International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 781–795
https://doi.org/10.21449/ijate.1368341
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Type I error and power rates: A comparative analysis of techniques in differential item functioning

Ayse Bilicioglu Gunes[1*], **Bayram Bicak**[2]

[1]TED University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]Akdeniz University, Faculty of Education, Department of Measurement and Assessment, Antalya, Türkiye

ARTICLE HISTORY
Received: Sep. 29, 2023
Revised: Dec. 02, 2023
Accepted: Dec. 09, 2023

Keywords:
Classical test theory,
Item response theory,
Differential item functioning,
Type I error,
Statistical power.

Abstract: The main purpose of this study is to examine the Type I error and statistical power ratios of Differential Item Functioning (DIF) techniques based on different theories under different conditions. For this purpose, a simulation study was conducted by using Mantel-Haenszel (MH), Logistic Regression (LR), Lord's χ^2, and Raju's Areas Measures techniques. In the simulation-based research model, the two-parameter item response model, group's ability distribution, and DIF type were the fixed conditions while sample size (1800, 3000), rates of sample size (0.50, 1), test length (20, 80) and DIF- containing item rate (0, 0.05, 0.10) were manipulated conditions. The total number of conditions is 24 (2x2x2x3), and statistical analysis was performed in the R software. The current study found that the Type I error rates in all conditions were higher than the nominal error level. It was also demonstrated that MH had the highest error rate while Raju's Areas Measures had the lowest error rate. Also, MH produced the highest statistical power rates. The analysis of the findings of Type 1 error and statistical power rates illustrated that techniques based on both of the theories performed better in the 1800 sample size. Furthermore, the increase in the sample size affected techniques based on CTT rather than IRT. Also, the findings demonstrated that the techniques' Type 1 error rates were lower while their statistical power rates were higher under conditions where the test length was 80, and the sample sizes were not equal.

1. INTRODUCTION

Measurement tools define the levels of traits or qualities that individuals possess. Therefore, the measures obtained from them must be accurate and precise. Two fundamental properties must be present in a measurement tool: reliability and validity. Reliability refers to the stability and consistency of measurements. On the other hand, validity is a matter of whether the instrument can measure the intended characteristic. Bias is one of the threats to validity (Clauser & Mazor, 1998; Zumbo, 1999). In the administration process of some tests, measurement bias may occur due to factors such as the characteristics of the participants or culture. The responses of individuals with the same ability level but in different subgroups to an item or a test may differ. This often results in the item or test functioning differently for people in different subgroups (Dorans & Holland, 1993). Exposure to item bias has been shown to have adverse effects on validity. Camilli and Shepard (1994) define bias as the systematic error in test scores

*CONTACT: Ayse Bilicioglu Gunes ayse.bilicioglu@tedu.edu.tr TED University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

toward a certain group. If this situation, which gives an advantage to one group and a disadvantage to another group at the same ability level, occurs in only a few items of the test, it is called item bias, and if it is a circumstance that occurs throughout the test, it is called test bias (Zumbo, 1999). In item bias, the probability of responding correctly to an item depends on belonging to a group rather than the level of ability that is being measured. This is one of the main signs of measurement error (Osterlind, 1983). If test scores do not reflect the intended construct in the same way between groups or situations (or indicate different constructs for different groups), score interpretations will be biased.

The first stage of bias studies is to determine differential item functioning (DIF) as an index of bias (Camilli & Shepard, 1994). According to Hambleton et al. (1991), DIF is the difference in the probability of responding correctly to an item when individuals in different groups are at the same ability level. Two groups -the focus group and the reference group- can be examined for determining DIF analysis. In the focus group, while answering the item, the unfavorable circumstances of people with similar abilities are studied. The reference group is the group that the focus group is compared to (Zumbo, 1999).

Camilli and Shepard (1994) note two potential reasons for the incidence of DIF. These are called item effect and item bias. Item effect is the actual difference between the probabilities of respondents in different groups giving a correct answer to an item (Zumbo, 1999). This difference can be attributed to the prior experience or knowledge of one of the groups. If the participants differ in terms of the knowledge they hold, it is expected that the responses to items will also reflect this difference. DIF is a necessary, but not sufficient, condition for item bias. Thus, there is no item bias if DIF is not apparent for a given item. But even if DIF is obvious, its existence alone does not prove item bias; additional item bias analyses (such as content analysis and empirical evaluation) are required to establish the existence of item bias (Zumbo, 1999). In item bias, factors other than the ability to be measured are included in the test during the measurement. The purpose of determining DIF is to explain the differences originating from this bias (Dorans & Holland, 1993). DIF occurs in two forms: uniform DIF and non-uniform DIF (Mellenbergh, 1983). When groups and ability levels do not interact, uniform DIF appears. Conversely, an interaction is involved in non-uniform DIF (Swaminathan & Rogers, 1990).

Many techniques are used to determine uniform and non-uniform DIF. The techniques used vary depending on whether the data are dichotomous or polytomous. Techniques are broadly categorized under two theories. While DIF determination techniques based on Classical Test Theory (CTT) compare the distribution of the groups' scores, those based on Item Response Theory (IRT) compare the probability of correct response of the groups to the related item. Among the frequently used DIF determination techniques based on CTT are Mantel-Haenszel (MH), Analysis of Variance, Transformed Item Difficulty, and Logistic Regression (LR). Also, some of the frequently used techniques based on IRT are the Likelihood Ratio Test, Raju's Area Measures, and Lord's χ^2 (Camilli & Shepard, 1994; Hambleton et al., 1991).

The MH technique was developed by Nathan Mantel and William Haenszel in the 1950s as a chi-square method. Afterwards, it was updated by Holland and Thayer (1998) and introduced to identify DIF. The MH technique is a frequently utilized technique as it does not require large samples, can provide effect size values, and calculations are not complex (Samuelsen, 2005). The values obtained with this technique are interpreted with the delta scale (Δ) taking into account the category levels recommended by The Educational Test Service (ETS). LR, which is another method based on CTT, is an expanded version of the MH technique and a sensitive technique for identifying both uniform and non-uniform DIF. The technique is based on a standard logistic regression model that uses independent variables to predict two dependent variables. LR assesses the presence of DIF by utilizing responses to the item and the total score. LR is a widely used technique due to its simplicity of programming and robustness in dealing

with non-uniform DIF (Clauser & Mazor, 1998; Swaminathan & Rogers, 1990). Raju's Area Measures is a technique that evolved based on IRT. This technique takes into account the Item Characteristic Curves (ICCs) in the determination of DIF (Raju, 1988). The area mentioned in Raju's Area Measures is the gap between the estimated ICCs of two groups (Camilli & Shepard, 1994). In Lord's χ^2 method, the differences in item parameters of two different groups for an item are compared (Lord, 2012). If there is a discrepancy between the two groups as a result of the comparison, it can be said that the relevant item functions in a different way. One of the virtues of the technique is that it can be used to determine both uniform and non-uniform DIF (Price, 2014).

As pointed out by Crocker and Algina (1986) and Dorans and Holland (1993), if the items in the test are biased, it means that the decisions based on these test scores also contain errors. Therefore, it is very pivotal to elaborate on how the techniques used to determine bias work and under which conditions. Questions such as "What is the main reason for an item to include DIF?" and "What should be done if an item contains DIF?" embody the essential questions of DIF studies. Researchers have discussed these questions and offered solutions like removing the relevant item from the test and adding a new one. However, this process is quite time-consuming and also affects the content validity of the test. Another argument is that the item showing DIF should be revised. This suggestion requires the researchers to implement the revised items to a group of students and then conduct a DIF analysis again (Ellis & Raju, 2003). Given these points, studies to detect the sources of DIF and mitigate the presence of item bias are imperative. Experts' opinion is the most widely used technique to investigate whether the item is biased or not as a consequence of DIF analyses. However, in some occasions, although the item is not biased, results can give the opposite information, and at that time the question appears “Why does the item have DIF?”. One possible reason for this may be Type I error rates. The presence of Type I error can be considered as a misidentification of DIF in the scope of item bias. In this case, even though the item does not actually possess DIF, it gives a statistically significant output as containing DIF. In other words, although the item displays similar performance across individuals in the focal and reference groups, the technique used indicates that the item displays DIF (Dainis, 2008). This problem has been the subject of research for various reasons (Jodoin & Gierl, 2010). The first reason is that identifying an item as presenting DIF is a waste of time and resources. If the DIF identification is faulty, it wastes resources in the test development process. Secondly, researchers waste their time on DIF analyses. Ultimately, from the perspective of the research, if the study intends to determine the strength of DIF detection techniques or to identify DIF items from real data, inferences drawn from comparisons (in terms of the quality and effectiveness of the technique) are not be valid due to Type I errors.

Studies on the effectiveness of DIF detection techniques taking certain variables into account have been released into the literature. Ankenmann et al. (1999) aimed to compare the Type I error and statistical power rates of MH and LR techniques. The results indicated that the MH technique had better statistical power compared to the LR technique, but both techniques were influenced by sample sizes. In terms of Type I error rates, it was found that the MH technique exceeded the nominal alpha level. In a research conducted by Gierl et al. (2000), the performance of MH, LR, and SIBTEST techniques was compared under different conditions such as the presence of DIF, sample size, and ability distribution. The results suggested that even in cases with small sample sizes, the Type I error rates of all three techniques were around the nominal alpha level. The SIBTEST technique, however, exhibited the highest statistical power. Dainis (2008) compared methods based on the CTT and IRT in terms of Type I error and power rates. The study found that the LR technique yielded low power and high Type I error rates. In a study by Demars (2009) comparing the Type I error rates of MH, LR, and SIBTEST techniques under different conditions, it was found that reducing test length and

increasing sample size led to inflated Type I error rates for MH and LR techniques. Vaughn and Wang (2010) investigated the Type I error rates of classification trees, MH, and LR techniques under conditions of sample size, DIF, and ability distribution. The study concluded that MH and LR techniques had low Type I error rates for three different sample sizes and ratios. Magis and De Boeck (2012) examined the performance of the MH technique under different conditions and found that the MH technique yielded inflated Type I error rates in situations where there was a between-group ability difference and when sample size increased. Atalay Kabasakal et al. (2014) compared the Type I error rates and power of MH, SIBTEST, and MTK-OO techniques under different conditions. The results indicated that in all conditions considered, the MH technique had the highest power, while the SIBTEST technique had the highest error. In a study by Sunbul and Omur Sunbul (2016), MH, LR, Lord's χ^2, and Raju's area measures were compared in terms of Type I error and power. The results suggested that techniques based on CTT were not significantly affected by varying conditions in Type I error rates, and both theory-based techniques showed an increase in power with an increase in sample size.

When the studies were examined, it was seen that a number of research has been published on the performance of DIF techniques considering various conditions, but most of it has been conducted with techniques based solely on the CTT or IRT (Ankenmann et al., 1999; Cohen et al., 1996; Demars, 2009; Gierl et al., 2000; Jodoin & Gierl, 2010; Kristjansson et al., 2005; Lim & Drasgow, 1990; Magis & De Boeck, 2012; Vaughn & Wang, 2010). Few published studies have used both (Atalay Kabasakal et al., 2014; Atar, 2007; Dainis, 2008; Erdem Keklik, 2012; Finch & French, 2007). Although MH, LR, Raju's Area Measures, and Lord's χ^2 techniques are frequently utilized in the literature, to date, there has been little comparative research conducted on Type I errors and powers of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques at once (Basman, 2023; Sunbul & Omur Sunbul, 2016). In addition, since the presence of Type I error can be considered as misidentification of DIF within the scope of item bias and statistical power shows the performance of the techniques, this study aimed to investigate the results of MH, LR, Raju's Area Measures and Lord's χ^2 techniques under different conditions and to compare the techniques with each other by considering Type I error and power ratios during the comparison of the techniques.

In this scope, it is thought that the study's results will help determine the appropriate DIF determination techniques that can be used for institutions working with large-scale tests, test developers, and those who make decisions based on the scores of the relevant tests.

1.1. Purpose of the Study

The main purpose of this study is to examine the Type I error and statistical power ratios of DIF determination techniques based on different theories under different conditions for items scored with two categories. In pursuit of this objective, the study seeks to address the following questions:

1. How do the Type I error rates of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques differ under conditions where the sample size is 1800 and 3000, the sample size ratio is 0.50 and 0.75, the number of items is 20 and 80, and there are no DIF items?
2. How do the statistical power ratios of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques differ under the conditions in which the sample size is 1800 and 3000, the sample size ratio is 0.50 and 0.75, the number of items is 20 and 80, and the ratio of DIF- containing items is 0.05 and 0.1?

2. METHOD

Within the scope of the research, Type I error and statistical power ratios were scrutinized by conducting DIF analyses with the data generated in the 2PL model under different conditions. The research follows a simulation-based model that provides the researcher with the opportunity to work under different conditions (Dooley, 2002).

2.1. Generation of Data

The conditions held constant in the current study are the IRT model used, the ability distributions of the focus and reference groups, and the type of DIF. The conditions manipulated include sample size and proportions, test length, and the proportion of items containing DIF.

2.1.1. *Fixed conditions*

In this study, participant responses, ability levels, and item parameters for the focus and reference groups were generated in compliance with the 2PL model. Since the data fit of the 2PL model was better than that of the 1PL model, the 1PL model was not included in this study. When the 3PL model was examined under real test conditions, it was noted that the standard error in the estimations increased due to the c parameter, and therefore, it is not a strong statistic in DIF determination studies (Hambleton et al., 1991). Considering these reasons, only the 2PL model was included in this study. The ability distribution of the groups was fixed using a normal distribution with a mean of 0 and a standard deviation of 1 (Dodeen, 2004; Hauck Filho et al., 2014; Roussos & Stout, 1996).

In generating the parameters, they were held constant for both groups, while the b parameter of the reference group was altered. Data were generated using R software. For both groups, the "a" parameter was obtained from a normal distribution with a mean of 0.8 and a standard deviation of 0.04 (Sunbul & Omur Sunbul, 2016), while the "b" parameter was randomly drawn from a distribution with minimum and maximum values of -2 and +2, respectively (Desa, 2012; Kogar, 2018). In studies conducted on simulated data in the literature, various values for the amount of DIF have been utilized, such as 0.10, 0.15, 0.25, 0.30, 0.32, 0.43, 0.53 (Atar, 2007; Fidalgo et al., 2000; Kristjansson, 2001; Wang & Su, 2004; Zwick et al.,1993). In the current study, production of DIF-containing items was conducted by adding 0.05 and 0.10 values to the b parameter. Uniform DIF was obtained in this way, and the research was carried out.

2.1.2. *Manipulated conditions*

Swaminathan and Rogers (1990) stated that one of the factors that may affect statistical estimation in studies is the sample size and sample size ratios of the focus and reference groups. Additionally, it has been noted that nonparametric techniques have enhanced power to identify items with DIF when the sample sizes of the groups are not equal (Kristjansson et al., 2005). When reviewing the literature, it is observed that techniques based on CTT have sufficient power when there are at least 200-250 individuals per group (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). In techniques based on IRT, at least 1000 samples are expected (Shepard et al., 1981). Similarly, according to the results obtained from their study, Gök and Kelecioğlu (2014) state that more stable estimates can be obtained as the sample size increases, and sample sizes of 1000 and above would be sufficient. Additionally, considering large-scale testing, it is known that very large sample sizes are used. Considering these factors and this study examined techniques belonging to two different theories, and one of these theories is IRT, which requires large samples, sample sizes of 1800 and 3000 were chosen in the study. The ratio of the sample size of the reference group to the total sample size was manipulated as R/T1= 0.50 and R/T2= 0.75. Another parameter analyzed within the scope of DIF is test length. In other studies in the literature, it is seen that the number of items is generally set as 20, 40, and 80 (Atalay Kabasakal et al., 2014; Narayanan & Swaminathan, 1994; Price, 2014; Wang et al.,

2013). Test lengths of 20 can be considered as small, 40 as medium, and 80 as considerable. Regarding national exams in Turkey, it is known that Liselere Giriş Sınavı (LGS) consists of 10-20 items, while exams such as Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES), Yabancı Dil Sınavı (YDS), and Yükseköğretim Kurumları Yabancı Dil Sınavı (YOKDIL) consist of 80-100 items. Given the large-scale exams in Turkey, two different test lengths, 20 and 80, were utilized in this study. The selection of the proportion of items containing DIF was based on the rates of 0, 0.05 and 0.10, as utilized in the studies by Atalay Kabasakal et al. (2012) and Sunbul and Omur Sunbul (2016).

Within the scope of this research, data were generated with 25 replications for a total of (2x2x2x3) 24 conditions, including two sample sizes, two sample size ratios, two test lengths, and three ratios of DIF-containing items, since it is known that errors decrease after 25 replications in data generation (Harwell et al., 1996).

2.2. Data Analysis

In the current study, data were generated and analysed using R software. In this study, LR and MH techniques based on CTT and Lord's χ^2 and Raju's Area Measures based on IRT were used for dichotomous data. Statistical analysis was performed using the "difR" package in the R software, which includes DIF detection techniques and facilitates the analysis of these techniques both independently and comparatively (Magis et al., 2018). The "ltm" package, which also provides model estimation, was used for the analysis of the techniques based on IRT (Rizopoulos, 2022).

DIF detection techniques were assessed with two criteria: Type I error and statistical power ratios. Following the analyses, for Type I error, the proportion of items that did not exhibit DIF but showed an analysis output indicating DIF was calculated and reported. In terms of statistical power ratios, the percentage of items that actually contained DIF, whereas the result of the analysis indicated the presence of DIF, was reported. For the Type I error rates of the results obtained from the techniques used, the classification in Table 1 by Bradley (1978) was considered. In terms of statistical power ratios, the criterion stipulated that techniques displaying values exceeding 0.80 were deemed sufficient and high, whereas those falling below this threshold were considered inadequate (Atalay Kabasakal et al., 2014).

Table 1. *Bradley's classification of type I error rates.*

Level	Value Range
Conservative	$\alpha < 0.045$
Maintained	$0.045 < \alpha < 0.055$
Inflated	$\alpha > 0.055$

3. FINDINGS

The findings related to the first sub-problem " How do the Type I error rates of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques differ under conditions where the sample size is 1800 and 3000, the sample size ratio is 0.50 and 0.75, the number of items is 20 and 80, and there are no DIF items?" are presented in Table 2.

Table 2. *Findings related to the first sub-problem.*

Conditions			DIF Techniques				
The ratio of Items Showing DIF	Sample Size	Number of Items	R/ T	MH	LR	LORD χ^2	RAJU
0 (There is no DIF)	1800	20	0.50	0.92	0.92	0.86	0.74
			0.75	0.90	0.89	0.79	0.68
		80	0.50	0.89	0.88	0.88	0.79
			0.75	0.86	0.86	0.86	0.79
	3000	20	0.50	0.92	0.92	0.89	0.78
			0.75	0.89	0.89	0.86	0.78
		80	0.50	0.91	0.91	0.90	0.84
			0.75	0.91	0.91	0.91	0.85

It can be seen from the data in Table 2 that the Type I error ratios under different conditions for 20 and 80 items with a sample size of 1800 vary between 0.68 and 0.92 and between 0.79 and 0.89, respectively. For 20 items, when sample size ratios are considered, the MH technique yielded the highest error, and Raju's Area Measures technique yielded the lowest error. This pattern was consistent for 80 items and a sample size ratio of 0.50. When the sample size ratio was set as 0.75, the Lord's χ^2 technique showed the highest error with 0.862, while Raju's Area Measures had the lowest error with α value of 0.791.

When scrutinizing the Type I error rates for 20 and 80 items under diverse conditions, with sample sizes of 3000, values ranged from 0.78 to 0.92 and 0.84 to 0.91, respectively. Notably, in both conditions involving 20 and 80 items, it is observable that the MH and LR methods consistently exhibited the highest error rates, whereas Raju's Area Measures technique consistently manifested the lowest error rates, particularly when sample size ratios were altered. However, upon closer analysis, it is evident that the error rates of techniques other than Raju's Area Measures are very close to each other.

The findings related to the second sub-problem " How do the statistical power ratios of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques differ under the conditions in which the sample size is 1800 and 3000, the sample size ratio is 0.50 and 0.75, the number of items is 20 and 80, and the ratio of DIF- containing items is 0.05 and 0.1?" are presented in Table 3.

Table 3. *Findings related to the second sub-problem.*

Conditions			DIF Techniques				
The Ratio of Items Showing DIF	Sample Size	Number of Items	R/ T	MH	LR	LORD χ^2	RAJU
0.05	1800	20	0.50	0.80	0.68	0.64	0.52
			0.75	0.64	0.44	0.56	0.44
		80	0.50	0.11	0.12	0.41	0.39
			0.75	0.73	0.71	0.68	0.70
	3000	20	0.50	0.12	0.08	0.28	0.20
			0.75	0.20	0.12	0.12	0.04
		80	0.50	0.08	0.12	0.32	0.28
			0.75	0.85	0.78	0.60	0.58
0.1	1800	20	0.50	0.60	0.54	0.50	0.38
			0.75	0.20	0.14	0.34	0.32
		80	0.50	0.09	0.10	0.31	0.24
			0.75	0.59	0.51	0.53	0.51
	3000	20	0.50	0.14	0.08	0.18	0.20
			0.75	0.08	0.04	0.42	0.38
		80	0.50	0.77	0.76	0.60	0.53
			0.75	0.13	0.14	0.48	0.44

As shown in Table 3, the statistical power ratios ranged between 0.14 and 0.80 for a sample size of 1800 and 20 items under varying conditions, including sample size and the proportion of items with DIF. When the proportion of items containing DIF was set at 0.05, it became evident that the MH method had the highest power ratios, while Raju's Area Measures technique had the lowest power. However, under conditions where the DIF item proportion increased to 0.1, and sample size ratios stood at 0.50 and 0.75, MH and Lord's χ^2 exhibited the highest power with 0.60 and 0.34, respectively. Conversely, when considering sample size ratios of 0.50 and 0.75, Raju's Area Measures and the LR technique displayed the least powerful performance.

The statistical power ratios for a sample size of 1800 and 80 items ranged between 0.11 and 0.73 under varying conditions of sample size and DIF-containing item ratios. When the proportion of DIF-containing items was 0.05 and the sample size ratios ranged between 0.50 and 0.75, Lord's χ^2 and MH techniques had the highest power, while MH and Lord's χ^2 techniques had the lowest power, respectively. For 80 items, when the proportion of items containing DIF increased to 0.1, and the sample size ratios were 0.50 and 0.75, Lord's χ^2 with 0.31 and MH with 0.59 were found to have the highest power. In addition, when the sample size ratio was 0.50, MH with 0.09, and when the sample size ratio was 0.75, LR and Raju's Area Measures with 0.51 were found to have the lowest power.

For a sample size of 3000 and 20 items, statistical power ratios ranged between 0.04 and 0.42 under varying conditions of sample size and DIF-containing item ratios. Raju's Area Measures technique and MH technique were found to have the highest power ratios when the DIF-containing item ratio was 0.05 and the sample size ratios were 0.50 and 0.75, respectively. For the 0.50 sample size ratio, the LR technique with 0.08 and for a 0.75 sample size ratio, Raju's Area Measures technique with 0.04 were found to have the lowest power. For 20 items, Raju's Area Measures and Lord's χ^2 techniques were found to have the highest power when the DIF-containing item ratio was 0.1 and the sample size ratios were 0.50 and 0.75, respectively, while the LR technique was found to have the lowest power with 0.08 for a sample size ratio of 0.50.

For a sample size of 3000 and 80 items, statistical power ratios ranged between 0.08 and 0.85 under varying conditions of sample size and the ratio of items showing DIF. Lord's χ^2 and MH techniques had the highest power when the proportion of items with DIF was 0.05 and the sample size ratios were 0.50 and 0.75, respectively. MH technique had the lowest power with 0.08 for a sample size ratio of 0.50, and Raju's Area Measures technique had the lowest power with 0.58 for a sample size ratio of 0.75. When the proportion of items containing DIF was 0.1 and the sample size ratios were 0.50 and 0.75, the MH technique and Lord's χ^2 had the highest power, respectively, while Raju's Area Measures had the lowest power with 0.53 and MH with 0.13 when the sample size ratios were 0.50.

4. DISCUSSION and CONCLUSION

Within the scope of the current research, Type I error and statistical power ratios were scrutinized by conducting DIF analyses with the data generated in the 2PL model under different conditions. As a result of the analyses performed for this purpose, it was found that Type I error rates were higher than the nominal error level (0.05) in all conditions. Based on these results, it can be inferred that there were inflated Type I errors. For the 1800 sample size, when all conditions are analyzed, the MH technique displayed the highest Type I error rates. When the techniques used were assessed within the conditions, it was found that the MH and LR techniques produced very similar values and exhibited higher errors in the condition where the number of items was 20 compared to the condition where the number of items was 80 for a sample size of 1800. Both techniques yielded the highest error rate for the condition where the number of items was 20 and the R/T ratio was 0.50, and the lowest error rate for the condition where the number of items was 80 and the R/T ratio was 0.75. These results align with the

findings of other studies. Kim (2010) studied the Type I error rates of LR, MH, DFIT, and Lord's χ^2 techniques under different conditions, and found that all techniques tended to inflate Type I errors in conditions where the test length was shorter, the sample size was larger, and the focal and reference groups were equal. Similarly, Demars (2009) observed that MH and LR techniques produced inflated Type I error rates with shorter test lengths and larger sample sizes. On the other hand, in a study conducted by Dainis (2008), a comparison of the Type I errors of the IRT-OO, DFIT, MH, and LR techniques revealed that the error rates of the MH technique were at an acceptable nominal level. Gierl et al. (2000) emphasized that the Type I error rates of the MH and LR techniques can be around or even below the nominal error levels even if the sample size is small with equal ability distributions. Ankenmann et al. (1999) also stated that the LR technique yields Type I error rates at the nominal error level under general conditions, while the MH technique yields error rates above this nominal error level. However, in the current study, both techniques yielded Type I error rates much higher than acceptable error rates in all conditions where the sample size varied, and the findings of the current study do not support the previous research (Ankenmann et al., 1999; Dainis, 2008; Gierl et al., 2000). The main reason for this situation might be the influence of the sample size on the employed techniques. When examining studies in the literature, it has been observed, especially for MH and LR methods, that inflated Type I error values are obtained as the sample size increases (Demars, 2009; Kim, 2010; Magis & De Boeck; 2012). In the context of this study, considering the methods based on MTK were employed, sample sizes exceeding 1000 were chosen. It is believed that the high Type I error values in the obtained findings are attributable to this choice.

For a sample size of 3000, upon comprehensive examination of all conditions, it is evident that LR had the highest Type I error rates. However, it should be noted that the differences in error rates among the various techniques were modest, indicating that the MH technique also exhibited high Type I error rates. All the values obtained, however, showed an inflated Type I error, just as in the conditions generated with a sample size of 1800.

In the conditions set up to determine the Type I error rates, broadening the sample size from 1800 to 3000 increased the Type I error values in all conditions except three. The three mentioned conditions and techniques are as follows: For 20 items with an R/T ratio of 0.75, the MH technique yielded a lower Type I error rate. Again, when the R/T ratio was 0.50 for 20 items, the MH technique produced the same error rate value for 1800 and 3000 sample sizes, while the LR technique produced the same error rate value when the R/T ratio was 0.75. When the techniques used within the context of the research were analyzed within the conditions, MH and LR techniques obtained very similar values in all conditions, as in the 1800 sample size. Both techniques yielded the lowest Type I error rates in the condition where the number of items was 20, and the R/T ratio was 0.75. When the findings of Lord's χ^2 and Raju's Area Measures techniques were examined, it was concluded that both techniques yielded the highest Type I error rates under the condition where the number of items was 80, and the R/T ratio was 0.75. The lowest Type I error rates were obtained under the condition where the number of items was 20, and the R/T ratio was 0.75, just as in the techniques based on CTT.

Comparing the Type I error findings obtained based on the theories they are grounded in, it can be interpreted, similar to the study by Kan et al. (2013), that the techniques based on CTT and IRT are similar in terms of the conditions under which they are affected. Techniques in both theories display lower Type I error rates under the condition of 0.75, where the sample sizes of the focal and reference groups are not equal. Erdem Keklik (2012) stated that lower Type I error rates were obtained when the sample size ratios were 1:2 rather than 1:1. Comparison of the findings with those of other studies (Demars, 2009; Magis & DeBoeck, 2012) confirms that Type I error rates tend to increase with growing sample size; however, this does not appear to

be the case for some studies (Dainis, 2008; Sunbul & Omur Sunbul, 2016; Vaughn & Wang, 2010) that report that error rates decrease.

In the context of the first sub-problem, it is plausible to attribute the elevated Type I error rates observed in the various conditions examined within the research to the specific circumstances established. These conditions can be seen as a key contributing factor to the deviation of Type I error rates from the nominal error level. These results are likely related to the sample size and test length used. The sample sizes used in the study were not small; large samples were employed, and the finding that large sample sizes increase Type I error rates is supported by other studies in the literature (Dainis, 2008; Demars, 2009; Gierl et al., 2000; Magis & De Boeck; 2012). Regarding test length, higher Type I error rates were obtained in the analyses with 20 items compared to the conditions with 80 items. It is possible that these results were influenced by the techniques based on CTT, yielding higher rates due to the theoretical structure of the related techniques. The first point to be mentioned is that the MH and LR techniques are based on the observed score. These techniques use true scores as the matching variable in the process of determining the DIF. As Zwick et al. (1997) stated in the evaluation of the techniques used in this respect, the measurement errors of the techniques based on observed scores may decrease the reliability of the test. Therefore, this leads to true scores that deviate from the mean. Another problem that can be referred to is the use of total test scores, working with observed score-based techniques such as MH and LR with data produced following IRT models may cause inflated Type I errors.

The findings related to the second sub-problem, statistical power ratios, revealed that the MH technique yielded the highest statistical power ratio value of 0.80 as a result of the analyses conducted for a sample size of 1800. The MH technique displayed the highest statistical power ratios in all conditions except for three conditions. Surprisingly, it reached its highest value when the number of items was 20, not 80 as expected. Similar to the results of the study, Atalay Kabasakal et al. (2014) also found that the MH technique had the highest statistical power ratios in all conditions. In another study, it was also confirmed that the MH technique was the most powerful technique (Kristjansson et al., 2005). Regarding the results based on the other techniques included in the study, LR, Lord's χ^2, and Raju's Area Measures techniques reached the highest statistical power ratios when the number of items was 80, the ratio of DIF-containing items was 0.05, and the R/T ratio was 0.75. However, the obtained ratios were distributed around 0.70 and could not reach the desired value of 0.80. The fundamental reason for this situation can be attributed to the sample size, similar to the Type I error rates. When examining studies in the literature, findings have been obtained indicating that the statistical power of DIF techniques is also negatively affected by the sample size (Ankenmann et al., 1999; Gierl et al., 2000; Atalay Kabasakal et al., 2014).

An overall evaluation of the techniques based on the theory they are grounded in reveals that, in five of the eight conditions within the scope of the second sub-problem, MH and LR techniques based on CTT were found to have higher values, while Lord's χ^2 and Raju's Area Measures techniques based on IRT were found to have higher values in three of them. From this perspective, since the techniques do not show a regular difference from condition to condition, based on the current study's results, it can be interpreted that the techniques based on CTT provide higher statistical power ratios at a sample size of 1800. Furthermore, a consistent trend is observed across all employed techniques, where alterations in the proportion of items containing DIF from 0.05 to 0.10, in conditions with both 20 and 80 items, negatively impact and reduce the statistical power ratios. However, in contrast to earlier findings, Atar and Kamata (2011) emphasized that the statistical power ratio decreases as the proportion of items containing DIF decreases, especially in the LR technique, if the sample size is small. Another study compared the performance of LR and MH techniques and found that the statistical power

ratios increased as the DIF-containing item ratio increased (Hidalgo & Lopez-Pina, 2004). Therefore, a definitive interpretation regarding the DIF-containing item ratio cannot be made.

Following the analysis conducted with a sample size of 3000, wherein a comparison of statistical power ratios among the techniques was performed, it is noteworthy that the MH technique yielded the highest statistical power ratios, consistent with the findings observed for the 1800 sample size, with a statistical power ratio of 0.85. However, the technique displayed lower performance than the other techniques in most of the conditions compared to the 1800 sample size. At this point, it can be said that the MH technique is affected by the sample size (Ankenmann et al., 1999). In the literature, there are studies that further support the idea that MH and LR techniques decrease statistical power ratios with increasing sample size (Erdem Keklik, 2012; Vaughn & Wang, 2010). On the other hand, Atar (2007) and Jodoin and Gierl (2010) state that increasing sample size leads to an increase in both statistical power ratios and Type I error rates. This differs from the findings presented here.

As a whole, when considering the findings derived from the evaluation of techniques within the research across various conditions, it becomes evident that higher statistical power ratios were consistently achieved under nearly all conditions when the number of items was set to 80. Techniques obtained the highest statistical power ratios when the number of items was 80, the ratio of DIF-containing items was 0.05, and the R/T ratio was 0.75. In this respect, it can be interpreted that increasing the number of items has a promising effect on the statistical power ratios. Another point that can be mentioned here is the impact of the R/T ratio on the power ratios. Although higher statistical power ratios were obtained when the ratio was 0.75, as stated in the current study, it was also stated in different studies that there may be inconsistencies in the techniques in conditions where the sample sizes of the groups are not the same (Jodoin & Gierl 2010; Narayanan & Swaminathan, 1994). Therefore, a definite interpretation regarding the R/T ratio cannot be made.

In line with the comparison in terms of theories, it is apparent that, in general, higher statistical power rates were achieved at a sample size of 1800. This observation promotes the findings reported by Sunbul and Omur Sunbul (2016), where it was asserted that augmenting the sample size led to an increase in the statistical power ratios for both CTT and IRT-based techniques. In fact, there exist other studies, such as Atar (2007) and Narayanan and Swaminathan (1994), which concur with this notion, indicating that larger sample sizes tend to enhance power ratios. Nevertheless, it is notable that in the present study, the majority of techniques exhibited superior performance with a sample size of 1800, with only a few exceptions.

Based on the Type I error and power rates findings obtained in this study, researchers studying with smaller samples may consider using techniques based on CTT while those working with larger samples may prefer techniques based on IRT. Additionally, considering the R/F ratios used in this study, it was found that Type I error rates were lower under conditions where the sample sizes of the focal and reference groups were not equal. Therefore, if practitioners have the flexibility in forming groups, it is recommended to create groups with unequal sample sizes.

Declaration of Conflicting Interests and Ethics

The authors declare 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 authors.

Authorship Contribution Statement

Ayse Bilicioglu Gunes: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, Validation, and Writing-original draft. **Bayram Bicak**: Supervision.

Orcid

Ayse Bilicioglu Gunes https://orcid.org/0000-0002-1603-8631

Bayram Bicak https://orcid.org/0000-0003-0860-9374

REFERENCES

Ankenmann, R.D., Witt, E.A., & Dunbar, S.B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistics in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277–300. https://doi.org/10.1111/j.1745-3984.1999.tb00558.x

Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioğlu, H. (2014). Değişen madde fonksiyonunun belirlenmesinde MTK olabilirlik oranı, SIBTEST ve mantel- haenszel yöntemlerinin performanslarının (I. Tip hata ve güç) karşılaştırılması [Comparison of the performance (Type I error and power) of the IRT likelihood ratio, SIBTEST, and mantel-haenszel techniques in determining the differential item functioning]. *Educational Sciences: Theory & Practice, 14*(6), 2175- 2193.

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures* (FSU_migr_etd-0248) [Doctoral dissertation, Florida State University]. http://purl.flvc.org/fsu/fd/FSU_migr_etd-0248.

Atar, B., & Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. *Hacettepe University Journal of Education*, (41), 36–47.

Basman, M. (2023). A comparison of the efficacies of differential item functioning detection methods. *International Journal of Assessment Tools in Education, 10*(1), 145-159. https://doi.org/10.21449/ijate.1135368

Bradley, J.V. (1978). Robustness. *British Journal of Mathematical and Statistical Psychology, 31*(2), 144–152. http://dx.doi.org/10.1111/j.2044-8317.1978.tb00581.x

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

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

Cohen, A.S., Kim, S.H., & Wollack, J.A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15–26.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. CBS College Publishing.

Dainis, A.M. (2008). *Methods for identifying differential item and test functioning: An investigation of type 1 error rates and power* (3323367) [Doctoral dissertation, James Madıson University]. ProQuest.

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* (2), 149- 170.

Desa, Z.N. (2012). *Bi-factor multidimensional item response theory modeling for subscores estimation, reliability, and classification* (3523517) [Doctoral thesis, University of Kansas]. ProQuest.

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement*, *41*(3), 261- 270.

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829-848). Blackwell

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). Lawrence Erlbaum.

Ellis, B.B., & Raju, N.S. (2003). Test and item bias: What they are, what they aren't, and how to detect them. *Educational Resources Information Center (ERIC).*

Erdem Keklik, D. (2012). *İki kategorili maddelerde tek biçimli değişen madde fonksiyonu belirleme tekniklerinin karşılaştırılması: Bir simülasyon çalışması [Comparison of techniques in detecting uniform differential item functioning in dichotomous items: A simulation study]* (311744) [Doctoral thesis, Ankara University]. YÖK, Ulusal Tez Merkezi.

Fidalgo, A.M., Mellenberg, G.J., & Muniz, 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, W.H., & 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

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. The Annual Meeting of the American Educational Research Association.

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

Harwell, M., Stone, C.A., Hsu, T.C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. https://doi.org/10.1177/014662169602000 2

Hauck Filho, N., Machado, W.D.L., & Damásio, B.F. (2014). Effects of statistical models and items difficulties on making trait-level inferences: A simulation study. *Psicologia: Reflexão e Crítica, 27*(4), 670- 678. https://doi.org/10.1590/1678-7153.201427407

Hidalgo, M.D., & Lopez-Pina, J.A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and mantel-haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915. https://doi.org/10.1177/0 013164403261769

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Ed.), *Test validity* (pp.129-145). Erlbaum.

Jodoin, M.G., & Gierl, M.J. (2010). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Psychological Measurement, 14* (4), 329- 349. https://doi.org/10.1207/S15324818AME1404_2

Kan, A., Sünbül, Ö., & Ömür, S. (2013). 6. - 8. sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi [Analysis of 6th - 8th grade placement exams subtests' differential item functioning by various methods]. *Mersin University Journal of the Faculty of Education, 9*(2), 207- 222.

Karasar, N. (2010). *Bilimsel araştırma yöntemleri [Research methods]*. Nobel Publication.

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* [Doctoral thesis, Georgia State University]. https://doi.org/10.57709/1642363

Koğar, H. (2018). An examination of parametric and nonparametric dimensionality assessment methods with exploratory and confirmatory models. *Journal of Education and Learning, 7*(3), 148-158. 10.5539/jel.v7n3p148

Kristjansson, E. (2001). *Detecting DIF in polytomous items: an empirical comparison of the ordinal logistic regression, logistic discriminant function analysis, Mantel, and*

generalized Mantel Haenszel procedures [Unpublished Doctoral Dissertation]. University of Ottawa.

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B.D. (2005). Comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953. https://doi.org/10.1177/0013164405275668

Lim, R.G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*(2), 164-174. https://doi.org/10.1037/0021-9010.75.2.164

Lord, F.M. (2012). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates.

Magis, D., Beland, B., & Raiche, G. (2018). difR: Collection of methods to detect dichotomous differential item functioning (DIF). https://cran.r-project.org/web/packages/difR/difR.pdf

Magis, D., & De Boeck, P. (2012). A robust outlier approach to prevent type I error inflation in differential item functioning. *Educational and Psychological Measurement, 72*(2), 291-311.

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

Narayanan, P., & Swaminathan, H. (1994). Performance of the mantel-haenszel and simultaneous item bias procedures for detecting differential. *Applied Psychological Measurement, 18*(4), 315-328. https://doi.org/10.1177/014662169401800403

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

Osterlind, S.J. (1983). *Test item bias*. Sage Publications.

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

Patton, M.Q. (1990). *Qualitative evaluation and research methods.* Sage Publications, Inc.

Price, E.A. (2014). *Item Discrimination, model-data fit, and type I error rates in DIF detection using lord's χ^2, the likelihood ratio test, and the mantel-haenszel procedure* [Doctoral thesis, Ohio University]. OhioLINK Electronic Theses and Dissertations Center. http://rave.ohiolink.edu/etdc/view?acc_num=ohiou1395842816

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

Rizopoulos, D. (2018). Latent trait models under IRT. https://cran.r-project.org/web/packages/ltm/ltm.pdf

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and mantel-haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 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*(2), 215-230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Samuelsen, K.M. (2005). *Examining differential item functioning from a latent class perspective* (3175148) [Doctoral thesis, University of Maryland]. PreQuest.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*(4), 317-375. https://doi.org/10.3102/10769986006004317

Simon, J.L. (1978). *Basic research methods in social science*. Random House.

Sünbül, Ö., & Ömür Sünbül, S. (2016). Değişen madde fonksiyonunun belirlenmesinde kullanılan yöntemlerde I. tip hata ve güç çalışması [Type I error and power study in methods used to determine differential item functioning]. *Elementary Education Online, 15*(3), 882- 897.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://www.jstor.org/stable/1434855

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

Wang, W.C., & Su, Y.H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*, 450–481.

Wang, W., Tay, L., & Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Applied Psychological Measurement, 37*(4), 316- 335. https://doi.org/10.1177/0146621613476156

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

Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.

Zwick, R., Thayer, D.T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items.* Educational Testing Service.