{\sum_j \frac{p_{r_j} q_{f_j} n_{r_j} n_{f_j}}{n_j}}{\sum_j \frac{q_{r_j} p_{f_i} n_{r_j} n_{fi}}{n_j}} \qquad (4)$$

$\beta_{MH} = \log_e(\alpha_{MH})$

p_{r_j} :Item difficulty of the related item at *j* score level,

q_{f_j} :Rate of incorrect answers for the related item at *j* score level.

Positive 'β' values obtained from these equations point to the existence of DIF in favor of reference group whereas negative values point to the existence of DIF in favor of focus group. β values of '0' show non-existing DIF (Angoff, 1993). Another criterion related to β values was developed by Educational Testing Service (ETS). Accordingly;

When D=-2.35* β_{MH} ;

A: Item without DIF is represented by $|D| < 1$

B: Item with significant DIF is represented by $1 < |D| < 1.5$

C: Item with high levels of DIF is represented by $|D| > 1.5$

(Zwick, 2012).

Literature includes various criteria for the interpretation of 'Δp_j'and β values. However, it is also suggested to apply statistical tests to determine whether these values are significantly different from '0' (Camilli & Shepard, 1994). Mantel-Haenszel Chi-Square Test, widely used in literature by Mantel-Haenszel (1959) was proposed for use to undertake this test (cited in: Dorans & Holland, 1993). Accepting 'H_0' hypothesis at the end of the test points to non-existence of DIF whereas rejecting the hypothesis points to existence of DIF. At least one focus-reference group is needed to conduct the test.

And another formula proposed by Mantel-Haenszel (1959) to test the statistical significance of α_{MH} and β_{MH} values and which shows χ^2 distribution at '1' degree of freedom is presented in Equation 5 (cited in: Dorans & Holland, 1993).

$$MH\,\chi^2 = \frac{\left\{\sum_{j=1}^{s}(Aj - E(Aj)) - 1/2\right\}^2}{\sum_{j=1}^{s} VAR(Aj)}$$

$$VAR(Aj) = \frac{n_{Rj}n_{Fj}m_{1j}m_{0j}}{T_j^2(T_j - 1)} \qquad (5)$$

$$E(Aj) = \frac{n_{Rj}m_{1j}}{T_j}$$

Aj : Number of correct answers observed at j score level for the reference group for the related item,

$E(Aj)$: Number of correct answers expected at j score level for the reference group for the related item

T_j : Number of total individuals at j score level

m_{1j} : Number of individuals in focus and reference groups who provided correct answers at j score level

m_{0j} : Number of individuals in focus and reference groups who provided incorrect answers at j score level (including blank answers).

To be used for items scored with weighted scoring, Mantel (1963) proposed a more comprehensive formula based on Mantel-Haenszel technique but obtained by extending the contingency table in terms of number of categories. Related formula is presented in Equation 6 (cited in: Gonzales et. al., 2010).

$$MANTEL = \frac{\left[\sum_{j=1}^{k} F_j - \sum_{j=1}^{k} E\left(F_j\right)\right]^2}{\sum_{j=1}^{k} Var\left(F_j\right)} \qquad F_j = \sum_{c=1}^{c} R_c N_{Fcj} \qquad (6)$$

F_j : Score of the reference group at j score level for the related item,

$E(F_j)$: Expected score of the reference group'‘s at j score level for the related item,

j : Paired score levels for both groups,

K : Number of paired scores for both groups.

Likelihood Ratio Test Technique (LRT): This technique examines whether there are significant differences between item parameters predicted separately for focus and reference groups. With this purpose, a limited model that assumes equal item parameters and an extended model that assumes the inequality of i. item parameter and equality of other parameters are formed and the difference between probability value algorithms of these two models is calculated. This value indicates χ^2 distribution (degree of freedom is equal to the number of parameters of the model that used).In cases where obtained value is higher than the table value, non-existence hypothesis is rejected and the related item is said to present DIF existence (Thissen, 2001).

Above mentioned DIF methods based on CTT and IRT have been developed in time with the help of the advances in computer usage and developments in theories. However, it is known

that all these techniques have strong and weak aspects and many of the techniques have been developed to overcome the weak aspects of the others. Many studies are included in the literature to present the operations of different DIF identification methods based on different variables. These studies show that findings obtained by different DIF identification methods are affected from variables such as number of items with DIF, test length, level of DIF, sample size, type of DIF and DIF detection method (Camilli & Shepard, 1994; Gelin & Zumbo, 2003; Gierl, Jodoin & Ackerman, 2000; Kim & Cohen, 1991; Narayanan & Swaminathan, 1994; Osterlind, 1983; Padilla et. al., 2012).

It is believed that scoring of the items is another variable that can change the findings obtained by DIF identification techniques. As a matter of fact, it is not possible for social sciences to expect equality of all items and options in psychological space. Studies also show that dichotomous scoring is insufficient to represent partial information of the individuals and weighted scoring better represents that information and provides more and reliable information about a larger range of skills dimension (Embretson & Reise, 2000; Ostini & Nering, 2006). Samejima's (1975, 1979) study also points to the fact that items scored by using weighted scoring provides more statistical information compared to items scored with dichotomous scoring (cited in: Ostini & Nering, 2006).

Therefore, this study intended to present the operations of widely used and easy to find software for application CTT and IRT based Mantel Haenzsel/Mantel, Standardization and Likelihood Ratio Test techniques in commonly used dichotomous, consensus based weighted (CBW) and emprical weighted (EW) scoring situations. Although there are many studies in literature about the variables such as different sample sizes, distribution of skills, number of items/ratio of items with DIF, level of DIF and numbers and expressions in categories used in item scoring, none of the accessed studies examined different scoring methods (such as dichotomous and weighted) in terms of DIF. Similarly, in Turkey, it has been observed that studies on DIF identification techniques generally focus on techniques used in the case of dichotomous scoring rather than polytomous scoring (Acar, 2008; Ateşok Deveci, 2008; Ayan,2011; Çepni, 2011; Doğan & Öğretmen, 2008; Gök et. al., 2010; Kalaycıoğlu, 2008; Kan, 2007; Kan, Sünbül & Ömür, 2013; Korkmaz, 2005; Kurnaz, 2006; Öğretmen, 1995; Öğretmen, 2006; Öğretmen, 2009; Yenal, 1995; Yurdugül, 2003; Yurdugül, 2010).

Hence, it is believed that a comprehensive study that includes different DIF identification techniques based on different theories that will be used for different scoring situations will extensively contribute to the literature in Turkey. Research findings are also believed to contribute to studies to ensure validity of measurement tools in the context of understanding the effects of different scoring situations that are used and to improve the test development process. These aspects reflecs the significance of this study.

In line with the purpose of the study, answers were sought to the questions listed below:

1. How are items with DIF distributed for university variable when they are developed based on Classical Test Theory and identified by using M-H/Mantel and Standardization techniques,

   a. In case of dichotomous scoring,

   b. In case of consensus based weighted scoring,

   c. In case of emprical weighted scoring?

2. How are items with DIF distributed for university variable when they are developed based on Item Response Theory and identified by using Likelihood Ratio Test technique,

   a. In case of dichotomous scoring,
   b. In case of consensus based weighted scoring,
   c. In case of emprical weighted scoring?

3. Are there differences between items with DIF identified with DIF identification techniques based on Classical Test Theory and Item Response Theory in cases of dichotomous scoring, consensus based weighted scoring and emprical weighted scoring?

   a. Do the items with DIF obtained with M-H/Mantel, Standardization and LRT techniques under different scoring situations differ?
   b. Do the items identified to include DIF by using different techniques differ in cases of dichotomous scoring, consensus based weighted scoring and emprical weighted scoring?
   c. What are the agreement percentages of items with DIF identified with the help of DIF identification techniques based on Classical Test Theory and Item Response Theory in cases of dichotomous scoring, consensus based weighted scoring and emprical weighted scoring?

4. Are there differences between the number of DIF items identified by DIF identification techniques based on based on Classical Test Theory and Item Response Theory in cases of dichotomous scoring, consensus based weighted scoring and emprical weighted scoring?

This study is limited to; M-H/Mantel, Standardization and LRT techniques, Verbal Reasoning Skills Testandthe data obtained by applying this test, dichotomous, consensus based and emprical weighted scoringconditionsandtheaccuracy of the softwares used in the calculations.

2. METHOD

2.1. Study Group

This study was conducted by using the data obtained from the single session Verbal Reasoning Skills Test administered to 1593 students attending various faculties and departments of Hacettepe (H.U) and Gazi (G.U) Universities in Ankara central province. Research data was obtained by Gözen Çıtak (2007) and used with the permission of Gözen Çıtak (2007). Since the study intended to investigate the operations of DIF techniques under different scoring situations instead of bias and sources of bias and Gözen Çıtak's (2007) research data were only considered in terms of university variable, therefore focus group was selected as Hacettepe University students whereas the reference groups was identified as Gazi University students. Table 1 presents information regarding the group that provided research data.

Table 1. Some Descriptive Information about the Group from Whom Research Data were Collected.

Variable	University		Gender	
	Hacettepe	Gazi	Male	Female
f	631	962	288	591
%	39.6	61.4	32.7	67.3
Total	1593		879	

2.2. Data Collection

Since the study made use of data collected by Gözen Çıtak's (2007) Verbal Reasoning Skills Test, this section provides brief information about the test. Verbal Reasoning Skills Test is preferred because it involves different scoring situations for the same items and because it overlaps with research problems of this research. Verbal Reasoning Skills Test was composed of 18 multiple choice items each of which offers four options with varying levels of accuracy and the items can be scored as weighted (it can be scored from 1 to 4). Score '4' implies the most correct answer whereas score '1' points to a correct yet farthest statement from the most accurate answer. These weights were identified from the means obtained from response option-total biserialcorrelation coefficients for EW scoring and from the weights provided by a ten-person expert group selected according to proximity of options to the correct answer for CBW scoring. Reliability of the test was calculated as 0.64 (KR-20) for '1-0' scoring, 0.68 (Cronbach α) for CBW scoring and 0.69 (Cronbach α) for EW scoring. t-test was used to examine the differences between the means of test scores for students who are registered to programs that receive students from verbal and quantitative sections to obtain proof for the validity of the test and 0.01 level of significance was obtained among means for all three scoring methods. Pearson product-moment correlation coefficient calculated for Turkish Verbal Expressions class scores for all three scoring methods was found to be 0.55 for '1-0' scoring; 0.52 for CBW scoring and 0.52 for EW scoring (Gözen Çıtak, 2007).

2.3. Analysis of Data

DIF analyses of the data used in the framework of the study were undertaken with the help of M-H/Mantel, Standardization and LRT techniques. As it is known, M-H/Mantel and Standardization techniques include assumptions that need not to be satisfied by parametric techniques. However, since LRT is one of the techniques based on IRT, the data should meet IRT's basic assumptions such as the unidimensionality and local independence (Embretson & Reise, 2000; Hambleton & Swaminathan, 1989; Ostini & Nering, 2006). Therefore, the first phase of data analysis included controlling whether these assumptions were met. Gözen Çıtak (2007) examined unidimensionality assumption, one of the main assumptions of IRT, by using principal components analysis based on inter-items tetrachoric correlation matrix and identified that 17 of the 18 items in the test were combined under the first component. This finding shows that data is unidimensional. Regarding the local independence, in the literature it is said that this assumption is linked to the unidimensionality and a data that is seen to be unidimensional meets also the local independence (Hambleton & Swaminathan, 1989: 25; Lord, 1980: 19). Based on these it is deemed that the study data also meets the local independence.

Second phase of data analysis examined model-data agreement to allow analyses based on IRT. -2 log likelihood value of the data as 13858 for H.U, 22127 for G.U for two parameter logistic model and as 61159.35 for weighted response model. Since the statistic with χ^2distribution was rather sensitive towards sample size and model-data agreement cannot be met for each model in big samples, *-2 log likelihood / (S-1) – 2n(r-1) ≤ 3.00* condition was taken into consideration for model-data fit.

-2 log likelihood value is also used as a measure of model-data fit in MULTILOG (Thissen, 1991) and it is shown in the literature that the compliance values providing the *-2 log likelihood / (S-1) – 2n(r-1) ≤ 3.00* condition can be regarded as a sign of a satisfactory fit forthe model (Bock, 1997; Drasgow, Levine, Tsien, Williams & Mead, 1995). Here, '*S'* is the number of response

patterns, *n* is the number of items and *r* is the response category. The probable number of response pattern in this study is 4^{18} based on the number of items and the response category. Bock (1997) reported that agreement values that meet the *-2 log likelihood / (S-1) - 2n(r-1) ≤ 3.00* condition was sufficient for model-data agreement (cited in: Gözen Çıtak, 2007). In line with these findings, data was found to be in agreement with two parameter logistic model for '1-0' scoring and with weighted scoring for Graded Response Model (GRM).

Easy-DIF software was used in the study for analyses based on M-H/Mantel and Standardization techniques. IRTLRDIF software was utilized for analyses based on LRT technique for two parameter logistic model for dichotomous scoring and GRM for weighted scoring.

To answer the first and second sub problems of the study, $_{MH\chi^2}$ values were calculated for M-H/Mantel technique and level of significance was examined for these values. (In this section Easy-DIF software automatically uses total score in the reference and focus group matches). Items that provide significant values were accepted to include DIF. SPD and SMD values were calculated for standardization technique and cutoff scores proposed by Gonzales et. al. (2010) were taken into consideration when determining the existence of DIF in the item. G^2 values were calculated for LRT technique. In the calculation of G^2 values, item which tested for DIF was taken as candidate item and other items as anchor items.

Since these values presented χ^2 distribution in degree of freedom that is consistent with the number of predicted parameters, critical values of χ^2 distribution here was accepted as (2 PLM) 5.99 ($\alpha = 0.05$, df=2) for dichotomous scoring, (GRM) 7.81 ($\alpha = 0.05$, df=3) for CBW and 9.48 ($\alpha = 0.05$, df=4) for EW scoring (Dişçi, 2012). In order to answer the third sub problem in the study, Cohran''s Q and McNemar tests were used to determine whether DIF items identified through various techniques presented significant differences.

The McNemar test is used to determine if there are differences on a dichotomous dependent variable between two related groups. And the Cochran's Q test is used to determine if there are differences on a dichotomous dependent variable between more than two related groups.

In this study Cohran's Q test was used to test whether the DIF items determined by M-H / Mantel, Standardization and LRT techniques differ significantly from each other at different scoring situations and McNemar test was used to investigate which techniques originated the difference if a significant difference was found with Cohran's Q test. Agreement between DIF items identified through various techniques and different scoring situations was investigated with simple agreement coefficient (percentage). Simple agreement coefficient (percentage) was obtained by proportioning the item numbers in agreement to total number of items (Erkuş, 2006). In order to analyses data related to the fourth sub problem n the study, chi-square test was used to identify the significance of differences between total number of items with DIF.

3. RESULTS

3.1. Results Related To the First and Second Sub Question

In order to find answers to the first sub problem, data for dichotomous, CBW and EW scoring were analyzed with the help of M-H/Mantel and Standardization techniques and the obtained information is presented in Table 2.

Table 2. Values Obtained by Different DIF Identification Techniques Based on Different Theories under Different Scoring Situations

Items	Classical Test Theory								Item Response Theory		
	M-H/Mantel					Std.			LRT		
	Dic. (*p*)	CBW (*p*)	EW (*p*)	MH D-DIF***	ETS Class. ****	Dic. (*SPD*)	CBW (*SMD*)	EW (*SMD*	Dic. (G^2)	CBW (G^2)	EW (G^2)
1	0.62	0.11	0.33	0,16	A	0.004	**0.061**	0.018	0.62	8.24	6.04
2	0.74	**0.01**	**0.00**	0,11	A	0.008	**0.095**	**0.091**	0.52	**23.54**	**24.44**
3	**0.01**	0.09	0.07	1,33	B	**-0.052**	**-0.058**	**-0.061**	**8.92**	6.74	5.14
4	0.22	**0.00**	**0.00**	0,42	A	0.021	**0.079**	**0.096**	5.82	**25.34**	**26.14**
5	0.36	0.48	0.65	0,25	A	0.020	0.002	-0.004	1.82	3.54	3.84
6	0.44	0.52	0.46	0,21	A	-0.018	-0.029	-0.036	1.92	3.44	2.94
7	0.91	0.33	0.37	0,02	A	-0.002	-0.048	**-0.055**	0.62	3.24	3.44
8	0.52	0.70	0.72	0,18	A	-0.005	0.011	0.016	3.32	7.84	7.84
9	0.34	0.16	0.07	0,25	A	0.034	-0.038	-0.050	5.42	**10.44**	**11.44**
10	**0.00**	**0.00**	**0.00**	0,75	A	**-0.077**	**-0.139**	**-0.141**	**15.12**	**17.14**	**17.14 ***
11	0.20	**0.00**	**0.00**	0,37	A	0.031	**0.109**	**0.069**	2.92	**32.74**	**32.54**
12	0.11	**0.01**	**0.01**	0,44	A	0.040	**0.106**	**0.112**	3.82	**13.34**	**13.54**
13	0.63	0.69	0.71	0,16	A	0.008	0.014	0.019	**13.02***	**13.04 ***	**13.64 ***
14	0.98	0.65	0.52	0,04	A	0.009	-0.001	-0.007	2.52	**13.64 ***	**13.14 ***
15	0.05	0.59	0.20	0,56	A	0.040	0.013	0.038	4.02	7.74	9.14
16	0.18	0.72	0.09	0,42	A	0.020	0.024	**0.071**	2.32	6.54	**25.54**
17	0.99	0.95	0.55	0,02	A	-0.009	-0.004	0.015	2.32	3.54	8.14
18	**0.00**	**0.00**	**0.00**	1,01	B	**-0.079**	**-0.198**	**-0.190**	**23.92 ***	**25.14**	**21.64**
	3**	6**	6**			3**	8**	9**	4**	9**	10**

*: Non-Uniform DIF, **: Total number of items with DIF, ***: $MH\ D\text{-}DIF = -2.35\ln(\hat{\alpha}_{MH})$ (This values are calculated for dic. Scored items to give an idea of the effect size), ****: ETS classification for designate items as A (negligible or non-significant DIF), B (slight to moderate DIF), or C (moderate to large DIF) (Zwick, 2012).

Examination of Table 2 shows the existence of significant DIF based on M-H/Mantel for dichotomous scoring situation in items 3, 10 and 18, existence of DIF that cannot be negligible based on standardization technique in items 3, 10 and 18. According to this information, a total of 3 items points to the existence of DIF based on M-H/Mantel technique, total of 3 items points to the existence of DIF based on standardization. In case of CBW, a total of 6 items points to the existence of DIF based on M-H/Mantel technique, total of 8 items points to the existence of DIF based on standardization. In case of EW, a total of 6 items points to the existence of DIF based on M-H/Mantel technique, total of 9 items points to the existence of DIF based on standardization technique.

In order to find answers to the second sub problem, data for dichotomous, CBW and EW scoring were analysed with the help of LRT technique and the obtained information is presented in Table 2. According to Table 2, examination of G^2 values obtained with LRT technique under dichotomous scoring conditions shows that items 3, 10, 13 and 18 pointed to DIF that cannot be negligible. Also, Non-Uniform DIF existence was observed in items 13 and 18. Examination of G^2values obtained by LRT technique under CBW scoring situation points to the fact that items 2, 4, 9, 10, 11, 12, 13, 14 and 18 included DIF existence of DIF that cannot be negligible. Also, Non-Uniform DIF existence was observed in items 13 and 14. Examination of G^2values obtained by

LRT technique under EW scoring situation points to the fact items 2, 4, 9, 10, 11, 12, 13, 14, 16 and 18 included DIF existence that cannot be negligible. Additionally, Non-Uniform DIF existence was observed in items 10, 13 and 14.

3.2. Results Related to the Third Sub Question

In order to find answers to the sub problem 3a, Cohran's Q test was used to determine whether there were differences between DIF identification techniques based on CTT and IRT in cases of dichotomous, CBW and EW scoring situations. The obtained information is presented in Table 3.

Table 3. Cohran's Q Test Results for Regarding the Differences between DIF Identification Techniques Based on CTT and IRT in Cases of Dichotomous, CBW and EW Scoring Situations.

Scoring Situations	M-H/Mantel, SPD/SMD and LRT
	Cohran's Q
Dichotomous	2.00
CBW	2.80
EW	4.33

Table 3 shows no significant differences between techniques in cases of dichotomous scoring ($p=0.37$, p>0.05); in cases of CBW ($p=0.24$, p>0.05); and EW ($p=0.11$, p>0.05).

In order to find answers to the sub problem 3b, Cohran's Q and McNemar tests were used to determine whether there were differences between DIF identification techniques based on CTT and IRT in cases of dichotomous, CBW and EW scoring situations. Since no items pointed to the existence of DIF. The obtained information is presented in Table 4.

Table 4 shows no significant differences between M-H/Mantel technique and scoring situations ($p=0.16$, p>0.05), significant differences between standardization technique and scoring situations ($p=0.01$, p<0.05) and significant differences between LRT technique and scoring situation ($p=0.02$, p<0.05).

Table 4. Cohran's Q and McNemar Test Results for Regarding the Differences between DIF Identification Techniques Based on CTT And IRT In Cases of Dichotomous, CBW and EW Scoring Situations.

DIF Procedures	Dichotomous CBW-EW		Dichotomous CBW	Dichotomous EW	CBW-EW
	Cohran's Q	p	McNemar p	McNemar p	McNemar p
M-H/Mantel	3.60	0.16	-	-	-
SPD/SMD	8.85	0.01	0.06	0.03	1.00
LRT	7.75	0.02	0.12	0.07	1.00

McNemar test was conducted to determine which scoring situation generated this difference and scoring situations were compared with each other. According to this analysis, there were no differences between dichotomous and CBW scoring situations (p=0.06, p>0.05); there were significant differences between dichotomous and EW scoring situations (p=0.03, p<0.05) and there were no differences between CBW and EW scoring situations (p=1.00, p>0.05) when

standardization technique was used. There were no differences between dichotomous and CBW scoring situations (p=0.12, p>0.05), there were significant differences between dichotomous and EW scoring situations (p=0.07, p>0.05) and there were no differences between CBW and EW scoring situations (p=1.00, p>0.05) when LRT technique was used.

In order to find answers to the sub problem 3c, simple agreement coefficients (percentages) were calculated for items with DIF identified with DIF identification techniques based on CTT and IRT under dichotomous, CBW and EW scoring conditions. The obtained information is presented in Table 5.

Table 5. Agreement Coefficients (Percentages) for DIF Items Identified with DIF Identification Techniques Based on CTT And IRT in Cases of Dichotomous, CBW and EW Scoring Situations.

Concordance Between Procedures	**M-H/Mantel and Std.**	**M-H/Mantel and LRT**	**Std. and LRT**	**M-H/Mantel, Std. and LRT**
Dichotomous	%100	%94	%94	%94
CBW	%89	%83	%78	%72
EW	%83	%78	%78	%67

Concordance Between Scoring Situations	**Dichotomous-CBW**	**Dichotomous-EW**	**CBW-EW**	**Dichotomous-CBW-EW**
M-H/Mantel	%72	%78	%100	%72
Standardization	%72	%67	%83	%61
LRT	%61	%56	%94	%56

According to Table 5, the highest agreement in dichotomous scoring situation was observed between M-H/Mantel and SPD techniques (100%) and the agreement in all other techniques was lower than that of agreement between M-H/Mantel and SPD techniques but was the same (94%). The highest agreement in CBW scoring situation was observed between M-H/Mantel and SPD techniques (89%) but this agreement was lower than the agreement obtained in cases of dichotomous scoring. It was also observed that agreement between M-H/Mantel and LRT, SPD and LRT and agreement among all three techniques was lower compared to agreement observed in the dichotomous case. The highest agreement in EW scoring situation was observed between M-H/Mantel and SPD techniques (83%) but with a general decrease in agreement between techniques compared to other scoring situations.

Agreement between scoring in Table 5 shows that the highest agreement was observed between CBW and EW scoring situations in M-H/Mantel technique (100%) and that agreement between dichotomous and CBW (72%) and dichotomous and EW (78%) was lower than that of CBW and EW scoring situations. The highest agreement was observed between CBW and EW scoring situations in Standardization technique (83%). Agreement between dichotomous and CBW (72%) and dichotomous and EW (67%) scoring situations was lower than the agreement between CBW and EW scoring situations and also agreement obtained with Standardization technique from different scoring situations was higher than the agreement obtained with M-H/Mantel technique. The highest agreement was observed between CBW and EW scoring situations in LRT technique (94%) and it was found that agreement between dichotomous and CBW (61%) and dichotomous and EW (56%) was lower than that of CBW and EW scoring situations and that the lowest agreement rate for dichotomous and weighted scoring situations was obtained in LRT technique.

3.3. Results Related to the Fourth Sub Question

Table 2 presents the items with DIF identified with DIF identification techniques based on CTT and IRT under dichotomous and weighted scoring situations.

Fisher''s Exact χ^2 test was conducted on the obtained data to find answers the related sub problem and it was found that DIF identification techniques that were used and scoring situations were independent of each other (p=0.55, p>0.05). Number of cells with expected values below 5 were calculated to undertake Fisher''s Exact χ^2 test and since this number is higher than the 20% of the total number of cells (33%), it was decided to combine the cells among themselves based on DIF identification techniques using CTT (M-H/Mantel, SPD/SMD) and weighted scoring situations. During this integration, it was ensured that items with DIF were counted only once for the three techniques and for both weighted scoring situations, in other words, clusters were integrated.

4. CONCLUSION

This study intended to investigate the operations of widely used M-H/Mantel, Standardization and Likelihood ratio tests based on CTT and IRT under dichotomous, CBW and EW scoring conditions. Although there were no significant differences between techniques for items with DIF under dichotomous scoring conditions, it was observed that techniques based on CTT generated more agreement among themselves compared to techniques based on IRT. The most important cause for this observation may be related to calculating the MH/Mantel and Standardization techniques by using the contingency tables and that they are based on the same theory. The fact that LRT technique was able to identify DIF in one more item in addition to MH/Mantel and Standardization techniques and the fact that this addition all item presented non-uniform DIF gives the impression that LRT technique is more sensitive to identify non-uniform DIF cases. It can be claimed that these findings are be parallel to the findings obtained by Atalayet. al. (2012).

Literature presents various studies that point to the fact that different techniques provide different results but also techniques based on CTT and IRT show more agreement among themselves (Abedlazeez, 2010; Doğan & Öğretmen, 2008; Kan, Sünbül & Ömür ,2013; Spray & Miller, 1994; Ward & Bennett, 2012). Results obtained from the current study are consistent with these studies in general.

Standardization technique was able to identify existence of DIF in two more items with the use of CBW in addition to M-H/Mantel technique and that very similar results were obtained by Standardization and LRT techniques in terms of total items with DIF creates the view at first that agreement between Standardization and LRT is higher than the agreement between M-H/Mantel and LRT. However, identification of agreement between M-H/Mantel and LRT techniques as 83%, agreement between Standardization and LRT techniques as 78% and agreement between M-H/Mantel and Standardization techniques as 89% shows that agreement between M-H/Mantel and LRTtechniques was higher than the agreement between Standardization and LRT techniques in terms of items identified to include DIF. This results shows that M-H/Mantel technique is more in agreement with LRT technique. This result may be related to the odds ratio of M-H/Mantel technique and calculating the Standardization technique through "*p"* values. As a matter of fact, literature mentions that using only *"p"* values as an indicator of DIF is problematic for DIF findings since *"p"* values are affected by mean group differences (real difference as proof of

validity) and by item discrimination (r_{jx}) (Angoff, 1982; Hunter, 1975; Lord, 1977, cited in: Camilli and Shepard, 1994).

Although findings obtained for dichotomous scoring conditions and similarly for CBW scoring conditions did not point to significant differences in terms of techniques that were used, it was observed that techniques based on CTT provided more agreement compared to techniques based on IRT. As was the case in dichotomous scoring situations, here the reason was believed to be generated from using contingency table to calculate M-H/Mantel and Standardization techniques and the fact that they were based on the same theory.

Identification of more DIF items with LRT technique compared to M-H/Mantel and Standardization techniques and that these items presented non-uniform DIF gives the impression that LRT technique can be more sensitive in this regard under CBW scoring conditions.

Compared to dichotomous scoring conditions, CBW scoring conditions provided increases in items with DIF which was more than double compared to items identified by all techniques. For instance, existence of identified DIF was 3 in dichotomous scoring condition for M-H/Mantel technique, 6 in CBW scoring condition, 3 in dichotomous scoring condition for Standardization technique, 8 in CBW scoring condition, 4 in dichotomous scoring condition for LRT technique and 9 in CBW scoring condition. Results of Cohran''s Q test presented significant differences for the items with DIF for Standardization and LRT techniques under dichotomous and CBW scoring conditions. This difference was not significant for M-H/Mantel technique. These findings are parallel to the findings obtained by Atalayet. al. (2012).

Findings display that DIF findings obtained for the same test with different DIF identification techniques under dichotomous and CBW scoring conditions may change. It is believed that this change is related to the fact that weighted scoring of items represent partial information of the individuals better, it provides information about a larger range of skills dimension and provides more reliable statistical information (Embretson & Reise, 2000; Ostini & Nering, 2006).

Obtaining similar findings in EW scoring situation to CBW scoring situation shows that techniques based on CTT show more agreement among themselves in all conditions (dichotomous, EW and CBW) compared to IRT based techniques and LRT technique identified existence of DIF in more items. Findings also show that agreement between DIF identification techniques deteriorate when scoring is moved form dichotomous conditions to CBW and EW scoring conditions. Agreement between items identified for CBW and EW scoring conditions were rather high for all techniques and no significant differences were observed for between these scoring conditions. Significant differences were observed between dichotomous and CBW scoring conditions for items with DIF for standardization technique. These findings show that weighted scoring conditions can be more consistent with each other compared to dichotomous scoring conditions.

Additionally, it was observed that total number of items identified by techniques based on CTT and IRT under dichotomous and weighted conditions are significantly independent of each other, these findings are consistent with the findings of Abedlazeez (2010).

Findings display that DIF findings obtained for the same items with the same DIF identification techniques for dichotomous and weighted scoring conditions can differ. This finding can show that DIF is not only the function of the item but also the function of the used scoring

method as well. Dichotomous scoring should not generate sufficient variability between individuals for the techniques that are used and is unable to represent real competence levels. These findings correspond with the findings of the study undertaken by Wetzel et. al. (2013) and Gelin & Zumbo (2003).

The highest number of DIF existence was identified by LRT technique in this study for all scoring conditions. It is believed that this result is dependent by the facts that M-H and Standardization techniques have insufficiencies in identifying non-uniform DIF existence, general advantages provided by IRT and adding the slope parameters in the calculation in LRT technique. Similarly, Embretssonand & Reise (2000) also reported that IRT is more advantageous than techniques based on CCT in tracing items with DIF in general.

5. REFERENCES

Abedlazeez, N. (2010). Exploring DIF: comparison of CTT and IRT methods. *International Journal of Sustainable Development*, 1(7), 11-46.

Acar, T. (2008). *Maddenin Farklı Fonksiyonlaşmasını belirlemede kullanılan Genelleştirilmiş Aşamalı Doğrusal Modelleme, Lojistik Regresyon ve Olabilirlik Oranı tekniklerinin karşılaştırılması.* Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı. Yayımlanmamış Doktora Tezi.

Atalay, K., Gök, B., Kelecioğlu, H., Arslan, N. (2012). Değişen Madde Fonksiyonunun belirlenmesinde kullanılan farklı yöntemlerin karşılaştırılması: bir simülasyon çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*. 43: 270-281.

Aiken, L. R. (2000). *Psychological testing and assessment.*Allyn and Bacon.

Anastasi, A., Urbina, S. (1997). *Psychological testing.* Prentice Hall.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. Holland, P.W. and Wainer, H. (Ed.), *Differential Item Functioning*. Lawrence Erlbaum Associates, Publishers, New Jersey.

Ateşok Deveci, N. (2008). *Üniversitelerarası Kurul Yabancı Dil Sınavı'nın Madde Yanlılığı bakımından incelenmesi.* Ankara Üniversitesi Eğitim Bilimleri Enstitüsü Eğitimde Psikolojik Hizmetler Anabilim Dalı. Yayınlanmamış Doktora Tezi.

Ayan, C. (2011). *PISA 2009 fen okuryazarlığı alt testinin değişen madde fonksiyonu açısından incelenmesi.* Hacettepe Üniversitesi Sosyal Bilimler Ensitüsü Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı. Yüksek Lisans Tezi.

Bock, R.D. (1997). The Nominal Categories Model. In W. J. van der Linden, and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (p.33-49). New York Inc.: Springer-Verlag.

Camilli, G., Shepard, L.A. (1994). *Methods for identifying biased test items.* Sage Publication, London.

Crocker, L., Algina, J. (1986). *Introduction to classical and modern test theory*. USA: Rinehart and Winston Inc.

Çepni, Z. (2011). *Değişen Madde Fonksiyonlarının SIBTEST, Mantel Haenzsel, Lojistik Regresyon ve Madde Tepki Kuramı Yöntemleriyle İncelenmesi.* Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı. Yayımlanmamış Doktora Tezi.

Drasgow, F., Levine, M.V., Tsien, S., Williams, B., Mead, A.D. (1995). Fitting Polytomous Item Response Thoery Models to Multiple-Choice Tests. *Applied Psychological Measurement*, 19 (2), 143-165.

Dişçi, R. (2012). *Temel ve Klinik Biyoistatistik.* İstanbul Tıp Kitapevi.

Doğan, N., Öğretmen, T. (2008). Değişen Madde Fonksiyonunu belirlemede Mantel–Haenszel, Ki-Kare ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, 33(148).

Dorans, N.J., Holland, P.W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. Holland, P.W. & Wainer (Ed.), *Differential Item Functioning*. Lawrence Erlbaum Associates, Publishers, New Jersey.

Embretson, S.E., Reise, S.P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Publishers, London.

Erkuş, A. (2006). *Sınıf Öğretmenleri İçin Ölçme ve Değerlendirme: Kavramlar ve Uygulamalar*, Ekinoks, Ankara.

Gelin, M.N., Zumbo, B.D. (2003). Differential item functioning results may change depending on how an item is scored: an illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement, 63:65,* DOI: 10.1177/0013164402239317.

Gierl, M.J., Jodoin, M.G., Ackerman, T.A. (2000). Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression When the Proportion of DIF Items is Large. Annual Meeting of the American Educational Research Association (AERA).

Gonzales, A., Padilla, J.L., Dolores, H., Gomez-Benito, J., Benitez, I. (2010). EASY-DIF: Software for analyzing Differential Item Functioning using the Mantel-Haenszel and Standardization procedures. *Applied Psychological Measurement*. doi:10.1177/0146621610381489.

Gök, B., Kelecioğlu, H., Doğan, N. (2010). Değişen Madde Fonksiyonunu belirlemede Mantel–Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, 35(156).

Gözen Çıtak, G. (2007). *Klasik Test Ve Madde-Tepki Kuramlarına Göre Çoktan Seçmeli Testlerde Farklı Puanlama Yöntemlerinin Karşılaştırılması.* Ankara Üniversitesi Eğitim Bilimleri Enstitüsü Eğitimde Psikolojik Hizmetler Anabilim Dalı. Yayınlanmamış Doktora Tezi.

Hambleton, R.K. and Swaminathan, H. (1989). *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff Publishing, Boston.

Kalaycıoğlu, D.B. (2008). *Öğrenci Seçme Sınavı'nın Madde Yanlılığı açısından incelenmesi.* Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı. Yayımlanmamış Doktora Tezi.

Kan, A. (2007). Test Yansızlığı: H.Ü. yabancı dil muafiyet sınavının cinsiyete ve bölümlere göre dmf analizi. *Eurasian Journal of Educational Research*, 29, 45-58.

Kan, A., Sünbül, Ö., Ömür, S. (2013). 6.- 8. sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, 9(2), 207-222.

Korkmaz, M. (2005). *Madde Cevap Kuramına Dayalı Olarak Çok Kategorili Maddelerde Madde ve Test Yanlılığının (İşlevsel Farklılığın) İncelenmesi.* Ege Üniversitesi. Yayınlanmamış Doktora Tezi.

Kim, S.-H., Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, 15, 269-278.

Kurnaz, F.B. (2006). *Peabody Resim Kelime Testinin Madde Yanlılığı açısından incelenmesi.* Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı. Yayınlanmamış Yüksek Lisans Tezi.

Lord, F. M. (1980). *Aplications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Magnusson, D. (1967). *Test Theory.* Massachusetts. Addison-Wesley.

Murphy, K.R., Davidshofer, C.O. (2005). *Psychological Testing: principles and applications.* Pearson/Prentice Hall.

Narayanan, P., Swaminathan, H. (1994). Performance of the Mantel-Haenszel and Simultaneous Item Bias procedures for detecting Differential Item Functioning, *Applied Psychological Measurement*, 18(4).

Padilla, J.L., Hidalgo, J.L., Benitez, I., Gomez-Benito, J. (2012). Comparison of three software programs for evaluating DIF by means of the Mantel-Haenszel procedure; EASY DIF, DIFAS and EZDIF, *Psicologica*, 33,135-156.

Osterlind, S.J. (1983). *Test Item Bias*. Sage Publication, London.

Ostini, R. Nering, M.L. (2006). *Polytomous Item Response Theory Models*. Sage Publications, Thousand Oaks, California.

Öğretmen, T. (1995). *Differential Item Functioning analysis of the verbal ablity section of the first stage of the University Entrance Examination in Turkey*. Ortadoğu Teknik Üniversitesi Sosyal Bilimler Enstitüsü. Yüksek Lisans Tezi.

Öğretmen, T. (2006). *Uluslararası okuma becerilerinde gelişim projesi (PIRLS) 2001 testinin psikometrik özelliklerinin incelenmesi: Türkiye-Amerika Birleşik Devletleri örneği*. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı. Doktora Tezi.

Öğretmen, T. (2009). Değişen Madde ve Test Fonksiyonunun belirlenmesinde Madde Tepki Kuramı'na dayalı parametrik yöntemlerin karşılaştırılması. *Eğitim ve Bilim*, 34(152), 113-125.

Santelices, M.V., Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: an issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement 72(1)* 5–36.

Selvi, H. (2013). *Klasik test ve madde tepki kuramlarına dayalı değişen madde fonksiyonu belirleme tekniklerinin farklı puanlama durumlarında incelenmesi*. Mersin Üniversitesi Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Anabilim Dalı. Yayınlanmamış Doktora Tezi, Mersin.

Selvi, H., Alici, D. (2016). *Investigating Differential Item Functioning Methods under Different Missing Value Reduction Methods Situations.* 5. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Antalya

Spray, J.,, Miller, T. (1994). *Identifying nonuniform DIF in polytomously scored test items.* American College Testing Research Report Series 94-1. Iowa City, IA: American College Testing Program.

Thorndike, R.L. (1982). *Aplied Psychometrics*. Houghton Mifflin Company, Boston.

Thissen, D.M. (1991). *Multilog User's Guide- Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory*. Chicago: Scientific Software, Inc.

Thissen, D. (2001). *IRTLRDIF (v.2.0b Sürümü)* [Computer Software]. Chapel Hill: L. L. Thurstone Psychometric Labaratory, University of North Carolina at Chapel Hill.

Ward, W.C., Bennett, R.E. (2012). *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment.*Routledge,Taylor ve Francis Group, London and New York.

Wetzel, E., Böhnke, J.R., Carstensen, C.H., Zeigler, M., Ostendorf, F. (2013). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*, 34(2), 69-81. doi: 10.1027/1614-0001/a000102.

Yenal, E. (1995). *Differantial Item Functioning analysis of the quantitative ability section of the first stage of the University Entrance Examination in Turkey*. Ortadoğu Teknik Üniversitesi Sosyal Bilimler Enstitüsü. Yayınlanmamış Yüksek Lisans Tezi.

Yurdugül, H. (2003). *Ortaöğretim Kurumları Seçme ve Yerleştirme Sınavı''nın madde yanlılığı açısından incelenmesi.* Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı. Yayınlanmamış Doktora Tezi.

Yurdugül, H. (2010). Farklı madde puanlama yöntemlerinin ve farklı test puanlama yöntemlerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1, 1-8.

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.*Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures*: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. Research Report. Educational Testing Service.