





<http://www.tayjournal.com>

<https://dergipark.org.tr/tr/pub/tayjournal>

Determination of Type I Error and Power Rate in Differential Item Functioning by Several Methods*

 Şeyma Erbay Mermer, Phd.
Bilecik Şeyh Edebali University, Türkiye
sey.erbay@gmail.com
Orcid ID: 0000-0002-7747-9545

 Yasemin Kuzu, Phd., Corresponding Author
Kırşehir Ahi Evran University, Türkiye
yaseminkuzu@ahievran.edu.tr
Orcid ID: 0000-0003-4301-2645

 Hülya Kelecioğlu, Prof. Dr.
Hacettepe University, Türkiye
hulyakelecioğlu@ahievran.edu.tr
Orcid ID: 0000-0002-0741-9934

Article Type: Research Article
Received Date: 09.09.2023
Accepted Date: 23.10.2023
Published Date: 30.10.2023

Plagiarism: This article has been reviewed by at least two referees and scanned via a plagiarism software
Doi: 10.29329/tayjournal.2023.610.09

Citation: Erbay-Mermer, Ş., Kuzu, Y., & Kelecioğlu, H. (2023). Determination of type I error and power rate in differential item functioning by several methods. *Türk Akademik Yayınlar Dergisi (TAY Journal)*, 7(3), 902-921.

*The study was presented in 4th International Academic Research Congress 2018 (INES 2018), Antalya, Türkiye, Oct, 30- Nov 3, 2018.

Abstract

In this study, the methods based on Classical Test Theory and Item Response Theory were used comparatively to determine Type I error and power rates in Differential Item Functioning. Logistic regression, Mantel-Haenszel, Lord's χ^2 , Breslow-Day and Raju's area index methods were used for the analyses, which were conducted using the R programming language. Determination of Type I error and power rates of these methods under certain conditions was carried out by simulation study. For data generation, analyzes were made under eight conditions in total by examining different sample sizes and DIF rates created with the WinGen 3 program. The results of the study indicate that, in general when the ratio of items containing DIF increased, Type I error increased and the power ratio decreased. Among the methods based on Item Response Theory, Lord's χ^2 and Raju's area index methods gave better results than other methods with low error and high power.

Keywords: IRT, DIF, Type I error, power.

Introduction

One of the most important issues emphasized by measurement and evaluation in education and psychology is large-scale exams at national (Public Personnel Selection Exam [KPSS], Academic Personnel and Graduate Education Entrance Exam [ALES], etc.) and international (Test of English as a Foreign Language [TOEFL], Programme for International Student Assessment [PISA], Trends in International Mathematics and Science Study [TIMMS], etc.) levels. The evaluation and interpretation of these exams, on the results of which important decisions are made about individuals and countries, are of great importance. Therefore, these exams should enable valid interpretations (Clauser & Mazor, 1998). In other words, in order for the decisions made on the test results to reflect the truth and the scores obtained from the test to reflect the actual performance of individuals, the measurements made need to be valid. Validity, which is a degree of theory and evidence that helps to demonstrate the accuracy of interpretations made on test scores or decisions made as a result of test scores (American Educational Research Association [AERA] et al., 1999), is one of the most important features that tests and other measurement tools should have. Tests should measure the construct with the same accuracy for all individuals without being affected by variables other than the measured trait (Sireci & Rios, 2013). Although validity is affected by many factors, the most important threats to tests are item and test bias (Clauser & Mazor, 1998). Item bias emerged in 1910 when Alfred Binet administered an intelligence test to children from low socioeconomic backgrounds. When Binet analyzed the test items, he found that some of the items measured cultural traits in addition to intelligence and deemed it appropriate to remove them from the test. In 1912, Stern showed in his study that different results emerged in different subgroups. Later on, the idea of preparing tests for a single group developed (Camilli & Shepard, 1994). Cleary introduced the concept of test bias by finding that predicted criterion scores were too high or too low in subgroups (Lee, 2003).

Bias is the interference of other variables (gender, school type, ethnicity, etc.) with the characteristics of individuals that we want to measure and leads to systematic errors that distort the results obtained from the test and the interpretations made based on these results (Gierl et al., 1999). The presence of biased items in a test that favor one group is a significant threat to the validity of the test (Kane, 2006; Messick, 1989). Therefore, it is very important to prepare the test in a way that does not give advantage to any subgroup (Gök et al., 2014). The first step in examining whether the items in a test are biased is to determine whether there is a differential item functioning (DIF) in the relevant

items. DIF is the difference in the probability of answering an item correctly by individuals with the same ability level according to their subgroups (Embretson & Reise, 2000; Hambleton et al., 1991). According to Zumbo (1999), DIF explains the differences in the probability of answering the item correctly for individuals in different groups in a comparison study for the level of ability targeted to be measured by the item. DIF analyses are a prerequisite for identifying biased items in a test, but they are also evidence for the validity of the test (Embretson, 2007). While a biased item can definitely be said to contain DIF, the presence of DIF in an item is not enough to say that the item is biased. For an item that is found to contain DIF, it can only be concluded that it is biased with expert opinion (Zumbo & Gelin, 2005), so it requires a qualitative evaluation based on item bias detection (Ellis & Raju, 2003; Furlow et al., 2009; Sireci & Allalouf, 2003).

DIF is considered in two different ways: uniform and non-uniform DIF. In uniform DIF, for an item containing DIF, the same group performs lower or higher at each ability level. Uniform DIF occurs when there is no interaction between ability level and group membership in terms of individuals' item performance. Therefore, in the presence of uniform DIF, only item difficulty parameters differ between groups. The fact that the differentiation in performance between groups is uniform across the entire ability domain means that the item contains a uniform DIF (Penfield & Lam, 2000).

In non-uniform DIF, for an item showing DIF, one group performs better at some ability levels and the other group performs better at some ability levels. If there is an interaction between ability level and group membership in terms of item performance of individuals, we can talk about non-uniform DIF (Camili & Shephard, 1994). In non-uniform DIF, unlike uniform DIF, item discriminations are different between subgroups. However, this is not true for the item difficulty parameter (Turhan, 2006).

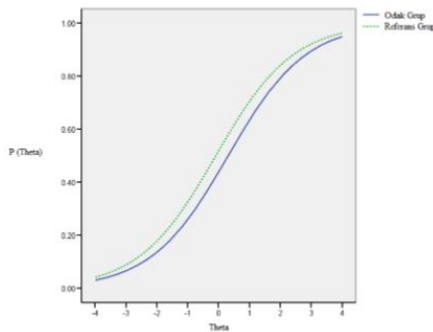


Figure 1. Uniform DIF functioning

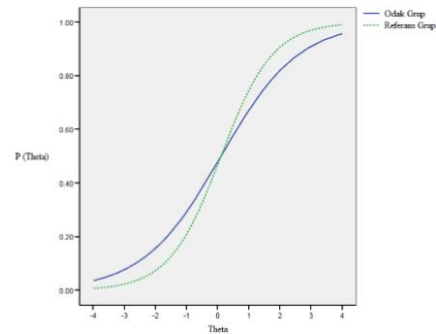


Figure 2. Non-uniform DIF functioning

When Figure 1 and Figure 2 are examined, it is seen that while the substance characteristic curves of the groups do not intersect each other for the uniform DIF, the substance characteristic curves overlap in the non-uniform DIF.

Although the presence of DIF threatens the validity of a test, it is a mistake to exclude an item from the test only because it contains DIF (Zumbo & Gelin, 2005). First, items containing DIF are subjected to statistical analyses, and as a result of the analyses, they are divided into A, B and C categories in terms of DIF level (Zieky, 1993). Here, category A means non-significant DIF, category B means moderate DIF and category C means high level DIF. Items in category A do not need to be removed from the test; items in category B can be used when they are important for the test. However, items in category C should be removed from the test if they are not very important for the trait to be measured (Zwick, 2012). Many DIF determination methods are mentioned in the literature. However, Karami and

Nodoushan (2011) stated that different methods determine DIF in different items for the same test, so it is not correct to analyze according to only one method and interpret the results according to only one method. Accordingly, if an item contains DIF according to more than one method, it is supported by different methods that the item has DIF.

DIF Determination Methods

There are many methods to determine whether items in a test contain DIF or not. Some of these methods, which are classified as based on Classical Test Theory (CTT) and Item Response Theory (IRT), are given below.

Table 1. DIF detection methods based on CTT and IRT

CTT Based Methods	IRT Based Methods
Variance Analysis	SIBTEST
Mantel-Haenszel	Hierarchical Generalized Linear Modeling (HGLM)
Logistic Regression	Item Characteristic Curve
	MIMIC
	Lord's χ^2 Test
	Raju's Area Index
	IRT Likelihood Ratio

While DIF determination is made on the basis of observed scores in CTT-based methods, in IRT-based DIF determination, ability estimation can be made independently of test items.

Mantel-Haenszel (MH) Method

It is a method used to determine uniform or non-uniform DIF and is based on the difference of "odd" values obtained from the scores of two groups at the same ability level (Mertler & Vannatta, 2005). The "odd" value is the ratio between the probability of an event occurring and the probability of it not occurring. $(0, \infty)$ values in the range. Therefore, the results are difficult to interpret. As a solution to this problem, it has been suggested to use the (delta) statistic, which corresponds to -2.35 times the natural logarithm of the MH statistic and is interpreted as the DIF effect size (Camilli & Shepard, 1994). The values to be taken as a criterion for interpreting the effect size (Dorans & Holland, 1993) are given below.

Table 2. DIF levels according to MH method

Level	Value	Amount of DIF
A	$\Delta MH < 1$	None or negligible
B	$1 \leq \Delta MH < 1.5$	Medium level
C	$\Delta MH \geq 1.5$	High level

According to Table 2, if the DIF effect size is 1.5 or above, the item must be removed from the test.

Logistic Regression (LR) Method

It is a parametric method based on the observed score, and in this method, the presence of DIF is examined over the responses of individuals to the item and the total score with the help of the models established. For the interpretation of the degree of DIF, the effect size ΔR^2 statistic is calculated. It is seen that different criteria are used for effect size (Jodoin & Gierl, 2001; Çepni, 2011). However, Zumbo and Thomas (1996) gave the following criteria for interpreting the DIF effect size.

Table 3. DIF levels according to LR method

Level	Value	Amount of DIF
A	$\Delta R^2 < 0.13$	None or negligible
B	$0.13 \leq \Delta R^2 < 0.26$	Medium level
C	$\Delta R^2 \geq 0.26$	High level

Lord's χ^2 Statistics

In this IRT-based method, item parameters and covariances for subgroups are first calculated. Then, these estimated parameters are scaled and Lord's χ^2 statistic is calculated (Camilli & Shepard, 1994). Finally, the presence of DIF is decided by comparing the observed values with the critical value (Osterlind, 1983).

Raju's Area Index

Determination of DIF by this method is based on substance characteristic curves. First, item characteristic curves of subgroups are drawn for an item. If there is a difference between the item characteristic curves, the presence of DIF is mentioned (Camilli & Shepard, 1994).

When DIF studies are analyzed in general, it can be said that domestic and international studies focus on comparing different DIF determination methods, identifying possible sources of bias, and calculating Type I error and power ratio. Li, Qin, and Lei (2017) used the hierarchical DIF approach to examine the effect of teachers' teaching performance at the item level. The items were taken from the TIMSS 2011 4th grade mathematics test in the United States. In the context of teaching responsiveness, individuals were grouped according to whether or not they received instruction on the content tested by a given item. Ultimately, seven of the 34 TIMSS items included in the study showed instructional responsiveness regardless of whether covariates were controlled for. Controlling for the overall scores for these seven items, students who received the relevant instruction were significantly more likely to respond correctly to these items than those who did not. Jeon, Rijmen, and Rabe-Hesketh (2013) provided a general overview of a multigroup bifactor model for assessing DIF in item-set-based tests. The proposed model has four main features. First, it accounts for group differences in the multidimensional latent space. Second, it relaxes the assumption that all dimensions are independent from the assumption that certain dimensions are conditionally independent of the overall dimension. Third, the proposed method is flexible and can be applied to a variety of measurement models, including item-set and quadratic models for dichotomously and multiply scored responses. Fourth, the model can efficiently predict for large problems with many items, item sets, and examinees using the full information ML method. The simulation study shows that ignoring group differences can bias item parameter estimates. In this case, especially DIF estimation can be biased.

Hou, la Torre and Nandakumar (2014) investigated the effectiveness of the Wald test in detecting both uniform and non-uniform DIF in the DINA model through a simulation study. The results of this study show that the Wald test has low Type I error rates. Furthermore, the performance of the Wald test in detecting uniform DIF was compared with the traditional MH and SIBTEST methods. The results of the comparison study show that the Wald test outperforms the MH and SIBTEST methods. Finally, the strengths and limitations of the proposed method are discussed and suggestions for future work are presented. Jodoin and Gierl (2001) discussed effect size measurement and classification for the LR DIF method in their study. A simulation study was conducted to determine whether effect size affects Type I error and power rates for the LR DIF method across sample sizes, ability distributions,

and the percentage of DIF items in a test. The results showed that the inclusion of an effect size measure when using a large sample - can significantly reduce Type I error rates, but there will also be a reduction in power rates. Chen et al. (2014) used hierarchical generalized linear models (HGLM) to assess DIF in their study. They described their new method as follows: identify the item that has almost no DIF in the test, such that the two groups can have different means and the other items can be evaluated for DIF. In this context, Simulation Study 1 compared various methods based on HGLMs for selecting DIF-free items. In Simulation Study 2, items rated as DIF-free were taken as anchors and other items were evaluated for DIF. This new method was compared with the traditional method based on HGLMs, where the two groups are assumed to have equal means in terms of Type I error rate and power ratio. As a result, the new method outperformed the traditional method when the means of the two groups were different.

Walker and Gocer-Sahin (2016) aimed to determine how much the secondary ability distributions should change before DIF is detected. Two-dimensional binary data sets were generated using a compensatory multidimensional IRT model and the correlation between the dimensions was systematically increased, while the mean difference in the second dimension was gradually changed between the reference and focus group. SIBTEST, MH and LR methods were used to test the DIF. The results showed that even with a very small mean difference on the second dimension, smaller DIF would be detected than in previous research. Although the smallest mean difference considered in this study was 0.25, statistically significant differences were found between the reference and focus groups in the subtest scores of the items measuring the secondary dimension. (2017) extend the MIMIC interaction model to detect DIF in the context of multidimensional IRT modeling and examine the performance of the multidimensional MIMIC interaction model with respect to Type I error and power rates under different simulation conditions. Simulation conditions include DIF type and size, test length, correlation between latent traits, sample size, and latent mean differences between focal and reference groups. The results of this study show that the power rates of the multidimensional MIMIC interaction model under uniform DIF conditions are higher than the power rates of non-uniform DIF conditions. As anchor item length and sample size increase, the power to detect DIF increases. Although the multidimensional MIMIC interaction model was found to be a very useful tool for identifying uniform DIF, its performance in detecting non-uniform DIF seems to be questionable.

Kabasakal and Kelecioğlu (2015) examined the effect of DIF items on test equalization in unidimensional and multidimensional IRT. In the study, the performance of three different equalization methods under 24 different simulation conditions were examined. The variables examined are sample size, test length, DIF size and test type. Multidimensional Item Response Models with DIF factors as parameters were compared with Stocking-Lord and simultaneous calibration methods, and differences were found in the performance of the methods in the conditions. Accordingly, multidimensional item response models were able to identify DIF items in a single analysis, apply equating methods and eliminate the bias caused by DIF. In addition, an increase in test length and sample size generally had a positive effect on item response models. When item response model-based methods were considered, it was found that separate calibration methods were more affected by the presence of DIF items than simultaneous calibration. This effect is most significant when DIF items are present in the common test and the DIF size is C.

When the literature is examined, studies comparing DIF determination methods are also found. Erdem-Keklik (2012) compared MH, LR and IRT Likelihood Ratio methods in determining the uniform DIF in a simulation study involving a sample size of ability distribution. The results showed that Item Response Theory Likelihood Ratio is better than other methods in controlling Type I error in different ability distributions. Şahin (2017) compared the objective (MH, LR and SIBTEST) and subjective methods used in DIF detection. The highest agreement regarding the presence of DIF was found between MH and SIBTEST methods (0,90; $\kappa = 0,79$), while the lowest agreement in objective methods was obtained between LR and SIBTEST methods (0,75; $\kappa = 0,50$) was found. The agreement between objective and subjective methods was found to be moderate. Awuor (2008), in his study comparing SIBTEST and MH methods, concluded that MH method is better than SIBTEST method in controlling Type I error at different sample sizes. Zheng et al., (2007) compared MH, LR and SIBTEST methods and concluded that DIF direction and magnitude are consistent in all methods. In their study, Kan, Sünbül, and Ömür (2013) compared the DIF determination methods of Transformed Item Difficulty, MH, LR, Lord's χ^2 and Raju's domain measure methods. As a result, while most of the items in the subtests did not contain DIF in the CTT-based methods, many items contained DIF in the IRT- based methods. The CTT-based methods were similar within themselves and the IRT-based methods were similar within themselves.

In the studies on DIF methods, it is seen that Type I error and power rates are studied under different conditions. However, it is noteworthy that similar methods are used in these studies. In this study, Type I error and power rates of methods based on CTT and IRT were studied. In this context, conditions and levels of conditions were changed in determining DIF. Therefore, the methods used differ from other studies in terms of the conditions and levels of the conditions. The following research problems were sought to be answered in the study.

1. What are the Type I error rates of LR, MH, Lord's χ^2 , Breslow-Day and Raju's area index methods under different conditions?
2. What are the power rates of LR, MH, Lord's χ^2 , Breslow-Day and Raju's area index methods under various conditions

Method

Research Model

In this study, the Type I error and power rates of the CTT and IRT methods used in DIF determination are comparatively analyzed under different conditions. Since it is a research that will contribute to the existing knowledge in the literature by providing information about the performance of the methods, the research model of the study is basic research.

Data Generation

In the present study, simulative data is used to determine the Type I error and power rates of different methods used in DIF detection under certain conditions. The same data set was used to determine the Type I error and power rates. WinGen 3 software was used for data generation. In order to calculate Type I error rates and power obtained under different conditions from DIF determination methods, data showing uniform DIF were generated for the reference and focus groups with sample sizes of 1000 (O=500, R=500), 1500 (O=500, R=1000), 1500 (O=1000, R=500), 2000 (O=1000, R=1000).

The number of items for each analysis was set as 25. A two-parameter logistic model was used in data generation. Item parameter was obtained from a normal distribution with a mean of 0.8 and a standard deviation of 0.02. Parameter b was randomly drawn from a uniform distribution with a minimum value of -3 and a maximum value of +3. The values of the ability distribution of the individuals were obtained from a normal distribution with a mean of 0 and a standard deviation of 1. In this way, the DIF item was obtained by creating a difference in difficulty levels for the reference and focus groups without differentiating the ability distribution of the individuals. Item parameters were common for the reference and focus groups. Data were generated as the proportion of items containing DIF (12%, 20%) and DIF level ($b = 0.75$). The amount of DIF, 0.75, was added to the parameter b as many times as the number of items desired to contain DIF.

A total of eight conditions were analyzed with four samples and two DIF rates generated by the simulation study. In both the Type I error study and the power study, 20 repetitions were performed for each condition, which was formed by crossing the levels of the criteria. In total, 160 replications were performed for all cases. The simulative data used in the study was generated with code written using the R.3.0.1 program.

Data Analysis

For the detection of DIF, a comparative analysis of methods based on CTC and MTC was performed. The "difR" package was used for the analysis. CTT and IRT based LR, Mantel-Haenszel, Lord's χ^2 , Breslow-Day (BD) and Raju's area index methods were used. The R.3.0.1 program and the "difR" package were used for the analysis of DIF detection. "difR" is an R package that contains indices for changing matter function detection methods (Magis et al., 2015).

In Type I error analyses, the proportion of items labeled as DIF when they did not contain DIF was determined after 20 replicates for each condition. In power analyses, the proportion of items labeled as containing DIF when they were not was determined.

Ethical Permits of Research

In this study, all the rules to be followed within the scope of the "Directive on Scientific Research and Publication Ethics of Higher Education Institutions" were followed. None of the actions specified under the second section of the Directive, "General Actions Contrary to Scientific Research and Publication Ethics", have been carried out.

Ethics Committee Permission Information:

This research does not require ethics committee permission.

Findings

Findings for Type I Error Rates

The results obtained from Type I error rates for all conditions are given in Table 4.

Table 4. Mean Type I error rates according to sample and item rates with DIF

Sample (F-R)	Item Rates with DIF (%)	LR	Lord's χ^2	MH	BD	Raju
500-500	12	0.141	0.072	0.159	0.031	0.041
	20	0.185	0.105	0.140	0.015	0.065
500-1000	12	0.186	0.095	0.177	0.041	0.045
	20	0.220	0.145	0.195	0.020	0.090
1000-500	12	0.177	0.090	0.163	0.040	0.050
	20	0.205	0.110	0.175	0.040	0.055
1000-1000	12	0.218	0.109	0.210	0.050	0.050
	20	0.270	0.150	0.220	0.040	0.075

When Table 4 is examined, the minimum value for LR is 0.141 and the maximum value is 0.27; Lord's χ^2 The minimum value for MH was 0.072 and the maximum value was 0.15; the minimum value for MH was 0.14 and the maximum value was 0.22; the minimum value for BD was 0.015 and the maximum value was 0.05; and the minimum value obtained from Raju's area index methods was 0.041 and the maximum value was 0.09.

The lowest Type I error rate was determined by averaging the repetition rates. When the results are examined, it is seen that in general, the Type I error rate is highest in LR and Lord's χ^2 methods and the least in the BD method. These findings suggest that MH and LR Type I error rates are similar when the ability distributions of the focal and reference groups are the same. LR is Lord's χ^2 and MH methods, Type I error rates fluctuate according to the sample size and increase as the sample size increases; in the BD method, they are highest when the sample size is large and the DIF rate is low (12%), while in Raju's area index method, the reference group is more than the focal group and the DIF rate is high (20%). The change graphs of Type I error rates according to the 12% and 20% DIF containing material conditions are given in Figure 3 and Figure 4.

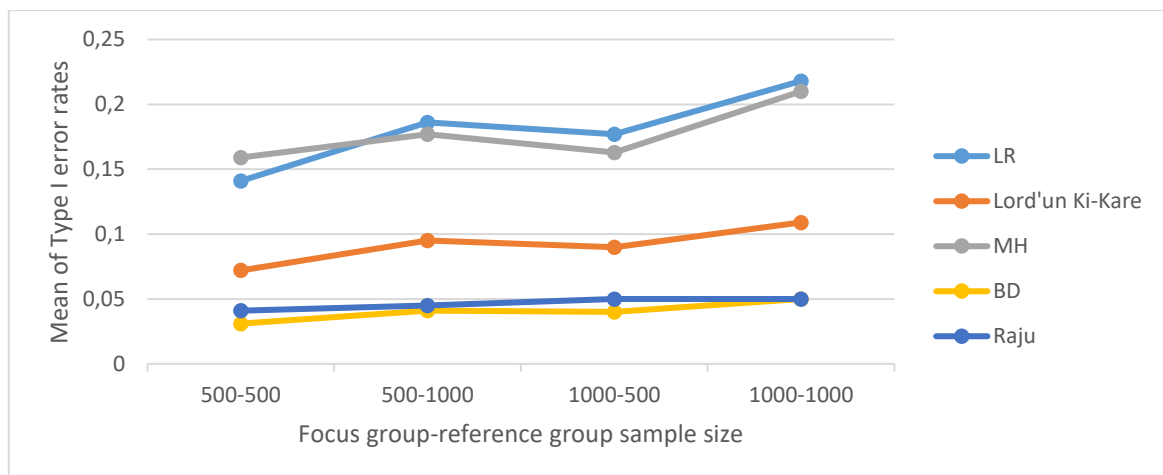


Figure 3. Type I error rates of the methods when the DIF content is 12%.

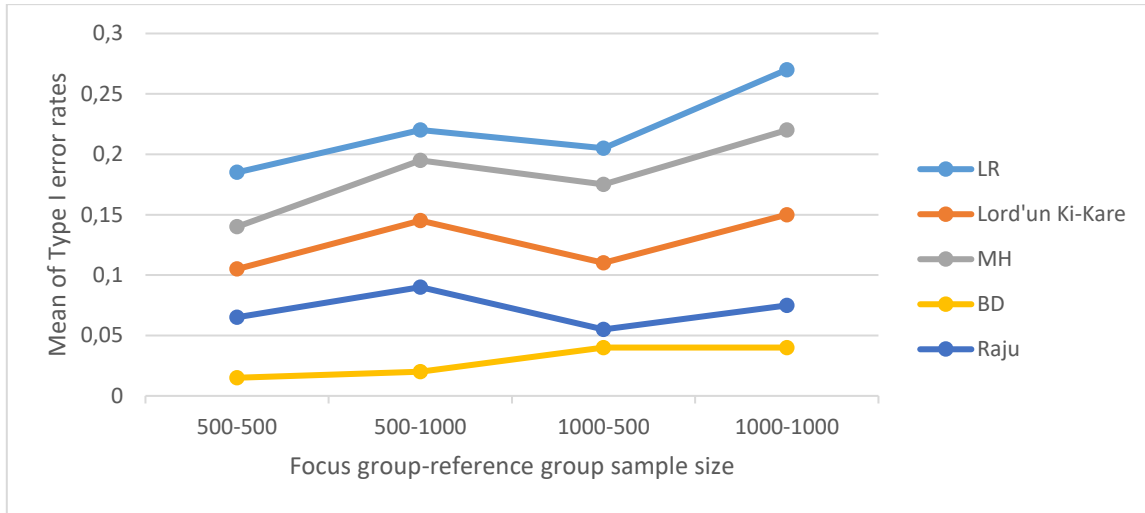


Figure 4. Type I error rates of the methods when the DIF content is 20%.

Findings for Power Rates

The results obtained from the power rates for all conditions are given in Table 5.

Table 5. Mean power rates according to sample and item rates with DIF

Sample (F-R)	Item Rates with DIF (%)	LR	Lord's χ^2	MH	BD	Raju
500-500	12	1.00	1.00	0.90	0.23	1.00
	20	0.90	0.94	0.76	0.52	0.88
500-1000	12	1.00	1.00	1.00	0.36	1.00
	20	0.94	0.98	0.78	0.38	0.98
1000-500	12	1.00	1.00	1.00	0.03	1.00
	20	0.98	0.98	0.74	0.68	0.90
1000-1000	12	1.00	1.00	1.00	0.20	1.00
	20	0.98	0.98	0.80	0.70	0.98

When Table 5 is examined, the minimum value for LR is 0.90 and the maximum value is 1. χ^2 The minimum value for MH was 0.94 with a maximum value of 1, the minimum value for MH was 0.74 with a maximum value of 1, the minimum value for BD was 0.03 with a maximum value of 0.70, and the minimum value obtained from Raju's area index methods was 0.88 with a maximum value of 1. The lowest power ratio was determined by averaging the repetition rates. When the results are examined, it is seen that the power ratio is highest for the 12% DIF material ratios and lowest for the BD method.

Looking at the results from the conditions, it is clear that the power ratio χ^2 method is the highest and the lowest in the BD method. In the 2000 sample of 12% DIF items, the power ratio is generally the highest for all methods and the lowest ratio is 0.2 in the BD method. It was also observed that the power ratio in the BD method was the highest in the case of a large sample and 20% DIF items. The variation graphs of the power ratios according to the 12% and 20% DIF containing material conditions are given in Figure 5 and Figure 6 below.

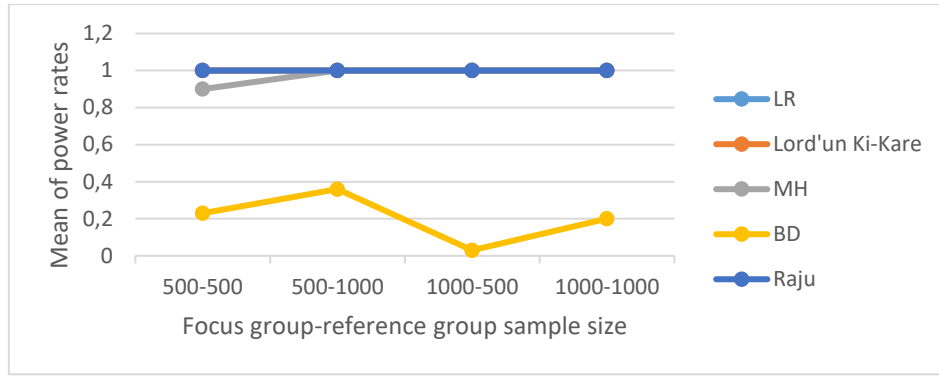


Figure 5. Power rates of the methods when the DIF content is 12%.

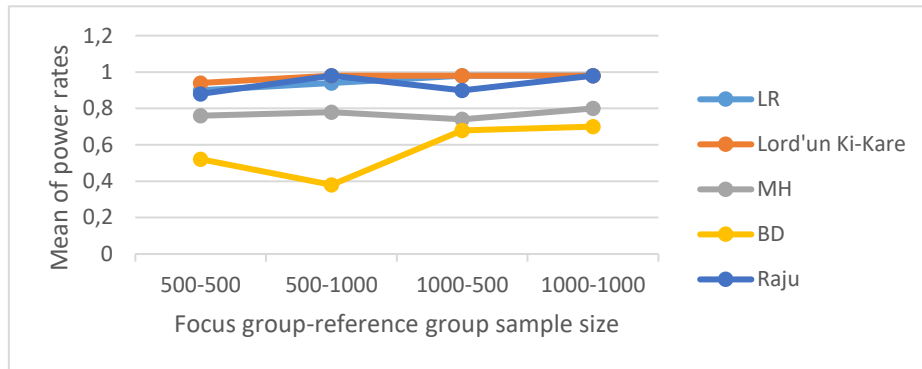


Figure 6. Power rates of the methods when the DIF content is 20%.

Discussion and Conclusion

Within the scope of the present study, the cases where the proportion of DIF-containing items differed and the sample sizes were observed for the changes in the means of Type I error and power ratios. For the observations, the LR and LRT methods were utilized. LR, Lord's χ^2 Type I error rates in BD and MH methods vary according to the sample size and increase as the sample size increases. In the BD method, the highest error rate was observed when the sample was large and the DIF rate was low (12%), while in Raju's area index method, the highest error rate was observed when the reference group sample was larger than the focal group sample and the DIF rate was high (20%). These results are in line with similar studies in the literature (Ankenmann et al., 1996; Atar & Kamata, 2011; Gierl et al., 2000; Rogers & Swaminathan, 1993; Roussos & Stout, 1996; Vaughn & Wang 2010).

In the condition where the sample size is the highest and the proportion of DIF items is the lowest; LR, Lord's χ^2 It is seen that the power ratio is the highest in MH and Raju's area index methods, while the lowest ratio is 0.2 in the BD method. Sünbül and Sünbül (2016), in their study on simulative data, stated that the power ratios of the methods increased with the increase in sample size. At the same time, the decrease in the power ratios of the methods with the increase in the proportion of DIF items supports this study.

In the results, the increase in the proportion of DIF items generally led to an increase in Type I error and a decrease in power. Similarly, Erdem-Keklik (2014) compared the Type I error and power ratios of MH and LR methods and found that Type I error was high in large sample sizes. In the analysis results, Lord's method, which is one of the ICA methods with low error and high power χ^2 and Raju's area index methods were found to give better results than the others.

Recommendations

In this study, the uniform DIF was analyzed. In addition to similar studies, non-uniform DIF can be examined by changing the ability parameters. In the study, the number of items was determined as 25, and the change in error and power ratios can be examined with different DIF determination methods when the number of items is less or more. In addition, the performance of DIF identification methods can be examined by manipulating variables such as sample size, sample size ratio, and item discrimination.

References

- American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- 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.
- 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.
- Awuor, R. A. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-based monte carlo study with sibttest and Mantel-Haenszel procedures* [Unpublished master thesis]. Virginia Polytechnic Institute and State University.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. SAGE.
- Chen, J. H., Chen, C. T., & Shih, C. L. (2014). Improving the control of Type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement, 38*(1), 18–36.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning items. *Educational Measurement: Issues and Practice, 17*, 31–44.
- Çepni, Z. (2011). *Değişen madde fonksiyonlarının sibttest, Mantel Haenszel, lojistik regresyon ve madde tepki kuramı yöntemleriyle incelenmesi* [Differential item functioning analysis using sibttest, Mantel Haenszel, logistic regression and item response theory methods]. [Unpublished master thesis], Hacettepe University.
- Ellis, B., & Raju, N. (2003). *Test and item bias: what they are, what they aren't, and how to detect them* (ED480042). ERIC. <https://eric.ed.gov/?id=ED480042>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.
- Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure?. *Educational Researcher, 36*(8), 449-455.
- 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]. (Tez No.311744). [Doctoral dissertation, Ankara University], National Thesis Center.
- Furlow, C. F., Ross, T. R., & Gagné, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement, 33*(6), 441–464.
- Gierl, M. J., Rogers, W. T., & Klinger, D. A. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *Alberta Journal of Educational Research, 45*(4), 353–376.
- 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]. *The Annual Meeting of the American Educational Research Association (AERA)*, New Orleans, Louisiana, USA.
- Gök, B., Kabasakal, K. A., & Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi [Analysis of attitude items in PISA 2009 student questionnaire in terms of differential item functioning based on culture]. *Journal of Measurement and Evaluation in Education and Psychology, 5*(1), 72–87.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE.
- Hou, L., de la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98–125.

- Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, 38(1), 32–60.
- 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.
- Kabasakal, K. A., & Kelecioğlu, H. (2015). Effect of differential item functioning on test equating. *Educational Sciences: Theory and Practice*, 15(5), 1229–1246.
- Kan, A., Sünbül, Ö., & Ömür, S. (2013). 6.- 8. Sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi [Investigating the differential item functions of the 6th-8th grade subtests of the Level Assessment Examination according to various methods]. *Mersin University Journal of the Faculty of Education*, 9(2), 207–222.
- Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Lawrence Erlbaum Associates.
- Karami H., & Nodoushan M. A. S. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies*, 5(4), 133–142.
- Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement*, 77(4), 545–569.
- Lee, K. (2003). *Parametric and nonparametric IRT models for assessing differential item functioning* [Unpublished doctoral dissertation]. Wayne State University.
- Li, H., Qin, Q., & Lei, PW. (2017). An examination of the instructional sensitivity of the TIMSS math items: a hierarchical differential item functioning approach, *Educational Assessment*, 22(1), 1–17.
- Mertler, C. A., & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Pyrczak.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.
- Penfield, R. D., & Lam, T. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, 19(3), 5–15.
- 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.
- 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.
- Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, 20(2), 148–166.
- Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, 19(2-3), 170–187.
- Sünbül, Ö., & Sünbül, S. Ö. (2016). Type I error rates and power study of several differential item functioning determination methods. *Elementary Education Online*, 15(3), 882–897.
- Şahin, M. G. (2017). Comparison of objective and subjective methods on determination of differential item functioning. *Universal Journal of Educational Research* 5(9), 1435–1446.
- Turhan, A. (2006). *Multilevel 2PL item response model vertical equating with the presence of differential item functioning* [Unpublished doctoral dissertation]. The Florida State University.
- Vaughn, B. K., & Wang, Q. (2010). DIF trees: Using classifications trees to detect differential item functioning. *Educational and Psychological Measurement*, 70(6) 941–952.
- Walker, C. M., & Gocer Sahin, S. (2016). Using a multidimensional IRT framework to better understand differential item functioning (DIF): A tale of three dif detection procedures. *Educational and Psychological Measurement*, 77(6), 945–970.

- Zheng, Y., Gierl, M. J., & Cui, Y. (2007). Using real data to compare DIF detection and effect size measures among Mantel-Haenszel, SIBTEST and logistic regression procedures [Paper presentation]. *NCME*, Chicago.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Lawrence Erlbaum Associates.
- 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.
- Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: bringing the context into picture by investigating sociological community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, 5(1), 23.
- Zumbo, B. D. A., & Thomas, D. R. (1996). A measure of dif effect size using logistic regression procedures [Paper presentation]. *National Board of Medical Examiners*. US, Philadelphia.
- Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, 2012(1).

BIOGRAPHICAL NOTES

Contribution Rate of Researchers

Author 1: 33%

Author 2: 33%

Author 3: 33%

The contribution rate of researchers to study is equal.

Conflict Statement

There is no conflict of interest the research.



Genişletilmiş Türkçe Özet

<http://www.tayjournal.com>

<https://dergipark.org.tr/tr/pub/tayjournal>

Değişen Madde Fonksiyonunda Tip I Hata ve Güç Oranının Farklı Yöntemlere Göre Belirlenmesi

Giriş

Eğitimde ve psikolojide ölçme ve değerlendirmenin üzerinde durduğu en önemli konulardan biri ulusal (Kamu Personel Seçme Sınavı [KPSS], Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı [ALES] vb.) ve uluslararası (Test of English as a Foreign Language [TOEFL], Programme for International Student Assessment [PISA], Trends in International Mathematics and Science Study [TIMSS] vb.) düzeyde yapılan geniş ölçekli sınavlardır. Sonuçları üzerinde bireyler ve ülkeler hakkında önemli kararlar alınan bu sınavların değerlendirilmesi ve yorumlanması büyük önem taşımaktadır. Dolayısıyla bu sınavların, geçerli yorumlar yapılmasına olanak sağlaması gerekmektedir (Clauser & Mazor, 1998). Diğer bir ifadeyle test sonuçları üzerine verilen kararların doğruları; testten alınan puanların ise bireylerin gerçek performanslarını yansıtması için, yapılan ölçmelerin geçerli olmasına ihtiyaç duyulmaktadır. Test puanları üzerine yapılan yorumların ya da test puanları sonucu alınan kararların doğruluğunu göstermeye yardımcı olan teori ve kanıtların bir derecesi (American Educational Research Association [AERA] vd., 1999) olan geçerlik, testlerin ve diğer ölçme araçlarının sahip olması gereken en önemli özelliklerdendir. Testler ölçtüğü yapıyı, ölçülen özellik dışındaki değişkenlerden etkilenmeden bütün bireyler için aynı doğrulukta ölçmelidir (Sireci & Rios, 2013). Geçerlik birçok faktörden etkilenmekle birlikte, bunların arasında testler için en önemli tehdit unsurları madde ve test yanlılığıdır (Clauser & Mazor, 1998). Madde yanlılığı, 1910 yılında Alfred Binet'in düşük sosyoekonomik düzeye sahip çocuklara zekâ testi uyguladığı çalışmayla ortaya çıkmıştır. Binet, test maddelerini incelediğinde bazı maddelerin zekâ haricinde kültürel özellikleri de ölçtüğünü ortaya koymuş ve testten çıkarmayı uygun görmüştür. 1912 yılında ise Stern, çalışmasında farklı alt gruplarda farklı sonuçların ortaya çıktığını göstermiştir. Daha sonrasında testlerin tek bir gruba yönelik hazırlanması yönünde düşünceler gelişmiştir (Camilli & Shepard, 1994). Cleary ise çalışmasında yordanan ölçüt puanların alt

gruplarda çok yüksek veya çok düşük olduğunu bularak, test yanlılığı kavramını ortaya atmıştır (Lee, 2003).

Yanlılık, bireylerin ölçmek istediğimiz özelliklerine (cinsiyet, okul türü, etnik köken vb.) başka değişkenlerin karışmasıdır ve testten elde edilen sonuçları ve bu sonuçlara dayalı olarak yapılan yorumları bozan sistematik hatalara yol açmaktadır (Gierl vd., 1999). Bir testte bir gruba avantaj sağlayan yanlı maddelerin bulunması testin geçerliği için önemli bir tehdittir (Kane, 2006; Messick, 1989). Dolayısıyla testin hiçbir alt gruba avantaj sağlamayacak şekilde hazırlanması çok önemlidir (Gök vd., 2014). Bir testteki maddelerin yanlı olup olmadığının incelenmesinde ilk basamak ilgili maddelerde DMF olup olmadığını tespit etmektir. DMF, yetenek düzeyi aynı olan bireylerin bir maddeyi doğru yanıtlama ihtimalinin, buldukları alt gruplara göre farklılaşmasıdır (Embretson & Reise, 2000; Hambleton vd., 1991). Zumbo'a (1999) göre DMF, madde ile ölçülmesi hedeflenen yetenek düzeyi için yapılan bir karşılaştırma çalışmasında farklı gruplarda bulunan bireylerin ilgili maddeyi doğru yanıtlama ihtimallerindeki farklılıkları açıklamaktadır. DMF analizleri bir testin içerdiği yanlı maddelerin tespitinde ön koşul olmakla birlikte aynı zamanda testin geçerliği için bir kanıt konumundadır (Embretson, 2007). Yanlı bir maddenin kesinlikle DMF içerdiği söylenebilirken; bir maddede DMF olması o maddenin yanlı olduğunu söylemekte yeterli değildir. DMF içerdiği tespit edilen bir madde için ancak uzman görüşüyle yanlı olduğu sonucuna ulaşılabilir (Zumbo ve Gelin, 2005), dolayısıyla madde yanlılığı tespiti temelinde nitel bir değerlendirme gerektirir (Ellis & Raju, 2003; Furlow vd., 2009; Sireci & Allalouf, 2003).

Literatürde birçok DMF belirleme yönteminden bahsedilmektedir. Ancak Karami ve Nodoushan (2011) aynı test için farklı yöntemlerin farklı maddelerde DMF belirlediğini, dolayısıyla yalnızca tek yöntemle göre analiz yapılarak sonuçların tek yöntemle göre yorumlanmasının doğru olmadığını belirtmişlerdir. Buna göre eğer bir madde birden çok yöntemle göre DMF içeriyorsa, maddenin DMF'li olduğu farklı yöntemlerle desteklenmiş olur. DMF yöntemleri ile ilgili yapılmış çalışmalarda, farklı koşullar altında Tip I hata ve güç oranlarının değişimi üzerine çalışıldığı görülmüştür. Bununla birlikte çalışmalarda benzer yöntemlerin kullanılması dikkat çekmektedir. Bu çalışmada Klasik Test Kuramı ve Madde Tepki Kuramı temelli yöntemlerin Tip I hata ve güç oranları üzerine çalışılmıştır. Bu bağlamda DMF'nin belirlenmesinde koşullar ve koşulların düzeyleri değiştirilmiştir. Dolayısıyla kullanılan yöntemler, ele alınan koşullar ve koşulların düzeyleri açısından diğer çalışmalardan ayrılmaktadır. Çalışmada aşağıdaki araştırma problemlerine cevap aranmıştır.

1. Farklı koşullar altında LR, MH, Lord'un χ^2 , Breslow-Day ve Raju'nun alan indeks yöntemlerinin Tip I hata oranları nasıldır?
2. Farklı koşullar altında LR, MH, Lord'un χ^2 , Breslow-Day ve Raju'nun alan indeks yöntemlerinin güç oranları nasıldır?

Yöntem

Bu çalışmada DMF belirlenmesinde kullanılan KTK ve MTK yöntemlerinin Tip I hata ve güç oranları farklı koşullar altında karşılaştırmalı olarak incelenmiştir. Yöntemlerin performansları hakkında bilgi vererek, literatürde mevcut bilgilere katkı sağlayacak bir araştırma olması nedeniyle, araştırmanın modeli temel araştırmadır.

Mevcut çalışmada, DMF tespitinde kullanılan farklı yöntemlerin belirli koşullardaki Tip I hata ve güç oranlarının belirlenmesi için simülatif veri kullanılmıştır. Tip I hata ve güç oranların

belirlenmesinde aynı veri seti kullanılmıştır. Veri üretimi için WinGen 3 yazılımından yararlanılmıştır. DMF belirleme yöntemlerinden farklı koşullarda elde edilen Tip I hata oranlarının ve gücün hesaplanması amacıyla, referans ve odak gruplar için örneklem büyüklükleri 1000(O=500, R=500), 1500(O=500,R=1000), 1500(O=1000, R=500), 2000(O=1000, R=1000) şeklinde tek biçimli DMF gösteren veriler oluşturulmuştur. Her analiz için madde sayısı 25 olarak belirlenmiştir. Veri üretiminde iki parametrelili lojistik model kullanılmıştır. Madde parametrelerinden a parametresi, ortalaması 0,8, standart sapması 0,02 olan normal dağılımla elde edilmiştir. b parametresi ise minimum değeri -3, maksimum değeri +3 olan tek biçimli dağılımdan random olarak çekilerek belirlenmiştir. Bireylerin yetenek dağılımına ait değerler de ortalaması 0, standart sapması 1 olan normal dağılımdan elde edilmiştir. Bu şekilde bireylerin yetenek dağılımını farklılaştırmadan referans ve odak gruplar için güçlük düzeylerinde farklılık oluşturarak DMF'li madde elde edilmiştir. Referans ve odak gruplar için madde parametreleri ortaktır. DMF içeren madde oranı (%12, %20) ve DMF düzeyi (b= 0,75) şeklinde veriler üretilmiştir. DMF içermesi istenen madde sayısı kadar b parametresine DMF miktarı olan 0,75 eklenmiştir.

DMF'nin tespitinde, KTK ve MTK temelli yöntemler karşılaştırmalı olarak kullanılmıştır. Analizler için "difR" paketi kullanılmıştır. KTK ve MTK temelli LR, Mantel-Haenszel, Lord'un χ^2 , Breslow-Day (BD) ve Raju'nun alan indeks yöntemleri kullanılmıştır. DMF tespitine yönelik analizler için R.3.0.1 programı ve "difR" paketi kullanılmıştır. "difR", değişen madde fonksiyonu belirleme yöntemlerine yönelik indeksleri barındıran bir R paketidir (Magis vd, 2015). Tip I hata analizlerinde, her koşul için ayrı ayrı gerçekleştirilen 20 tekrar sonucunda, DMF içermediği halde DMF'li olarak işaretlenen maddelerin oranı belirlenmiştir. Güç analizlerinde ise, DMF içerirken, DMF'li olarak işaretlenen maddelerin oranı belirlenmiştir.

Bulgular

Yöntemlerin örneklem ve DMF'li madde oranlarına göre güç oranları ortalamaları incelendiğinde LR için sonuçlarda minimum değer 0,141 maksimum değer 0,27; Lord'un χ^2 için minimum değer 0,072 maksimum değer 0,15; MH için minimum değer 0,14 maksimum değer 0,22; BD için minimum değer 0,015 maksimum değer 0,05, Raju'nun alan indeksi yöntemlerinden elde edilen minimum değer 0,041 maksimum değer 0,09 olarak hesaplanmıştır.

En düşük Tip I hata oranı tekrar oranlarının ortalaması alınarak belirlenmiştir. Sonuçlara bakıldığında genel olarak Tip I hata oranının en fazla LR ve Lord'un χ^2 yöntemlerinde en az BD yönteminde olduğu görülmektedir. Bu bulgular, odak ve referans grupların yetenek dağılımları aynı olduğunda MH ve LR Tip I hata oranlarının benzer olduğu göstermektedir. LR, Lord'un χ^2 ve MH yöntemlerinde Tip I hata oranlarının örneklem büyüklüğüne göre dalgalanma gösterdiği ve örneklem büyüklüğü arttıkça arttığı, BD yönteminde büyük örneklem ve DMF oranının az olduğu (%12) durumda en yüksek olduğu, Raju'nun alan indeksi yönteminde ise referans grubunun odak grubun sayısından fazla DMF oranının yüksek olduğu (%20) durumda olduğu görülmüştür.

Diğer yandan yöntemlerin örneklem ve DMF'li madde oranlarına göre güç oranları ortalamaları incelendiğinde LR için minimum değer 0,90 maksimum değer 1, Lord'un χ^2 için minimum değer 0,94 maksimum değer 1, MH için minimum değer 0,74 maksimum değer 1, BD için minimum değer 0,03 maksimum değer 0,70, Raju'nun alan indeksi yöntemlerinden elde edilen minimum değer 0,88 maksimum değer 1 olarak hesaplanmıştır. En düşük güç oranı tekrar oranlarının ortalaması alınarak

belirlenmiştir. Sonuçlara bakıldığında genel olarak güç oranının en fazla %12 DMF'li madde oranlarında olduğu durumda en az BD yönteminde olduğu görülmüştür.

Koşullardan elde edilen sonuçlara bakıldığında güç oranının genel olarak Lord'un χ^2 yönteminde en fazla, BD yönteminde ise en az olduğu görülmektedir. 2000 kişilik örnekleme %12 oranında DMF'li madde bulunduğu koşulda güç oranının genel olarak bütün yöntemler için en fazla olduğu durum olduğu ve en düşük oranın BD yönteminde 0,2 olduğu görülmektedir. Ayrıca BD yönteminde güç oranının en fazla büyük örneklem ve DMF'li madde oranının %20 olduğu durumda olduğu görülmüştür.

Tartışma ve Sonuç

Mevcut çalışma kapsamında DMF içeren madde oranlarının farklı olduğu durumlar ve örneklem sayıları, Tip I hata ve güç oranları ortalamalarının değişimleri için gözlenmiştir. Gözlemler için KTK ve MTK yöntemlerinden yararlanılmıştır. LR, Lord'un χ^2 ve MH yöntemlerinde Tip I hata oranlarının örneklem büyüklüğüne göre değiştiği ve örneklem büyüklüğü arttıkça arttığı görülmüştür. BD yönteminde örneklemin büyük ve DMF oranının az olduğu (%12) durumda en yüksek olduğu, Raju'nun alan indeksi yönteminde ise en yüksek hata oranı referans grubun örnekleminin odak grubun örnekleminde büyük, DMF oranının yüksek olduğu (%20) durumda görülmüştür. Bu sonuçlar literatürdeki benzer çalışmalarla da paralellik göstermektedir (Ankenmann vd., 1996; Atar & Kamata, 2011; Gierl vd., 2000; Rogers & Swaminathan, 1993; Roussos & Stout, 1996; Vaughn & Wang 2010).

Örneklem büyüklüğünün en yüksek, DMF'li madde oranının düşük olduğu koşulda; LR, Lord'un χ^2 , MH ve Raju'nun alan indeksi yöntemlerinde güç oranının en fazla olduğu; en düşük oranın ise BD yönteminde 0,2 olduğu görülmektedir. Sünbül ve Sünbül (2016), simülatif veri üzerinden yaptıkları çalışmada örneklem büyüklüğünün artmasıyla yöntemlerin güç oranlarının arttığının belirtmişlerdir. Aynı zamanda DMF'li madde oranının artmasıyla yöntemlerin güç oranlarının azalması bu çalışmayı destekler niteliktedir.

Sonuçlarda genel anlamda DMF'li madde oranının artması, Tip I hatanın artmasına ve gücün azalmasına sebep olmuştur. Benzer şekilde Erdem-Keklik (2014), MH ve LR yöntemlerinin Tip I hata ve güç oranlarını karşılaştırdığı çalışmada geniş örneklem büyüklüklerinde Tip I hatanın yüksek olduğunu belirtmiştir. Analiz sonuçlarında düşük hata ve yüksek güç ile MTK yöntemlerinden olan Lord'un χ^2 ve Raju'nun alan indeksi yöntemlerinin diğerlerine nazaran daha iyi sonuçlar verdiği görülmüştür.

Öneriler

Bu çalışma kapsamında tek biçimli olan DMF incelenmiştir. Benzer çalışmalara ek olarak tek biçimli olmayan DMF, yetenek parametrelerinin değiştirilmesiyle incelenebilir. Çalışmada madde sayısı 25 olarak belirlenmiş madde sayısının daha az veya daha fazla olduğu durumlarda hata ve güç oranlarının değişimi farklı DMF belirleme yöntemleriyle de incelenebilir. Ayrıca örneklem büyüklükleri, örneklem büyüklüğü oranı, madde ayırt ediciliği gibi değişkenler manipüle edilerek DMF belirleme yöntemlerinin performansları incelenebilir.