Journal of Pedagogical Research
Volume3,Issue 3, 2019
http://dx.doi.org/10.33902/jpr.v3i3.137

Research Article

Determination of differantial item functioning (DIF) according to SIBTEST, Lord's χ^2, Raju's area measurement and Breslow-Day methods

Fatma Gökçen Ayva Yörü[1] and Hakan Yavuz Atar[2]

[1]*Gazi University, Faculty of Education, Ankara, Turkey (ORCID: 0000-0002-4555-1987)*
[2]*Gazi University, Faculty of Education, Ankara, Turkey (ORCID: 0000-0001-5372-1926)*

The aim of this study is to examine whether the items in the mathematics subtest of the Centralized High School Entrance Placement Test [HSEPT] administered in 2012 by the Ministry of National Education in Turkey show DIF according to gender and type of school. For this purpose, SIBTEST, Breslow-Day, Lord's $\chi^2$and Raju's area measurement methods were used to determine the DIF of the 20 items in the mathematics subtest of HSEPT in 2012, and it was determined whether the items show DIF according to these methods or not. The research was conducted on the basis of the data obtained from HSEPT that the eighth grade students took in 2012. After the missing data were removed from the data set, DIF analyses were performed for the mathematics subtest of 1,063,570 (n_{female}=523,939 and n_{male}=539,631; $n_{state\ school}$=1,025,979 and $n_{private\ school}$ =37,591) in total. Since it is aimed to examine the current situation, this study is based on a descriptive research design. According to the methods used, the number and DIF levels of the items with DIF differed depending on the variables of gender and type of school. In line with the findings, this research suggests the researchers to use at least two methods to determine the DIF.

Keywords: Differential Item Functioning; SIBTEST; Breslow-Day; Lord's χ^2; Raju's Area Measurement

Article History: Submitted 26 August 2019; Revised 29 November 2019; Published online 8 December 2019

1. Introduction

In Turkey, various examinations (TEOG, KPSS, LGS and YKS etc.) are administered at different grade levels and course areas with the purposes of diagnosis, selection and placement. The accuracy of the decisions made according to the results obtained from the tests used in these examinations directly affects the individuals' lives. Therefore, it is very important that the results obtained from the tests are valid and reliable. The biggest concern in test development and implementation is to prove that the use of the scores obtained from the test and the comments made are valid (Bachman, 1990).

It is expected that students who have the same ability level in terms of the structure measured through the assessment tools in education will receive a similar score from a certain item. In addition, the test should be fair for different participants. DIF is one of the statistical approaches proposed to determine whether test items function differently between the sub-groups taking the

Address of Corresponding Author

Fatma Gökçen Ayva Yörü, Gazi University, Gazi Faculty of Education, Measurement and Evaluation in Education Department, 06500, Ankara / Turkey.

korelasyon_@hotmail.com

How to cite: Ayva-Yörü, F. G. & Atar, H. Y. (2019). Determination of differantial item functioning (DIF) according to SIBTEST, Lord's χ^2, Raju's area measurement and Breslow-Day methods. *Journal of Pedagogical Research, 3*(3), 139-150.

test and to determine the sources of variance (Geranpayeh & Kunnan 2007). DIF occurs when the probability of responding a particular item correctly differs after the individuals in different subgroups are matched at the level of same ability that the item intends to measure (Holland & Wainer, 1993). If an item is identified as having DIF, this is due to a source of variance not related to the structure measured by the test. In other words, due to the grouping factor, groups perform differently in a given item, and this is a source of variance not related to structure (Messick, 1989, 1994).

In DIF determination studies, the performances of at least two groups, focal and reference, are compared (Finch & French, 2013; Karami, 2012). The focal group is called the minority group, and the reference group is called the majority group (Finch & French, 2013). Two types of DIF are identified, uniform and non-uniform. Uniform DIF occurs when one group performs better at all ability levels than the other group. More specifically, almost all members of a group perform better than members of the other group (Karami, 2012). The non-uniform DIF situation arises when the group members' probability of correctly responding to a particular item is not the same throughout their ability levels (Camilli & Shepard, 1994; Zumbo, 1999). In other words, there is an interaction between grouping and ability levels (Karami, 2012).

DIF studies have an important role in assessing the validity of test scores (French & Finch, 2013; French & Finch, 2015). Because, it is accepted that the presence of DIF in the test items may reduce the validity of the test (French & Finch, 2015; Li & Zumbo, 2009). In addition, since studies on DIF have led to the determination of bias, it has received increasing interest in measurement practices in education and psychology over the last two decades (Millsap & Everson, 1993). Extensive research has been conducted to identify DIF and methods have been developed in this direction (Wiberg, 2007). There are also many studies on the development of statistical methods used in determining DIF (Clauser & Mazor, 1998; Hidalgo & Gomez-Benito, 2010; Osterlind & Everson, 2009; Raju, et al., 2009; Steinberg & Thissen, 2006; Zumbo, 2007). Basically, these methods assume that individuals with similar abilities or knowledge will perform similarly in relevant test items (Dorans & Holland, 1993). There are methods on DIF determination such as Mantel-Haenszel statistics (Holland & Thayer, 1988), Logistic Regression (Swaminathan & Rogers, 1990), Simultaneous Item Bias Test (SIBTEST) (Shealy & Stout, 1993), Raju's Area Measurement (Raju, 1988, 1990), Breslow-Day (Breslow & Day) (1980), Standardization (Dorans & Holland, 1993; Dorans and Kulick, 1986), Confirmatory Factor Analysis and multidimensional approaches (Jöreskog & Sörbom, 1989). Because of the advantages and disadvantages of these methods, it is recommended to use more than one method in DIF studies (Camilli & Shepard, 1994; Holland & Wainer, 1993). In this respect, there are many studies in literature on these statistical methods used to determine DIF (Clauser & Mazor, 1998; Hidalgo & Gomez-Benito, 2010; Osterlind & Everson, 2009; Raju, et al., 2009; Steinberg & Thissen, 2006; Zumbo, 2007). In the scope of the research; SIBTEST and Breslow-Day based on the Classical Test Theory and Lord's χ^2and Raju's Area Measurement methods based on Item Response Theory were chosen. In the DIF analyses, these methods based on the CTT match the groups according to the observed scores, whereas those based on the IRT match the groups according to the implicit variable. These methods have been applied in the scope of the research since they use different matching criteria. The explanations for the four DIF determination methods discussed are given below.

1.1. Simultaneous Item Bias Test (SIBTEST) Method

It is a statistical method proposed by Shealy and Stout (1993) to determine DIF in Dichotomous data. In the method, the latent score rather than the observed score is used as the matching criterion (Clauser & Mazor, 1998). Using the true score estimation through the observed scores as matching criteria makes it possible to control Type I error in determining DIF (Gierl, 2005; Shealy & Stout, 1993). The magnitude of DIF is determined by the β_U statistic obtained from the analysis. The β_U statistic is given in equation 1.

$\widehat{\beta u} = \int B_O(\theta)\, f_F(\theta) d\theta$ Equation (1)

The values of $B_O(\theta)$ and $f_F(\theta)$ are respectively the probability that the individuals in the group respond to the item and the density function of the probability to respond to the item correctly.

The criteria proposed by Roussos and Stout (1996) for the interpretation of this magnitude level;
A level: $I\beta_U I < .059$;
B level: $.059 \leq I\beta_U I < .088$ and
C level: $I\beta_U I \geq .088$.

1.2. Lord's χ^2 Method

In this method, the difference between the item parameter values of the focus and reference group is tested (Magis, Beland, Teurlinckx & Boeck, 2010). In this method, variance covariance values of difficulty and discrimination parameters are examined, and the area between the item characteristic curves of the groups is calculated (Hambleton, Swaminathan & Rogers, 1991). The method has been proposed by Lord to determine both uniform and non-uniform DIF (Wiberg, 2007). Lord's χ^2 statistic is given in equation 2.

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

The values of $a_{diff}, b_{diff}\ ve\ c_{diff}$ are item parameter values for the focal group.

1.3. Raju's Area Measurement Method

In this method, the area between the item characteristic curves of the focus and reference group is examined to determine whether the item is DIF or not (Magis, Beland, Teurlinckx & Boeck, 2010). If the area between the item characteristic curves is zero, it indicates that the item is not with DIF. As the area between the curves moves away from zero, bias increases in item (Lord, 1980; Raju, 1988). In the determination of DIF, different methods are used in the calculation of the area between the curves, including marked and unmarked area indices, weighted and unweighted marked and unmarked area indices (Crocker & Algina 1986; Raju & Arenson 2002). The Z statistic for one parameter logistic model is given in Equation 3.

$Z = \frac{b_{JR} - b_{JF}}{\sqrt{\hat{\sigma}^2_{JR} + \hat{\sigma}^2_{JF}}}$ Equation (3)

b_{Ji} and $\hat{\sigma}^2_{Ji}$ represents the item parameter estimates and the standard error values of the estimate, respectively.

The criteria proposed by Wright and Oshima (2015) for the evaluation of NCDIF statistics obtained with Raju's area measurement method analysis are as follows;

A level: NCDIF < .003;
B level: .003 ≤ NCDIF < .008;
C level: NCDIF ≥ .008.

1.4. Breslow-Day Method

This method, developed by Breslow and Day (1980), was proposed to evaluate the homogeneity of the relationship between focal and reference group membership and item responses in the total test score range. In the absence of homogeneity, there is uniform DIF (Aguerri, Galibert, Attorresi & Marañón, 2009). The method has a distribution of χ^2 with a degree of freedom of 1. In addition, Breslow-day method has superior statistical power and Type I error rate compared to other proposed methods (Penfield, 2003). Breslow-Day statistics are given in equality 4 (Aguerri, Galibert, Attorresi & Marañón, 2009):

$$BD = \sum \frac{[A_j - E(A_j)]^2}{var\ (A_j)}$$ Equation (4)

In this study, whether the items in SBS 2012 mathematics subtest showed DIF according to students' gender and school type was examined by Breslow-Day, SIBTEST, Lord's χ^2 and Raju's area measurement methods. When the literature is examined, there are research studies conducted in DIF and bias by using the SBS data in terms of gender and school types in Turkey (Arıkan, Uğurlu, & Atar, 2016; Kan, Sünbül, & Ömür, 2013; Karakaya, 2012; Karakaya & Kutlu, 2012; Kelecioğlu, Karabay, & Karabay, 2014; Terzi & Yakar, 2018; Toprak, & Yakar, 2017; Yıldırım & Büyüköztürk, 2018). Karakaya (2012) conducted the DIF analysis of the items according to gender in the SBS 6th, 7th and 8th grade science and technology and mathematics subtests in 2009 using Mantel-Haenszel method. Then, it was determined that none of the items with DIF were biased as a result of the expert opinion. Karakaya and Kutlu (2012) carried out DIF analyzes by using Logistic Regression and Mantel-Haenszel methods according to gender and school type in 2009 SBSTurkish subtest, and then they conducted bias study based on expert opinion. As a result of the research, they found that only one of the items including DIF by gender and school type was biased according to the gender. Kan, Sünbül and Ömür (2013) used Transformed Item Difficulty, Mantel-Haenszel, Logistic Regression, Lord's χ^2 and Raju's area measurement methods. According to the results of the study, the majority of the items in the sub-tests did not contain DIF in the methods based on the Classical Test Theory; however, the majority of the items in the sub-tests contained DIF in the methods based on the Item Response Theory. Kelecioğlu, Karabay and Karabay (2014) performed DIF analyzes using SIBTEST, Mantel-Haenszel and logistic regression methods. In the study conducted on the 8th grade SBS data in 2009, they examined the existence of DIF in Turkish, mathematics, science and technology and social studies subtests according to school type and gender variables. In addition, they conducted bias study with expert opinions on DIF-containing items according to at least two of the methods used in the study. Arıkan, Uğurlu and Atar (2016) carried out DIF analyzes using Mantel-Haenszel, SIBTEST, MIMIC and Logistic Regression methods and then conducted biased studies based on expert opinion. They carried out whether the items in 8th grade science and technology sub-tests in SBS (2009) showed DIF by gender on the sub-samples of 300, 600, 1200, 2000 participants. As a result of the study, they found that different number of items contain DIF. Toprak and Yakar (2017) conducted a DIF study using Logistic Regression, Mantel-Haenszel, SIBTEST, Likelihood Ratio and Wald istatistic methods. In the study, they determined the existence of DIF in terms of gender in the 8th grade Turkish subtest of SBS (2011). Terzi and Yakar (2018) conducted DIF analysis by gender using Mantel-Haenszel and Logistic Regression methods. Yıldırım and Büyüköztürk (2018) carried out DIF determination and bias studies according to gender and school type by using Mantel-Haenszel and Logistic Regression methods.

When the related researches are examined, it is shown that there are research studies that use SIBTEST and Mantel Haenzsel methods together (Arıkan, Uğurlu & Atar; 2016; Kelecioğlu, Karabay, & Karabay, 2014; Toprak & Yakar, 2017) and Raju's area measurement and Lord's χ^2 methods together (Kan, Sünbül & Ömür; 2013). In this study, DIF analyses were made by using SIBTEST, Breslow-Day, Lord's χ^2 and Raju's area measurement methods according to gender and school type variables. When performing DIF analyses, the groups are matched according to the observed score (total score) in SIBTEST and Breslow-Day methods based on CTT. In the Lord's χ^2 and Raju's area measurement methods based on IRT, DIF analyses are performed by matching the groups according to the implicit variable. Using the true score estimation as matching criterion through observed scores makes it possible to control Type I error in DIF determination (Gierl, 2005; Shealy & Stout, 1993). It is attempted to define the similarities and differences of the methods through the comparison of the results by using these methods together based on CTT and IRT that match observed scores and implicit variable. Research is important in this aspect since there are no studies, which use these four methods together to determine DIF, in the literature. For this reason, it is thought that this research conducted empirical data set will contribute to the literature.

In the research, it is intended to answer the following questions:

1. In the math subtest of SBS (2012), are there any items with DIF by gender in the analyses made with SIBTEST, Lord's χ^2, Raju's area measurement and Breslow-Day methods?
2. In the 2012 SBS math subtest, are there any items with DIF by school type in the analyses made with SIBTEST, Lord's χ^2, Raju's area measurement methods and Breslow-Day methods?

2. Method

2.1. Research Model

In the study, it was examined whether the items in the mathematics subtest in 2012 SBS show DIF by gender and school type according to various methods. Since it is aimed to examine the current situation, this study is based on a descriptive research design.

2.2. Participants

The research was conducted with the eighth grade students who took the SBS exam in 2012. After the missing data were removed from the dataset, DIF analyzes were performed on the basis of the responses of 1,063,570 (n_{female} = 523,939 and n_{male} = 539,631; $n_{state\ school}$ = 1,025,979 and $n_{private\ school}$ = 37,591) eighth grade students to the math subtest.

2.3. Data Collection

In this research, the data of the eighth grade mathematics subtest of 2012 Placement Test which was applied in selecting students to secondary education were used. The data used in the research were accessed with the permission of the Ministry of National Education Innovation and Educational Technologies General Directorate. The mathematics subtest consists of 20 questions and shows a one-dimensional structure with normal distribution.

2.4. Data Analysis

In order to search for answers to the questions in the sub-objectives, firstly, the suitability of the data for the analysis of DIF determination methods was examined. LISREL 8.51 and SPSS 21 package programs were used to determine whether the items in the mathematics subtest met the unidimensionality and normality assumptions. According to the results, it was found that mathematics subtest data provided one-dimensionality (RMSEA = .03; CFI = .90; GFI = .98 and AGFI = .97) and normality assumptions.

Descriptive statistics related to mathematics subtest are given in Table 1.

Table 1.

Descriptive Statistics Related to Gender and School Type Variables of SBS 2012 Math Subtest

		Gender		School Type	
	All group	Female	Male	Private school	State school
Number of students	1,063,570	523,939 (%49.3)	539,631 (%50.7)	37,591 (%3.5)	1,025,979 (%96.5)
Mean	7.17	7.22	7.11	14.08	6.91
Standard deviation	4.75	4.76	4.73	.17	4.54
Kurtosis	.06	.13	-.01	-.54	.25
Skewness	.93	.96	.90	-.75	.97
Average difficulty	.35	.35	.36	.70	.34
Average discrimination	.51	.51	.51	.60	.49
KR-20	.86				

As can be seen in Table 1, while the average difficulty of the test was .35 for female students, this value was found to be approximately the same .36 for male students. While the average

difficulty of the test for private schools shows that it is an easy test, this value shows that it is difficult for state schools. By examining the skewness and kurtosis coefficients, it can be said that the measurements show normal distribution. In addition, the reliability coefficient of the measurements was found to be sufficiently high with .86 for the whole group.

It was found that mathematics subtest scores of students showed significant difference according to gender (t= -11.74, p<.01). Female students' math test score means were higher than male students' scores. It was determined that mathematics subtest scores of students showed significant difference according to school type (t= 298.97, p<.01). The math test scores of private school students were higher than the mean scores of the students attending state schools.

In this research, DIF analyses were carried out with Breslowday method based on the CTT and SIBTEST, Lord's χ^2 and Raju's area measurement methods based on IRT. In terms of gender variable, female students were specified as the focal group, and male students were specified as the reference group ($n_{female=}$ 523,939, $n_{male=}$539,631). In terms of school type, private schools were specified as the focal group, and stated schools were specified as the reference group ($n_{private-school}$=37,591, $n_{stated-school}$=1,025,979). In the DIF analysis based on the Item Response Theory, the estimations were made according to the 2PL model. All the analyses of DIF determination methods were made by means of R "difR" and "mirt" packages in R Studio program.

3. Results

The results of the DIF analyses by gender of SIBTEST, Raju's area measurement, Lord's χ^2 and Breslowday methods of the items in HSEPT 2012 math subtest are given in Tables 2, 3 and 4.

Table 2.
SIBTEST Findings of the Items in HSEPT 2012 Math Subtest by Gender

Items	β	SE	χ^2	p	DIF Level	Advantage group
1	.001	.00	.305	.58		
2	-.058	.00	5,160.84	.00 *	A	Female
3	-.002	.00	5.99	.01*	A	Female
4	-.116	.00	17,294.15	.00 *	C	Female
5	.011	.00	241.18	.00 *	A	Male
6	.007	.00	95.59	.00 *	A	Male
7	.007	.00	61.62	.00 *	A	Male
8	.025	.00	986.09	.00 *	A	Male
9	.001	.00	1.55	.21		
10	.027	.00	1,597.36	.00 *	A	Male
11	.014	.00	382.82	.00 *	A	Male
12	-.031	.00	1,542.20	.00 *	A	Female
13	-.009	.00	179.64	.00 *	A	Female
14	.003	.00	25.08	.00 *	A	Male
15	.000	.00	.07	.79		
16	.020	.00	546.01	.00 *	A	Male
17	-.012	.00	210.06	.00 *	A	Female
18	.001	.00	4.07	.04 *	A	Male
19	.102	.00	14,830.39	.00 *	A	Male
20	.000	.00	.07	.78		

Table 2 shows the SIBTEST findings of the items in the HSEPT 2012 math subtest according to the variable of gender. According to Table 2, 15 of the 20 items in the mathematics subtest show DIF at the A level, and one item is at the C level according to gender. 10 of the items with DIF at A level are in favor of males, and 5 items are in favor of females. It is found that the 4th item with DIF at the C level is in favor of females.

Table 3.
Raju's Area Measurement Method Findings of the Items in HSEPT 2012 Math Subtest by Gender

Items	Statistics	p	Δ_{Raju}	Δ_{M-H}	NCDIF	DIF Level
1	21.50	.00 *	-.27	-.00	.000	A
2	48.27	.00 *	.44	.85	.003	B
3	71.98	.00 *	.07	.01	.000	A
4	56.73	.00 *	1.39	1.43	.009	C
5	52.15	.00 *	.19	-.21	.000	A
6	70.31	.00 *	.09	-.14	.000	A
7	38.84	.00 *	-.08	-.09	.000	A
8	80.29	.00 *	-.13	-.38	.001	A
9	66.39	.00 *	.08	-.03	.000	A
10	-15.35	.00 *	-4.69	-.55	.001	A
11	71.13	.00 *	.00	-.28	.000	A
12	54.03	.00 *	.33	.46	.001	A
13	61.11	.00 *	.23	.16	.000	A
14	53.47	.00 *	.22	-.07	.000	A
15	57.27	.00 *	.09	-.00	.000	A
16	90.18	.00 *	-.08	-.28	.000	A
17	37.68	.00 *	.08	.17	.000	A
18	50.62	.00 *	.03	-.03	.000	A
19	150.33	.00 *	-.66	-1.41	.009	C
20	22.59	.00 *	-.30	.02	.000	A

Table 3 shows Raju's area measurement method findings of the items in the HSEPT 2012 math subtest by gender. According to Table 3, 17 of the 20 items included in the mathematics subtest show DIF at the A level, one item is at the B level, and two items are at the C level according to gender.

Table 4.
Findings of the Lord's χ^2 and Breslow-Day Method of the Items in HSEPT 2012 Math Subtest by Gender

	Lord'un χ^2			Breslow-Day
Items	χ^2	p	χ^2	p
1	757.88	.00 *	735.36	.00 *
2	2,340.17	.00 *	214.89	.00 *
3	5,707.71	.00 *	216.59	.00 *
4	5,462.58	.00 *	135.34	.00 *
5	3,696.23	.00 *	569.42	.00 *
6	6,273.22	.00 *	764.90	.00 *
7	2,258.06	.00 *	177.39	.00 *
8	7,534.07	.00 *	238.59	.00 *
9	5,319.98	.00 *	260.10	.00 *
10	1,489.29	.00 *	804.07	.00 *
11	6,623.39	.00 *	409.94	.00 *
12	3,040.09	.00 *	149.20	.00 *
13	4,268.07	.00 *	498.31	.00 *
14	3,550.51	.00 *	682.93	.00 *
15	3,744.31	.00 *	480.41	.00 *
16	8,993.59	.00 *	237.63	.00 *
17	1,751.15	.00 *	573.83	.00 *
18	3,254.01	.00 *	91.16	.00 *
19	23,458.09	.00 *	174.44	.00 *
20	574.87	.00 *	72.20	.00 *

Table 4 presents the findings of the Lord's χ^2 and Breslow-Day method by gender in the items in the HSEPT 2012 math subtest. Table 4 shows that χ^2 and p values. According to Lord's χ^2 and Breslow-Day methods, it was found that each item included in the mathematics subtest contains DIF by gender.

The results of the DIF by school type analysis of SIBTEST, Raju's area measurement, Lord's χ^2 and Breslowday methods of items in HSEPT 2012 math subtest are given in Tables 5, 6 and 7.

Table 5.
SIBTEST Findings of the Items in the HSEPT 2012 Mathematics Subtest by School Type

Items	β	SE	χ^2	p	DIF Level	Advantage group
1	-.047	0.00	174.41	.00 *	A	Private school
2	.043	0.00	101.50	.00 *	A	State school
3	.043	0.00	101.08	.00 *	A	State school
4	.009	0.01	4.22	.04 *	A	State school
5	-.039	0.00	196.02	.00 *	A	Private school
6	-.038	0.00	138.19	.00 *	A	Private school
7	-.004	0.01	0.61	.44		
8	.026	0.00	41.13	.00 *	A	State school
9	.079	0.00	357.07	.00 *	A	State school
10	-.019	0.00	45.18	.00 *	A	Private school
11	-.035	0.00	127.77	.00 *	A	Private school
12	-.070	0.00	446.57	.00 *	A	Private school
13	-.057	0.00	416.29	.00 *	A	Private school
14	-.058	0.00	549.01	.00 *	A	Private school
15	.016	0.01	13.03	.00 *	A	State school
16	.132	0.00	966.23	.00 *	A	State school
17	.041	0.00	97.29	.00 *	A	State school
18	-.016	0.00	34.58	.00 *	A	Private school
19	.043	0.00	97.07	.00 *	A	State school
20	-.011	0.00	14.19	.00 *	A	Private school

Table 5 shows the SIBTEST findings of the items included in the HSEPT 2012 math subtest by school type. Table 5 shows that 19 of the 20 items in the mathematics subtest show DIF by school type at A level. 10 of the items with DIF at A level are in favor of private schools, and 9 of them are in favor of state schools.

Table 6.
Raju's Area Measurement Method Findings According to School Type of Items in HSEPT 2012 Math Subtest

Items	χ^2	p	Δ_{Raju}	Δ_{M-H}	NCDIF	DIF Level
1	-27.22	.00 *	3.90	.19	.000	A
2	16.55	.00 *	3.33	.54	.001	A
3	17.82	.00 *	3.31	.38	.001	A
4	11.69	.00 *	2.46	-.58	.002	A
5	-30.64	.00 *	5.10	-.89	.004	B
6	24.95	.00 *	3.60	-.28	.000	A
7	-24.58	.00 *	3.44	-.25	.000	A
8	36.34	.00 *	3.49	1.16	.006	B
9	29.24	.00 *	3.85	.60	.002	A
10	-35.57	.00 *	11.29	-1.22	.007	B
11	22.12	.00 *	3.57	1.03	.005	B
12	-27.53	.00 *	3.22	.09	.000	A
13	-33.14	.00 *	3.74	-.56	.001	A
14	-49.86	.00 *	4.89	-.36	.001	A

Table 6 continued

Items	χ^2	p	Δ_{Raju}	Δ_{M-H}	NCDIF	DIF Level
15	22.66	.00 *	3.08	-.02	.000	A
16	22.59	.00 *	3.59	.32	.001	A
17	16.81	.00 *	3.58	.19	.000	A
18	39.95	.00 *	3.67	.54	.001	A
19	25.23	.00 *	3.36	.38	.001	A
20	-11.62	.00 *	7.30	-.58	.002	A

Table 6 shows the findings of Raju's area measurement method according to school type of items in HSEPT 2012 math subtest. According to Table 6, 16 of the 20 items included in mathematics subtest show DIF by school type at level A, and four items are at B level.

Table 7.

Findings Related to Lord's χ^2 and Breslow-Day Method by School Type of Items in HSEPT 2012 Math Subtest

	Lord'un χ^2		Breslow-Day	
Items	χ^2	p	χ^2	p
1	737.88	.00 *	75.57	.00 *
2	672.65	.00 *	61.38	.00 *
3	850.19	.00 *	81.61	.00 *
4	431.16	.00 *	42.23	.00 *
5	788.63	.00 *	247.36	.00 *
6	827.26	.00 *	80.59	.00 *
7	406.82	.00 *	220.95	.00 *
8	6,953.13	.00 *	111.89	.00 *
9	3,249.42	.00 *	182.31	.00 *
10	1,115.52	.00 *	243.27	.00 *
11	567.50	.00 *	86.47	.00 *
12	757.88	.00 *	113.32	.00 *
13	1,118.26	.00 *	152.97	.00 *
14	1,634.86	.00 *	271.43	.00 *
15	2,707.26	.00 *	65.34	.00 *
16	689.69	.00 *	63.56	.00 *
17	635.45	.00 *	31.08	.03 *
18	2,461.20	.00 *	23.08	.09
19	3,088.94	.00 *	158.03	.00 *
20	174.16	.00 *	85.79	.00 *

Table 7 shows the findings of the Breslow-Day and Lord's χ^2 methods according to the school type of the items in the HSEPT 2012 math subtest. Table 7 shows $\chi 2$ and p values. It was found that each item except for the 18th item in mathematics subtest showed DIF in Breslow-Day method according to school type ($p<.05$). All items show DIF according to Lord's χ^2 method ($p<.05$).

4. Discussion and Conclusion

In this study, whether the items in the 8th grade HSEPT mathematics sub-test of 2012 showed DIF according to gender and school type variables was examined by Breslow-Day, SIBTEST, Lord's χ^2 and Raju's area measurement methods. The findings obtained can be summarized as follows.

As a result of the analyses, when the gender variable was examined, according to SIBTEST method, one item was at high level, and fifteen items were at low level; however, according to Raju's area measurement method, two items were at high level, one was at moderate level, and seventeen items had low level DIF. According to the Lord's χ^2 and Breslow-Day method, it was determined that all items showed DIF. In addition, according to SIBTEST method, ten of the items

with DIF were in favor of males, six of them were in favor of females, and also one item with a high level of DIF was in favor of females. When the school type variable was examined, according to SIBTEST method, nineteen items had low level DIF; however, according to Raju's area measurement method, four items showed moderate level DIF, and sixteen items showed low level DIF. According to the Lord's χ^2 method, all items showed DIF whereas nineteen items showed DIF according to Breslow-Day method. In addition, according to the SIBTEST method, ten of the items with low level of DIF were found to be in favor of private schools, and nine of them were in favor of state schools. In line with these findings, it can be interpreted that all four methods give moderately similar results in determining DIF. In addition, when the results obtained were examined, it was found that items with DIF were similar according to the four methods used, but DIF levels differed according to the four methods. It is considered that the reason for this difference may be due to the different criteria used in the evaluation of DIF methods.

When DIF studies, which are given in the literature and conducted on the data obtained from HSEPT administered in different years, were examined (Arıkan, Uğurlu & Atar, 2016; Kan, Sünbül & Ömür, 2013; Karakaya, 2012; Karakaya & Kutlu, 2012; Kelecioğlu, Karabay & Karabay, 2014; Terzi & Yakar, 2018; Toprak & Yakar, 2017; Yıldırım & Büyüköztürk, 2018), it was found that the items showed DIF in different numbers and levels according to the results obtained from different DIF determination methods. These results are consistent with our study. In addition, Karakaya and Kutlu (2012) concluded that Logistic Regression and Mantel-Haenszel methods showed moderate similarity in their studies examining the items in the 2009 HSEPT Turkish subtest in terms of gender and school type variables. Arıkan, Uğurlu and Atar (2016) examined the similarities and differences of MIMIC, SIBTEST, Logistic Regression and Mantel-Haenszel methods over different samples (300, 600, 1000, 1200 and 2000). As a result of the study, they found that the number of items with DIF determined by SIBTEST method increased as the sample size increased. Çepni (2011) carried out DIF studies with SIBTEST, Mantel-Haenszel, Logistic Regression and methods based on Item Response Theory and stated that methods determined DIF in similar and different items. According to the findings of this study, when gender and school type variables were taken into consideration, it was seen that the four methods used in the study determined DIF in almost all items. This finding is consistent with the finding that hypothesis tests were meaningful for almost every item when Çepni (2011) and Kan, Sünbül and Ömür (2013) used IRT methods. In addition, Breslow-Day method, which was used less frequently in literature, was used in this study. Similar to the results of the other three methods, DIF was also determined by Breslow-Day method in almost all items.

4.1. Suggestions

According to the four methods used in this study, it was determined that items with DIF showed different levels of DIF. In the light of the findings obtained from the research, it may be suggested that researchers should use at least two methods in the studies to determine DIF.

In line with the purpose of this study, it was investigated whether the items showed DIF by using different methods or not, but no bias studies were conducted in this research. Even if different criteria are considered according to the same DIF determination method, whether an item has DIF or not, or the level of DIF may vary. For this reason, in further studies, whether the items with DIF are biased or not can be determined through judgemental process. Besides, DIF and bias studies can be conducted by considering different variables in addition to gender and type of school (socio-economic status, region, etc.).

References

Aguerri, M. E., Galibert, M. S., Attorresi, H. F., & Marañón, P. P. (2009). Erroneous detection of nonuniform DIF using the breslow-day test in a short test. *Quality & Quantity*, *43*(1), 35-44.

Arıkan, Ç. A., Uğurlu, S., & Atar, B. (2016). MIMIC, SIBTEST, lojistik regresyon ve mantel-haenszel yöntemleriyle gerçekleştirilen DMF ve yanlılık çalışması. *Hacettepe University Journal of Education, 31*(1), 34-52.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Breslow, N. E., & Day, N. E. (1980). Statistical methods in cancer research. Volume I - The analysis of case-control studies. *IARC Sci Publ*, (32), 5–338

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

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.

Crocker, L., & Algina J. (1986). *Introduction to classical and modern test theory*. Orlando: Harcourt Brace Jovanovich Inc.

Çepni, Z. (2011). *Değişen madde fonksiyonlarinin SIBTEST, mantel haenzsel, lojistik regresyon ve madde tepki kuramı yöntemleriyle incelenmesi.* (Doctoral dissertation, Hacettepe University). Retrieved from https://tez.yok.gov.tr

Dorans, N. J., & Holland, P. W. (1993). *DIF detection and description: Mantel-Haenszel and standardization.* In P. W. Holland & H. Wainer (Eds.), *Differential item functioning (pp. 35-66).* Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance of female candidates on the scholastic aptitude test. *Journal of Educational Measurement, 23*(4), 355–368.

French, B. F., & Finch, W. H. (2013). Extensions of mantel-haenszel for multilevel dif detection. *Educational and Psychological Measurement*, *73*(4), 648–671.

French, B. F., & Finch, W. H. (2015). Transforming SIBTEST to account for multilevel data structures. *Journal of Educational Measurement*, *52*(2), 159-180.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced english examination. *Language Assessment Quarterly, 4*(2), 190-222.

Gierl, M. J. (2005). Using dimensionality-based dif analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice*, *24*(1), 3-14.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J., (1991). *Fundamentals of item response theory.* London: Sage.

Hidalgo, M. D., & Gomez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education*, (Vol. 4, pp. 36-44). Oxford: Elsevier.

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

Holland, P. W., & Thayer, D. T. (1988*). Differential item performance and the mentel-haenszel procedure*. In H. Holland & H.I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Jöreskog, K., & Sörbom, D. (1989). *LISREL 7 User's Reference Guide.* Chicago: Scientific Software International.

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.

Karakaya, İ. (2012). An Investigation of item bias in science and technology subtests and mathematic subtests in in level determination exam. *Theory and Practice in Educational Sciences*, *12*(1), 215–229.

Karakaya, İ., & Kutlu, Ö. (2012). An Investigation of item bias in Turkish subtests in level determination exam. *Journal of Education and Science, 37*(165), 348–362.

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment*, *11*(2), 59-76.

Kelecioğlu, H., Karabay, B., & Karabay, E. (2014). Investigation of placement test in terms of item biasness. *Elementary Online*, *13*(3), 934–953.

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicologica: 30*(2), 343-370.

Magis, D., Beland, S., Teurlinckx, F., & Boeck, P. (2010). A general framework and an R package forthe detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297–334.

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

Penfield, R. D. (2003). Applying the breslow-day test of trend in odds ratio heterogeneity to the analysis of nonuniform dif. *The Alberta Journal of Educational Research, 49*(3), 231-243.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502.

Raju, N. (1990). Determinig the significance of estimated signed and unsigned areas between two item response functions**.** *Applied Psychological Measurement,* 14(2), 197-207.

Raju, N. S., & Arenson, E. (2002). *Developing a common metric in item response theory: An area-minimization approach*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33*(2), 133–147.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/ dif from group ability differences and detects test bias / DTF as well as item bias / DIF. *Psychometrika*, *58*(2), 159–194.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402–415.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Terzi, R., & Yakar, L. (2018). Differential item and differential distractor functioning analyses on Turkish high school entrance exam. *Journal of Measurement and Evaluation in Education and Psychology*, *9*(2), 136-149.

Toprak, E., & Yakar, L. (2017). Analysis of SBS 2011 Turkish subtest items in terms of differential item functioning by different methods. *International Journal Of Eurasia Social Sciences, 8*(26), 220-231.

Wiberg, M. (2007). *Measuring and detecting differential item functioning incriterion-referenced licensing test: A theoretic comparison of methods* (EM No. 60). Umea University Department of Educational Measurement, Umea.

Wright, K. D., & Oshima, T. C. (2015). An effect size measure for Raju's differential functioning for items and tests. *Educational and Psychological Measurement*, *75*(2), 338-358.

Yıldırım, H., & Büyüköztürk, Ş. (2018). Using the delphi technique and focus-group interviews to determine item bias on the mathematics section of the Level Determination Exam for 2012. *Educational Sciences: Theory & Practice, 18*(2), 447-470.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223-233.