International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 145–159
https://doi.org/10.21449/ijate.1135368
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

A Comparison of the efficacies of differential item functioning detection methods

Munevver Basman[1,*]

[1]Marmara University, Faculty of Education, Department of Educational Sciences, Türkiye

ARTICLE HISTORY

Received: June 24, 2022
Accepted: Mar. 02, 2023

Keywords:
Crossing simultaneous item bias test,
Differential item functioning,
Logistic Regression,
Lord's chi-square,
Mantel-Haenszel,
Raju's area measure.

Abstract: To ensure the validity of the tests is to check that all items have similar results across different groups of individuals. However, differential item functioning (DIF) occurs when the results of individuals with equal ability levels from different groups differ from each other on the same test item. Based on Item Response Theory and Classic Test Theory, there are some methods, with different advantages and limitations to identify items that show DIF. This study aims to compare the performances of five methods for detecting DIF. The efficacies of Mantel-Haenszel (MH), Logistic Regression (LR), Crossing simultaneous item bias test (CSIBTEST), Lord's chi-square (LORD), and Raju's area measure (RAJU) methods are examined considering conditions of the sample size, DIF ratio, and test length. In this study, to compare the detection methods, power and Type I error rates are evaluated using a simulation study with 100 replications conducted for each condition. Results show that LR and MH have the lowest Type I error and the highest power rate in detecting uniform DIF. In addition, CSIBTEST has a similar power rate to MH and LR. Under DIF conditions, sample size, DIF ratio, test length and their interactions affect Type I error and power rates.

1. INTRODUCTION

Tests are tools that contain systematic processes used to evaluate latent traits (Linn & Gronlund, 2000). With the results obtained from the tests, groups with different traits can be compared, and various decisions can be made based on the comparison results. However, if the test items are biased in favor of a group and not fair, the validity of the test is affected (Kane, 2006; Messick, 1989). For this reason, studies on the reliability and validity of the tests are carried out.

One way to ensure the validity of the tests is to check that all items work similarly across different groups of individuals. Differential item functioning (DIF) occurs, however, when individuals with equal ability levels from various groups perform differently on the same test item. In other words, DIF is the differentiation of the probability of subgroups with the same ability to correctly answer the item (Gao, 2019; Hambleton et al., 1991). While determining DIF in bias studies, two groups can be studied as the focus and the reference groups. The focus group is the one in which the negative situations of individuals with the same ability are examined while responding to the item. The group to which the focus group is compared is

*CONTACT: Munevver Başman munevver.rock@gmail.com Marmara University, Faculty of Education, Department of Educational Sciences, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

called the reference group (Zumbo, 1999). The focus group is also called the minority, and the reference group as the majority (de Ayala, 2009). When comparing the item parameters and the item characteristic curves (ICC) of the groups, it is checked whether they are different.

DIF occurs in two forms: uniform DIF and non-uniform DIF (Mellenbergh, 1983). The item examined in the uniform DIF has a situation where a certain group works in favor of the other group at every ability level. In other words, it is a situation where the percentage of a group answering an item correctly at each ability level is consistently high (Osterlind & Everson, 2009). The ICCs of both groups are different and do not overlap with each other. Uniform DIF is indicated when item difficulty (b-parameters) differs between groups (reference and focus group). In non-uniform DIF, the item studied is in favor of one group in a certain skill level range, while it works in favor of the other group in another ability range (Camilli & Shepard, 1994; Hambleton et al., 1993; Swaminathan & Rogers, 1990). The ICC of both groups are different, but they overlap at some point on the ability (theta) scale. Non-uniform DIF is detected when item discrimination (a-parameters) or both a and b parameters differ across groups.

DIF detection methods are basically classified according to the Classical Test Theory (CTT), which takes into account the observed score group, and Item Response Theory (IRT), which takes into account the latent variable group. Since the test score in the CTT is dependent on the item sample, there are limitations in the generalization of the DIF results. Therefore, there are trends toward IRT in later studies (Embretson & Reise, 2000; Hambleton et al., 1993). When DIF determination methods according to IRT and CTT are compared, the estimation of the item parameters with IRT gives more meaningful results than the CTT, the differences in item functions can be defined more meaningfully by plotting the differences in the IRT compared to the CTT, and it is easier with the IRT than the CTT, to understand whether the item shows DIF or not (Camilli & Shepard, 1994; Narayanan & Swaminathan, 1996). However, DIF detection methods based on IRT require large sample size and assumptions may be difficult to meet in practice (Narayanan & Swaminathan, 1994). DIF determination methods based on CTT can be preferred because CTT can also be used in small samples and assumptions are easier to meet in practice than IRT.

According to CTT, analysis of variance, chi-square, transformed item difficulty, the Mantel-Haenszel (MH) method, and the Logistic Regression (LR) procedure are some methods for detecting DIF. Some DIF detection methods based on IRT are Lord's chi-square (LORD), Raju's area measure (RAJU), the IRT Likelihood Ratio test (IRT-LR), Lord's IRT Wald test, the crossing simultaneous item bias test (CSIBTEST), and the Multiple Indicators Multiple Causes (MIMIC) model (Camilli & Shepard, 1994; Gao, 2019; Oshima & Morris, 2008). This research compares MH, LR, LORD, RAJU, and CSIBTEST methods. MH and LR methods among CTT methods are the most used methods in research due to their ease of use and interpretation (Kelecioğlu et al., 2014). Among the IRT methods, LORD based on chi-square method, RAJU based on ICC and CSIBTEST not requiring item calibration were chosen because they use different procedures.

Mantel-Haenszel method, proposed by Holland and Thayer (1988), is a test statistic based on chi-square. In this method, two levels are used for the item score variable (correct and incorrect response), two levels are used for group membership (focal and reference groups), and k levels are used for the matching variable. It is tested whether the probability of having the correct response for an item at a given level of the matching variable differs between the groups across all k levels of the matching variable (Dorans & Holland, 1992). The MH statistic based on chi-square is computed and logarithmic transformation is applied to facilitate the interpretation of MH results.

Logistic Regression, proposed by Swaminathan and Rogers (1990), can detect uniform and non-uniform DIF within dichotomous data. While determining the DIF with this method, a likelihood ratio is used (Camilli, 2006). Group belonging and total test score are the independent variables, and item score (0,1) is the dependent variable. It uses the total test score to estimate the traits of reference and focal groups and compares their response probabilities considering their ability differences.

Lord's chi-square, proposed by Lord (1980), is used to simultaneously check the differences in the item parameters between focal and reference groups. The chi-square statistic is calculated using item parameter differences and the variance-covariance matrix for these differences. A decision is made whether to reject or not reject the null hypothesis of no DIF by comparing the chi-square statistic with a critical value.

Raju's area measure, proposed by Raju (1988), detects DIF considering item characteristic curves. ICC for reference and focal groups are drawn according to the correct response probability, and the areas between these curves are compared with each other.

Simultaneous item bias test (SIBTEST) uses a latent score and does not need item calibration even though it is based on the IRT framework. The crossing simultaneous item bias test (CSIBTEST), proposed by Li and Stout (1996), is an extension of SIBTEST (Shealy & Stout, 1993). It is capable of detecting both uniform and non-uniform DIF, while SIBTEST can detect only uniform DIF.

In the literature, studies exist about the performances of DIF detection methods considering some variables. Holmes Finch and French (2007) compared SIBTEST, LR, IRT-LR, and confirmatory factor analysis (CFA) changing different factors. They found no significant differences in Type I error rates within the methods across the values used for the underlying model, group ability, and sample size. In addition, they found that power rates increased with increasing sample sizes and decreased with decreasing percentages of DIF for LR and IRT-LR. Güler and Penfield (2009) compared LR and a combination of MH and Breslow-Day (BD) procedures called the combined decision rule (CDR) to simultaneously detect both uniform and nonuniform DIF under the condition of different sample sizes and unequal ability distributions for focal and reference groups. Type I error rates and CDR and LR power rates were higher when the sample size was larger. DeMars (2009), Li et al. (2012) and Erdem Keklik (2014) compared MH and LR methods under different conditions. Type I error rates of MH and LR were found to be similar when the reference and focus group ability distributions showed a unit normal distribution. Kim (2010) compared MH, LR, LORD, and the Differential Functioning Item and Test (DFIT). A larger sample size inflated all methods' Type I error rates, and a longer test inflated the Type I error rates of MH and LR. Lopez (2012) compared the efficacy of CSIBTEST, IRT-LR, and LR. LR showed the highest predictive power and the lowest average Type I error rate. IRT-LR and CSIBTEST showed higher values than the nominal alpha level of .05. Atalay Kabasakal et al. (2014) compared the Type I errors and powers of MH, SIBTEST, and IRT-LR methods by using different values for test length, sample size, percentage of DIF, ability differences between groups, and underlying models. Type I error of SIBTEST and power rates of MH had the highest values. The factors' main and interaction effects can differentiate the methods' power and Type I error rates. Gao (2019) compared MH, LR, MIMIC model, Lord's IRT-based Wald test, IRT-LR, and a Randomization Test based on an R-square change statistic. The MIMIC model had the highest power rates. The LR had higher Type I error rates for larger sample sizes and shorter tests.

When the studies were examined, it was seen that the performances of DIF determination methods were examined by considering some variables. Although many DIF detection models have been developed and extensively studied in binary data, there are still ongoing studies in the literature on the limitations and advantages of these models and under what conditions they

can be used for which data. On the other hand, it is seen in the literature that comparison studies are done under limited conditions due to their nature (Jodoin & Gierl, 2001; Li et al., 2012; Narayanan & Swaminathan, 1994; Roussos & Stout, 1996). Within the scope of this study, Type I error, and power ratios of methods based on CTT and IRT used in determining DIF were tried to be determined by considering both the main effects and the interaction effects of various conditions. From this point of view, it contributes to the field since the methods used, the conditions used and their levels are differentiated, and also the interaction effects of the factors are discussed together with the main effects. This is the first study in the literature that compares MH, LR, LORD, RAJU, and CSIBTEST methods at the same time considering different sample sizes, test lengths and proportions of DIF items.

In examining the performance of DIF methods, it was necessary to examine the uniform DIF determination processes, which commonly occur in real situations, under the conditions of ability distribution, sample size and sample size ratios, which are especially used by comparison criteria such as Type I error and which can affect the results of DIF analysis. For these reasons, in the presence of a uniform DIF underlying the 3PL model, this paper answers the following questions:

a) How do the Type I error rates of MH, LR, LORD, RAJU, and CSIBTEST methods change in conditions where the sample size is 500 and 2000; test length is 10, 20 and 30; percentage of items showing DIF is 10% and 20%?

b) How do the statistical power levels of MH, LR, LORD, RAJU, and CSIBTEST methods change in conditions where the sample size is 500 and 2000; test length is 10, 20 and 30; percentage of items showing DIF is 10% and 20%?

2. METHOD

This study compares five DIF detection methods using simulation, considering their power and Type I error rates. The model of the research is basic research since it is a research that will contribute to the previous knowledge in the literature by providing information about the performances of MH, LR, LORD, RAJU, and CSIBTEST methods (Karasar, 2021).

These DIF methods can demonstrate different conclusions according to different variables (e.g. trait distribution differences, sample sizes, length of the test, ratio of items with DIF, model type, and DIF type). The procedures performed to examine these five DIF methods in this study are presented below.

2.1. Simulation Conditions

A Monte Carlo simulation is utilized to analyze the Type I error rates and power of five DIF detection methods by changing independent variables: the sample sizes for the focal and reference groups, the test length, and the proportion of items showing DIF.

Sample size: Sample size per group can affect DIF detection rates. Hidalgo et al. (2016) indicate that the sample sizes are 250 per group for small size and 1000 per group for large size, and these sample sizes reflect situations in practice. Kaya et al. (2015) state that the small sample size is 250 per group in simulation studies to investigate DIF. Güler and Penfield (2009) identify 200-250 individuals per group as the small sample size and 1000 individuals per group representing the large sample size. Jodoin and Gierl (2001) used 250 per group for small and 1000 per group for large sample sizes in their simulation study. In this study, the sample size was simulated at 250 and 1000 per group for small and large sample sizes, respectively.

Test length: Test length can also affect DIF detection rates. If the number of items increases, more reliable results and more precise estimation of ability can be obtained (Narayanan & Swaminathan, 1996). Herrera and Gomez (2008) simulated 10 items, Rockoff (2018) simulated 10, 20 and 40 items, Gao (2019) simulated 20 and 40, Lopez (2012) simulated 15 and 30 items;

Glas and Meijer (2003) and Uysal et al. (2019) used 30 items in their simulation study. In this study, the test length was set at 10, 20 and 30 items, considered short tests in the literature (Narayanan & Swaminathan, 1994).

The proportion of DIF items: The proportion of items exhibiting DIF can affect DIF detection rates similar to test length but in the opposite direction. When the proportion of DIF items increases, DIF detection rates are likely to decrease (Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1996). Demars and Lau (2011) stated that the percentages of DIF items were generally no more than 30%. Narayanan and Swaminathan (1994) indicated that the proportion of items showing DIF was either 10% or 20% in the simulation studies conducted to determine the effect of the proportion of items with DIF. Apinyapibal et al. (2015), Gao (2019), Holmes Finch and French (2007), Jodoin and Gierl (2001), Narayanan and Swaminathan (1996) used 10% and 20% DIF items in their study. In this study, the proportion of DIF items was 10% and 20%. Although these rates have been discussed in other studies, the performances of these five methods considering 10% and 20% DIF items have not been examined previously. The trait distribution (normal distribution), the model type (3PL), and the DIF type (uniform) remain constant in this research even though they also affect DIF detection methods.

In the simulation, sample size (500, 2000), test length (10, 20, 30), and percentage of items showing DIF (10%, 20%) are considered as manipulated conditions, while uniform DIF and 3PL models were considered as fixed conditions.

2.2. Data Generation

Data were generated using the 3PL IRT model which considers the case of answering correctly by chance. Item parameters were obtained through the WinGen3 software (Han & Hambleton, 2014). Tests consisting of 10, 20 and 30 items were created using the distributions of the item parameters obtained from an administration of the TIMSS 2019 paper-based Mathematics Test, a real test application to generate the data. The slope and the location parameters were generated using normal distributions with means of 1.3 and 0.531 and standard deviations of 0.357 and 0.52, respectively; the guessing parameters were set at 0.20 for all items because this parameter is near the upper end of its typically observed range (Reise & Waller, 2002). Lopez (2012) states that guessing is a realistic possibility in many testing applications and it is difficult to interpret the manipulations involving c-parameters in the context of DIF studies. Since fixing the c-parameters reduces Monte Carlo noise, a constant value of 0.20 is used for the guessing parameters.

A normal distribution with a mean of 0 and a standard deviation of 1 was used to generate the ability parameters. The differences between the location parameters of focus and reference groups for DIF items were taken as 0.60. Uniform DIF was simulated by randomly determining items. Items with DIF were applied using WinGen3 thus, 1-0 data were obtained for the focus and reference groups.

The simulation design consisted of 12 DIF conditions in total, which combined three different test lengths, two different sample sizes, and two different proportions of DIF items. Under each condition, 100 replications were made because it is common to obtain stable results (Kim, 2010). Thus, a total of 1200 data was generated. DIF analyses were performed for each data set with the five DIF methods mentioned before.

2.3. Data Analysis

The distributions of the slope and the location parameters obtained from an administration of the TIMSS 2019 paper-based Mathematics Test, a real test application, were determined using ARENA Input Analyser program. The test included number, algebra, geometry, data and probability items and was applied to 8th grade students. According to the results, the slope and location parameters distributions were normal distributions with means of 1.3 and 0.531 and

standard deviations of 0.357 and 0.52, respectively. According to these distributions, the data were generated by the software WinGen3.

The data were analyzed and the methods were compared using dichoDif in the difR package (Magis et al., 2022) for data analysis of the R statistical software (version 4.0.2, R Core Team, 2022). Type I error, and the power rates were used to compare the performances of the methods. Type I error is the decision that the item shows DIF even though it does not actually show DIF. The power is the decision that the item showing DIF is determined as having DIF due to the analysis. Methods with high power rates and low Type I error rates are preferred for determining whether or not an item has DIF. According to Bradley (1978; as cited in Hidalgo et al., 2016), the Type I error rate should be between 0.025 and 0.075. The power of methods should be at least .80 to be sufficient and this criterion is widely used in the literature (Atar, 2007).

To compare the performances of the methods, a one-way ANOVA (assumptions have been met as the data for each group have a normal distribution, and these distributions have the homogeneity of variance) was also used for each study criteria to facilitate interpretations. In addition, factorial ANOVA was used to examine the interaction effects of the factors. The statistical significance findings of the respective analyses and post hoc comparisons were examined.

3. RESULTS

This research compares the DIF detecting methods using Type I error and their power rates under various conditions. For these 1200 data, it was examined whether there were significant differences between the performances of the DIF detection methods. The results in each condition are shown in Table 1.

As seen in Table 1, Type I error rates for small sample sizes in all conditions by MH and LR methods range from .034 to .081 and generally are lower than .075 and higher than .025, while Type I error rates for large sample sizes range from .063 to .174. When all methods are compared according to sample size, it is seen that Type I error rates are higher for large sample sizes.

Table 1. *Type I Error Rate and Power Rate by Study Procedures.*

Sample Size (Reference/ Focal)	Test length	%DIF	Number of DIF items	MH		LR		CSIBTEST	
				Type I	Power	Type I	Power	Type I	Power
500 (250/250)	10	10	1	.051	**.680**	.060	**.620**	**.096**	**.710**
		20	2	**.081**	**.640**	**.081**	**.575**	**.134**	**.670**
	20	10	2	.052	**.625**	.058	**.620**	**.079**	**.620**
		20	4	.034	**.030**	.039	**.060**	.056	**.060**
	30	10	3	.037	**.390**	.048	**.417**	.056	**.357**
		20	6	.069	**.457**	.067	**.440**	.071	**.383**
2000 (1000/1000)	10	10	1	.072	1.00	.074	1.00	**.128**	1.00
		20	2	**.174**	.985	**.160**	.980	**.271**	.985
	20	10	2	.068	.965	.063	.975	**.093**	.955
		20	4	**.131**	.838	**.114**	.890	**.173**	.853
	30	10	3	**.079**	1.00	.073	1.00	**.093**	1.00
		20	6	**.117**	.885	**.100**	.915	**.150**	.885

Table 1. *Continues*

Sample Size (Reference/ Focal)	Test length	%DIF	Number of DIF items	LORD		RAJU	
				Type I	Power	Type I	Power
500 (250/250)	10	10	1	**.023**	**.440**	**.023**	**.440**
		20	2	.051	**.405**	.051	**.405**
	20	10	2	.049	**.390**	.049	**.390**
		20	4	**.024**	**.025**	**.024**	**.025**
	30	10	3	**.080**	**.273**	**.080**	**.273**
		20	6	.046	**.293**	.046	**.293**
2000 (1000/1000)	10	10	1	.064	.990	.064	.990
		20	2	**.146**	.980	**.146**	.980
	20	10	2	.051	.925	.051	.925
		20	4	**.313**	**.735**	**.313**	**.735**
	30	10	3	.072	.990	.072	.990
		20	6	**.137**	**.768**	**.137**	**.768**

Note. MH=Mantel-Haenszel, LR= Logistic Regression, CSIBTEST=crossing simultaneous item bias test, LORD=Lord's chi square (χ^2), RAJU=Raju's area measure.

Power rates of all methods for small sample sizes are lower than .80 in all conditions and power rates of MH, CSIBTEST and LR methods for large sample sizes are above .80 in all conditions. Power rates of LORD and RAJU for large sample sizes are acceptable values, generally more than .80. When all methods are compared according to sample size, it is seen that power rates are higher for large sample sizes. In addition, the Type I error increases, and the power rate decreases in all methods as the ratio of the item with DIF increases for large sample sizes.

The comparison of the methods depending on the sample size can be seen more clearly in Figure 1 and Figure 2.

Figure 1. *Type I error rates for sample size.*

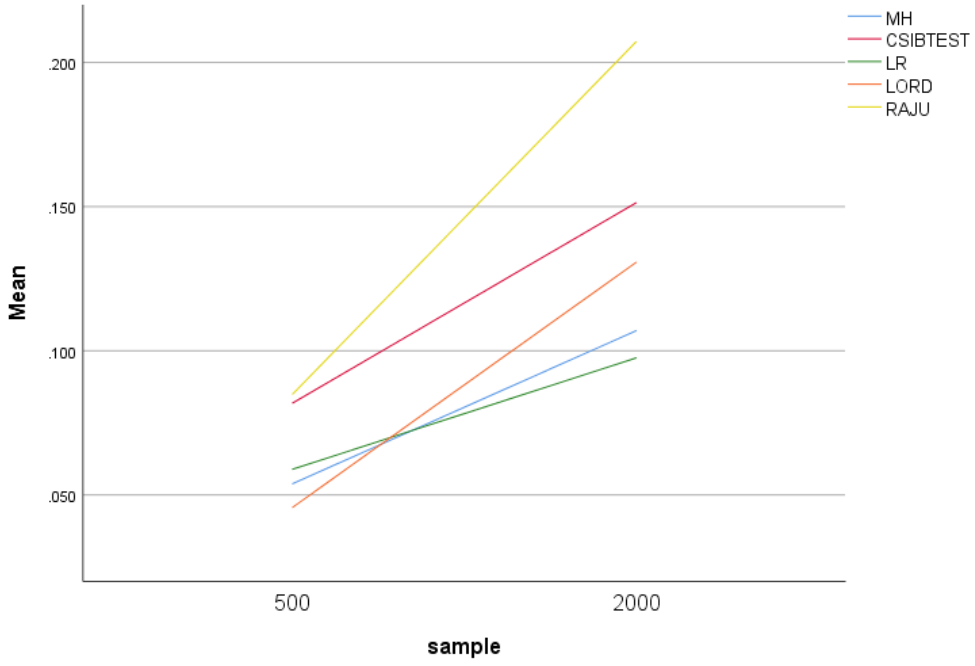
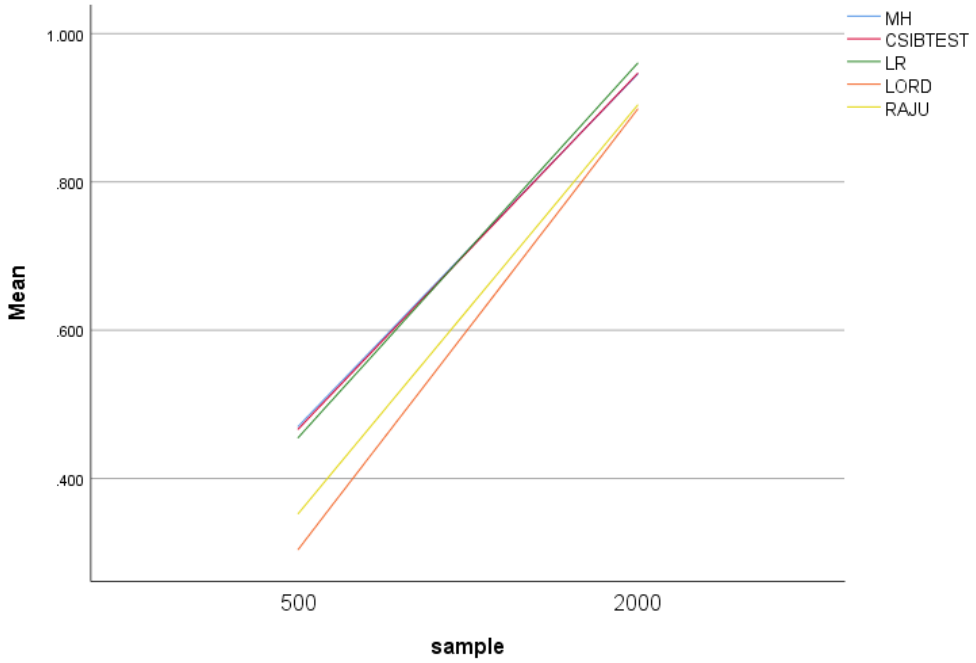


Figure 2. *Type I error rates for sample size.*



When Figure 1 and Figure 2 are examined, it can be seen more clearly that the Type I error increases and the power ratio decrease in all methods in the large sample than in the small sample. In addition, it is seen that RAJU has the highest Type I error, and MH, CSIBTEST and LR demonstrate significantly higher power rates than LORD and RAJU for both small and large sample sizes.

To facilitate interpretation, analyses of variance (ANOVA) for each procedure by manipulation were applied. The results for Type I error rates and power rates are shown in Tables 2 and 3, respectively.

It is found that the average Type I error rates of the methods are significantly different ($F(4.5995) = 67.721$, $p<.05$). Post hoc tests show that RAJU demonstrates significantly higher error rates (.146) than the other procedures. Then, it is found that CSIBTEST (.117) produces a significantly higher error rate than other methods except for RAJU. In addition, LR shows the lowest average Type I error rate, but there is no significant difference between MH, LR, and LORD.

Table 2. *ANOVA Results for Type I Error Rate by Study Procedures*

		MH		LR		CSIBTEST		LORD		RAJU	
	df	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2
S	1	180.690*	.132	82.666*	.065	206.370*	.148	152.614*	.114	216.266*	.154
T	2	13.275*	.022	13.965*	.023	71.180*	.107	10.604*	.018	18.107*	.030
P	1	108.662*	.084	53.230*	.043	114.539*	.088	83.303*	.066	118.776*	.091
*S*T*	2	1.005	.002	1.456	.002	2.835	.005	20.727*	.034	26.279*	.042
*S*P*	1	45.015*	.037	31.174*	.026	74.646*	.059	113.937*	.088	113.508*	.087
*T*P*	2	10.675*	.018	7.200*	.012	16.633*	.027	19.044*	.031	15.160*	.025
*S*T*P*	2	8.785*	.015	5.494*	.009	4.601*	.008	27.182*	.044	46.651*	.073

Note. MH=Mantel-Haenszel, LR= Logistic Regression, CSIBTEST=crossing simultaneous item bias test, LORD=Lord's chi square (χ^2), RAJU=Raju's area measure. *$p<.05$

As seen in Table 2, ANOVA results for Type I Error Rate for all procedures show that the main effects of sample size, test length, and proportion of DIF items are significant. Furthermore, significant *sample size x proportion of DIF items, test length x proportion of DIF items*, and *sample size x test length x proportion of DIF items* interactions are found for all methods. A significant *sample size x test length* interaction is found in LORD and RAJU, while it is not a significant interaction effect in other DIF detection methods.

Table 3. *ANOVA Results for Type I Error Rate by Study Procedures*

		MH		CSIBTEST		LR		LORD		RAJU	
	df	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2
S	1	1322.417*	.527	1314.089*	.525	1426.104*	.546	1903.149*	.616	1586.204*	.572
T	2	91.293*	.133	106.056*	.151	47.503*	.074	63.776*	.097	47.024*	.073
P	1	110.971*	.085	103.059*	.080	92.598*	.072	96.417*	.075	98.439*	.077
*S*T*	2	31.249*	.050	44.828*	.070	19.497*	.032	1.627	.003	.483	.001
*S*P*	1	15.721*	.013	18.436*	.015	23.855*	.020	.260	.000	.068	.000
*T*P*	2	73.230*	.110	55.534*	.085	52.614*	.081	30.717*	.049	62.915*	.096
*S*T*P*	2	53.690*	.083	45.536*	.071	43.585*	.068	20.045*	.033	34.049*	.054

Note. MH=Mantel-Haenszel, LR= Logistic Regression, CSIBTEST=crossing simultaneous item bias test, LORD=Lord's chi square (χ^2), RAJU=Raju's area measure. *p<.05

It is found that the average power rates of the methods significantly differ, too (F(4.5995) = 22.298, p<.05). Post hoc tests show that MH (.708), CSIBTEST (.707), and LR (.708) demonstrate significantly higher power rates than LORD (.601) and RAJU (.628). There are no significant power rate differences between MH, CSIBTEST and LR, and between LORD and RAJU. As seen in Table 3, ANOVA results for power rate for all methods are affected by sample size, test length, and the proportion of items exhibiting DIF. In addition, *sample size x test length* and *sample size x proportion of DIF items* are found to be statistically significant for the MH, CSIBTEST, and LR methods, while all other interactions are found to be statistically significant for all methods. The significant *test length x proportion of DIF items* and *sample size x test length x proportion of DIF items* interactions are found for all methods.

When the main factors are examined by independent samples t-test and one-way ANOVA, Type I errors and power rates of all methods for large samples are significantly higher than Type I errors and power rates for small samples. Type I errors and power rates of all methods for the shortest test length (10 items) are significantly higher than Type I errors and power rates for others (20 and 30 items), except LORD and RAJU for Type I error rates. For these methods, they are significantly lower than others. However, there are no significant differences for Type I error rates and power rates in all methods between 20 and 30 items, except RAJU (The Type I error rate of 20 items is higher than 30 items). There is a significant Type I error rate difference for RAJU and a significant power level difference for MH between 20 and 30 items. Type I errors and power rates of all methods for 20% DIF items are significantly higher than those for 10% DIF items, except MH and CSIBTEST. There is no significant difference between 10% and 20% for them.

4. DISCUSSION and CONCLUSION

The existence of differential item functioning indicates that some situations need attention in a test. If items show DIF in a test, it indicates that different undesirable factors may affect the feature that the test intends to measure (Shealy & Stout, 1993). Therefore, it is important to identify procedures that can effectively detect DIF.

This study examines the efficacy of five DIF determination methods; MH, LR, LORD, RAJU, and CSIBTEST, considering various conditions. For this purpose, a simulation study was conducted considering real data parameters from an administration of the TIMSS 2019 paper-based Mathematics Test.

According to the results, it is found that the Type I error is low for the MH method and it gives acceptable results under many conditions (Marañón et al., 1997; Shealy & Stout, 1993). Guilera et al. (2013) discussed the Type I error and power of the MH method using the meta-analysis technique and found similar results for the MH method to this study. LR demonstrates the lowest average Type I error rate, and methods show slightly greater error rates than the nominal .075 error rate. These findings support the results of the study by Lopez (2012), which compared the efficacy of CSIBTEST, IRT-LR, and LR. LR had the lowest average Type I error rate, and CSIBTEST and IRT-LR demonstrated error rates that were greater than the nominal .05 level (Lopez, 2012). In addition, no significant differences between LR, MH, and LORD according to Type I error rate are found in this study. These findings are consistent with similar studies in the literature (DeMars, 2009; Erdem Keklik, 2014; Gierl et al., 2000; Rogers & Swaminathan, 1993; Uyar, 2015; Vaughn & Wang, 2010). According to Type I error and power rate, MH and LR have the lowest Type I error rate and the highest power rate. This finding supports the results of the research of Erdem Keklik (2014), which found that the MH and LR methods were similar and had lower Type I errors than IRT-LR when the trait distributions are normally distributed. It can be concluded that MH and LR are more sensitive to detecting items with DIF than other methods in this study.

When the methods are examined under different conditions, it is seen that their Type I errors, and power rates can differ according to the conditions. Swaminathan and Rogers (1990) indicated that the sample size affects the power of DIF detection procedures. In this study, when the small sample is compared with the large sample, it is seen that the Type I error and power ratios are higher in the large sample. Contrary to Holmes Finch and French (2007), these findings are in agreement with DeMars (2009), Güler and Penfield (2009), Li et al. (2012), and Roussos and Stout (1996).

It is expected that longer tests are likely to show more reliable scores. The power of the DIF methods is likely to increase with increasing test lengths (Narayanan & Swaminathan, 1996). However, Guilera et al. (2013) demonstrated that MH for tests with lengths from 20 to 40 items showed lower Type I error and power than shorter tests. In this study, Type I errors and power rates of tests with 20 and 30 items are found to be significantly lower than the shorter test (10 items), which is consistent with Guilera et al. (2013), Kim (2010), Lopez (2012) and Uttaro and Millsap (1994).

Fidalgo et al. (2000) stated that the greater the number of items with DIF, the greater the Type I error. The finding that the Type I error and power rates of all methods increase as the ratio of the item with DIF increases is consistent with the results in the literature (Atalay Kabasakal et al., 2014; Finch, 2005; Guilera et al., 2013; Holmes Finch & French, 2007; Uyar, 2015).

When the interaction effects are examined, it is seen that Type I errors and the power rates differ according to the interactions of *test length x proportion of DIF items* and *sample size x test length x proportion of DIF items* for all methods. Type I errors, and the power rates differ according to *sample size x test length* interactions of the LORD and RAJU. Type I errors differ according to the interactions of *sample size x proportion of DIF items* for all methods, while the power rates differ for only MH, CSIBTEST, and LR. It can be concluded that the interaction effect of the variables can differentiate the Type I errors and power ratios of the methods. Thus, it is thought that interaction effects should be taken into account when using the methods.

To sum up, when the results obtained from this study and other relevant research results are evaluated together, LR and MH are used as a reason for preference, especially in small samples,

as they have the lowest Type I error and the highest power rate in detecting uniform DIF. It is seen that Type I errors, and power rates of the methods can differ according to the conditions. So, the preferred DIF determination methods should be chosen considering the applied situations and requirements of the theories (e.g. IRT-based DIF detection methods require a large sample size). It can be stated that which DIF methods to use should be decided by considering the conditions. As stated by Kelecioğlu et al. (2014) and Ayva Yörü and Atar (2019), at least two different DIF detection methods are suggested to be used to improve the reliability of the results, as different methods are seen to provide different results in certain situations. The methods to be used can be selected based on the properties of the application, such as sample size, test length etc. (e.g. LR and MH can be used if the sample size is small).

This study examined the efficacy of MH, LR, LORD, RAJU, and CSIBTEST methods considering various conditions. These DIF detection methods and used conditions are limitations of the study. Further studies may compare other DIF detection procedures based on CTT and IRT and the differences between them can be analysed according to Type I errors and their power rates. The sample sizes were 250 and 1000 per group and the proportions of DIF items were 10% and 20%. Different sample sizes and ratios of items with DIF can be used. It would be better comparing especially 10% to 30%. Test lengths were taken 10, 20 and 30 as short test lengths. Short and long test lengths can be also used. The normal trait distribution, the 3PL model type, and the uniform DIF type remain constant. Further studies may examine non-uniform DIF with these procedures by changing the values of slope and location parameters. In addition, different trait distributions and model types (1PL or 2PL) can be researched in the future.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Münevver Başman https://orcid.org/0000-0003-3572-7982

REFERENCES

Apinyapibal, S., Lawthong, N., & Kanjanawasee, S. (2015). A comparative analysis of the efficacy of differential item functioning detection for dichotomously scored items among logistic regression, SIBTEST and raschtree methods. *Procedia-Social and Behavioral Sciences*, *191*, 21-25. https://doi.org/10.1016/j.sbspro.2015.04.664

Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioğlu, H. (2014). Comparing performances (type I error and power) of IRT likelihood ratio SIBTEST and Mantel-Haenszel methods in the determination of differential item functioning, *Educational Sciences: Theory and Practice, 14*(6), 2175-2193. https://doi.org/10.12738/estp.2014.6.2165

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures* [Unpublished doctoral dissertation]. University of Florida State.

Ayva Yörü, F.G., & Atar, H.Y. (2019). Determination of differential item functioning (DIF) according to SIBTEST, Lord's [Chi-squared], Raju's area measurement and Breslow-Day Methods. *Journal of Pedagogical Research*, *3*(3), 139-150. https://doi.org/10.33902/jpr.v3i3.137

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

Camilli, G. (2006). Test fairness. In R.L. Brennan (Ed), *Educational Measurement* (4th ed., pp. 221–257). Rowman & Littlefield.

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

DeMars, C.E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*, 149-170. https://doi.org/10.3102/1076998607313923

DeMars, 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. https://doi.org/10.1177/0013164411404221

Dorans, N.J., & Holland, P.W. (1992). DIF detection and description: Mantel-Haenszel and standardization. *ETS Research Report Series*, *1992*(1), i-40. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Embretson, S.E., & Reise, S.T. (2000). *Item response theory for psychologists*. Lawrance Erlbaum Associates.

Erdem Keklik, D. (2014). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel ve lojistik regresyon tekniklerinin karşılaştırılması [Comparison of Mantel-Haenszel and logistic regression techniques in detecting differential item functioning]. *Journal of Measurement and Evaluation in Education and Psychology*, *5*(2), 12-25. https://doi.org/10.21031/epod.71099

Fidalgo, A.M., Mellenbergh, G.J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, *5*(3), 43-53.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295. https://doi.org/10.1177/0146621605275728

Gao, X. (2019). *A comparison of six DIF detection methods* [Unpublished master's thesis]. University of Connecticut.

Gierl, M.J., Jodoin, M.G., & Ackerman, T.A. (2000, April 24-27). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large* [Paper presentation] In Annual Meeting of the American Educational Research Association (AERA), New Orleans, LA, United States.

Glas, C.A., & Meijer, R.R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*(3), 217-233. https://doi.org/10.1177/0146621603027003003

Guilera, G., Gomez-Benito, J., Hidalgo, M.D. & Sanchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods, 18*(4), 553-71. https://doi.org/10.1037/a0034306

Güler, N., & Penfield, R.D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, *46*(3), 314-329. https://doi.org/10.1111/j.1745-3984.2009.00083.x

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

Hambleton, R.K., Clauser, B.E., Mazor, K.M., & Jones, R.W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1-18.

Han, K.T., & Hambleton, R.K. (2014). User's manual for WinGen3: Windows software that generates IRT model parameters and item responses (*Center for Educational Assessment Report No. 642)*. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Herrera, A., & Gomez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity, 42*(6), 739-755. https://doi.org/10.1007/s11135-006-9065-z

Hidalgo, M.D., López-Martínez, M.D., Gómez-Benito, J., & Guilera, G. (2016). A comparison of discriminant logistic regression and Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (IRTLRDIF) in polytomous short tests. *Psicothema*, *28*(1), 83-88. https://doi.org/10.7334/psicothema2015.142

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

Holmes Finch, W., & French, B.F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement*, *67*(4), 565-582. https://doi.org/10.1177/0013164406296975

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education*, *14*(4), 329-349. https://doi.org/10.1207/S15324818AME1404_2

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17– 64). Rowman & Littlefield.

Karasar, N. (2021). *Bilimsel araştırma yöntemleri* [Scientific research methods]. Nobel Yayınları.

Kaya, Y., Leite, W., & Miller, M.D. (2015). A comparison of logistic regression models for DIF detection in polytomous items: the effect of small sample sizes and non-normality of ability distributions. *International Journal of Assessment Tools in Education*, *2*(1), 22-39. https://doi.org/10.21449/ijate.239563

Kelecioğlu, H., Karabay, B., & Karabay, E. (2014). Seviye belirleme sınavı'nın madde yanlılığı açısından incelenmesi [Investigation of placement test in terms of item biasness]. *Elementary Education Online*, *13*(3), 934-953.

Kim, J. (2010). *Controlling Type I error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing*. Dissertation, Georgia State University.

Li, Y., Brooks, G.P., & Johanson, G.A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, *72*(5), 847-861. https://doi.org/10.1177/0013164411432333

Li, H.H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647-677. https://doi.org/10.1007/BF02294041

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th Ed.). Upper Saddle River.

Lopez, G.E. (2012). *Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-likelihood ratio test, crossing-SIBTEST, and logistic regression procedures* [Unpublished doctoral dissertation]. University of South Florida.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Magis, D., Beland, S., &Raiche, G. (2022). Collection of methods to detect dichotomous differential item functioning (DIF). Package 'difR'.

Marañón, P.P., Garcia, M.I.B., & Costas, C.S.L. (1997). Identification of nonuniform differential item functioning: A comparison of Mantel-Haenszel and item response theory analysis procedures. *Educational and Psychological Measurement*, *57*(4), 559-568. https://doi.org/10.1177/0013164497057004002

Mellenbergh, G.J. (1983). Conditional item bias methods. In S.H. Irvine & J.W. Berry (Eds.), *Human assessment and cultural factors* (pp. 293-302). Springer.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13-103). MacMillan.

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), 315-328. https://doi.org/10.1177/014662169401800403

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show non-uniform DIF. *Applied Psychological Measurement, 20*(3), 257-274. https://doi.org/10.1177/014662169602000306

Oshima, T.C., & Morris, S.B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43-50. https://doi.org/10.1111/j.1745-3992.2008.00127.x

Osterlind, S.J., & Everson, H.T. (2009). *Differential Item Functioning*. Sage.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. http://www.R-project.org/

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502. https://doi.org/10.1007/BF02294403

Reise, S.P., & Waller, N.G. (2002). Item response theory for dichotomous assessment data. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 88–122). Jossey-Bass.

Rockoff, D. (2018). *A randomization test for the detection of differential item functioning* [Unpublished doctoral dissertation]. The University of Arizona.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105–116. https://doi.org/10.1177/014662169301700201

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*, 215-230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Shealy, R., & Stout, W. (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. https://doi.org/10.1007/BF02294572

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Uttaro, T., & Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*(1), 15–25. https://doi.org/10.1177/014662169401800102

Uyar, Ş. (2015). Gözlenen gruplara ve örtük sınıflara göre belirlenen değişen madde fonksiyonunun karşılaştırılması [Comparing differential item functioning based on manifest groups and latent classes] [Unpublished doctoral dissertation]. University of Hacettepe.

Uysal, İ., Ertuna, L., Ertaş, F.G. & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 133-148. https://doi.org/10.21031/epod.534312

Vaughn, B.K., & Wang, Q. (2010). DIF trees: using classifications trees to detect differential item functioning. *Educational and Psychological Measurement, 70*(6) 941–952. https://doi.org/10.1177/0013164410379326

Zumbo, B.D.A. (1999). *Handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and likert type item scores*. Ottowa.