ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(1); 1-12

Performances of MIMIC and Logistic Regression Procedures in Detecting DIF *

Seçil UĞURLU ** Burcu ATAR ***

Abstract
In this study, differential item functioning (DIF) detection performances of multiple indicators, multiple causes (MIMIC) and logistic regression (LR) methods for dichotomous data were investigated. Performances of these two methods were compared by calculating the Type I error rates and power for each simulation condition. Conditions covered in the study were: sample size (2000 and 4000 respondents), ability distribution of focal group [N(0, 1) and N(-0.5, 1)], and the percentage of items with DIF (10% and 20%). Ability distributions of the respondents in the reference group [N(0, 1)], ratio of focal group to reference group (1:1), test length (30 items), and variation in difficulty parameters between groups for the items that contain DIF (0.6) were the conditions that were held constant. When the two methods were compared according to their Type I error rates, it was concluded that the change in sample size was more effective for MIMIC method. On the other hand, the change in the percentage of items with DIF was more effective for LR. When the two methods were compared according to their power, the most effective variable for both methods was the sample size.

Key Words: Differential item functioning, MIMIC model, Logistic regression, Uniform DIF, Type I error rate and power.

INTRODUCTION

Test items may be biased since they may contain constructs that are undesired to be measured along with the desired ones. Any item may also be in relation with a second or more factors other than the one which is of interest. Those factors that are irrelevant to the construct being measured may affect the performances of individuals. This issue is known as test bias. While test bias focuses on test scores and is interested in fairness of a test, item bias focuses on the relationship between answering an item correctly and group membership. And hence, item bias is related to a specific item. Differential item functioning (DIF), which is a statistical method used in item bias analysis, has been the subject of a vast majority of recent studies (Zumbo, 1999).

DIF occurs when respondents who are at the same ability level but from different groups have different item response probabilities on a specific item (Crane, Belle & Larson, 2004; Mazor, Kanjee & Clauser, 1995). In other words, the expression of DIF is that an item displays different statistical properties in different groups for individuals who are at the same ability levels (Holland & Wainer, 1993). Many methods have been developed for detecting test items with DIF. Some DIF detection methods used for dichotomously scored items are; chi-square test based on item response theory (Lord, 1980), standardization (Dorans & Kulick, 1986), Mantel-Haenszel (MH) (Holland & Thayer, 1988), item response theory likelihood ratio test (IRT-LRT) (Thissen, Steinberg & Wainer, 1988), logistic regression (LR) (Swaminathan & Rogers, 1990), simultaneous item bias test (SIBTEST) (Shealy & Stout, 1993), and multiple indicators, multiple causes (MIMIC) model (Finch, 2005; Oort, 1998).

Fleishman, Spector, and Altman (2002) mentioned in their study that when there are more than two groups, methods get very complicated for testing DIF in IRT framework. As they mentioned in their study, the MIMIC model has an advantage of including multiple exogenous variables to the analysis

* This study is based on Seçil Uğurlu's master thesis titled "Performance of Multiple Indicators Multiple Causes and Logistic Regression Procedures in Detecting Differential Item Functioning".

** Res. Assist., Hacettepe University, Faculty of Education, Ankara-Turkey, secilarslan@hacettepe.edu.tr, ORCID ID: 0000-0002-3495-7797
*** Assoc. Prof. Ph.D., Hacettepe University, Faculty of Education, Ankara-Turkey, burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Uğurlu, S., & Atar, B. (2020). Performances of MIMIC and logistic regression procedures in detecting DIF. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 1-12. doi: 10.21031/epod.531509
Received: 23.02.2019
Accepted: 22.11.2019

simultaneously. Because of allowing a simultaneous analysis of several groups in a single framework, MIMIC model seems to be very useful (Muthen, 1988). This method has become an interesting research subject when its advantages on DIF researches are considered. MIMIC method is quite new with respect to the other methods mentioned above, and especially regarding dichotomous data, there are few studies in the literature involving MIMIC method (see Finch, 2005). Some recent studies on this method were conducted by Fleishman et al. (2002), Woods (2009), Wang, Shih, and Yang, (2009), Woods, Oltmanns and Turkheimer (2009), and Wang and Shih, (2010). Considering these studies, it is reasonable to investigate that under which circumstances MIMIC method is more effective in DIF detection. The aim of the current study is to compare the performance of MIMIC method with LR method - a commonly used method - in detecting items with DIF and interpret the results of these two methods. The DIF detection methods used in this study was explained in detail in the following sections:

Logistic Regression DIF Detection Method

As specified by Swaminathan and Rogers (1990), in detection of differential item functioning, LR model for the two groups of interest can be expressed as:

$$P(u_{ij}=1|\theta_{ij})= \frac{e^{(\beta_{0j}+\beta_{1j}\theta_{1j})}}{1+e^{(\beta_{0j}+\beta_{1j}\theta_{1j})}}, \; i=1, \ldots, n_j, j = 1, 2. \quad (1)$$

u_{ij}: response of ith individual in jth group to the item,

β_{0j}: intercept parameter for jth group,

β_{1j}: slope parameter for jth group,

θ_{ij}: ability of ith individual in jth group.

In Equation 1, if logistic regression curves are the same for the two groups, i.e., $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$, no DIF is present. However, if $\beta_{11} = \beta_{12}$ and $\beta_{01} \neq \beta_{02}$, since the LR curves are parallel, it can be concluded that uniform DIF exists. If $\beta_{01} = \beta_{02}$ and $\beta_{11} \neq \beta_{12}$, since the curves are not parallel, it can be concluded that nonuniform DIF exists (Swaminathan & Rogers, 1990).

MIMIC DIF Detection Method

MIMIC method, which is newer than LR, is based on confirmatory factor analysis (CFA) (Finch, 2005). As outlined by Finch (2005), in DIF context, MIMIC model is as Equation 2:

$$y_i^* = \lambda_i \eta + \beta_i z_k + \varepsilon_i \quad (2)$$

where y_i^* is the latent response variable for ith item (when $y_i^* > \tau_i$, y_i is equal to 1, otherwise y_i is equal to 0; τ_i is the threshold parameter and is related to item difficulty for ith item), η is latent trait variable that is aimed to be measured by the test, λ_i is the factor loading, ε_i is random error, z_k is grouping variable that indicates the group membership and β_i is the slope that relates z_k with y_i^* (Finch, 2005; Wang et al., 2009).

MIMIC is a method that allows conducting DIF analyses with multiple grouping variables, and the z symbol in Figure 1 is defined as a vector of the aforementioned multiple grouping variables. The z vector may have continuous or categorical values. Thus, it can be said that MIMIC method is more flexible than traditional DIF detection methods (MH, SIBTEST, IRT-LRT, etc.) that use just only one categorical grouping variable (Wang et al., 2009).

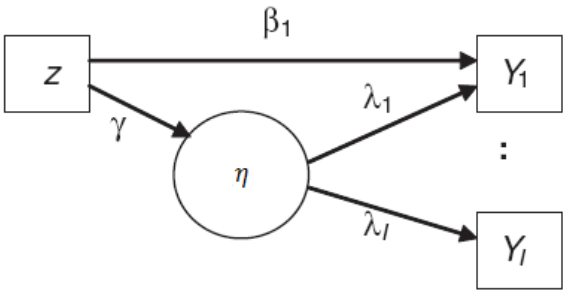

Figure 1. Detecting Differential Item Functioning in Item Y1 with the MIMIC Method. Adapted from "The MIMIC Method with Scale Purification for Detecting Differential Item Functioning" by W. C. Wang, C. L. Shih and C. C. Yang, 2009, *Educational and Psychological Measurement, 69*(5), p. 717. Copyright 2009 by SAGE Publications.

The underlying base method for DIF detection by MIMIC method involves evaluation of both direct and indirect effects for a grouping variable. By investigating the indirect effect of the grouping variable (z) on item responses through the latent trait (η), it is indicated whether the mean of this latent variable differs across the groups or not; thus, computations are carried out for group differences on the latent trait. By investigating the direct effect of the grouping variable (z) on item responses (Yi), i.e. $\beta_1 \neq 0$, it is indicated whether any difference in response probabilities exists across the groups or not. This relation, after checking the differences in the mean of latent trait for groups, is the test of uniform DIF (Finch, 2005).

DIF detection models to be used in bias studies must be appropriate for the test used and for the properties of the groups to which the test is applied. This study used different conditions for dichotomous data to investigate the circumstances under which the MIMIC method produces more accurate results in DIF detection. The conditions used in the current study differ from previous studies in terms of the levels of these three conditions: sample size, ability distribution across groups, and percentage of items with DIF. It is an important question whether the MIMIC method works similarly in cases with different sample sizes (Wang & Shih, 2010). Therefore, different sample sizes in the study were compared. The data used in the study were produced according to the three-parameter logistic model (3PLM), and the test length was taken as 30 items to show similarity with actual applications. In addition, the focus of this study was on the assessment of uniform DIF.

In this study, the MIMIC method was compared to the LR method, which is a relatively more traditional method. This study compared how Type I error rates and power of MIMIC and LR DIF detection methods changed according to sample size, ability distributions of the groups, and percentage of items with DIF. In summary, the goal of this study was to investigate the performances of MIMIC and LR methods under various conditions according to their type I error rates and power when detecting DIF items on dichotomous tests. The research questions were as the following:

1. How do Type I error rates and power of MIMIC and LR DIF detection methods differ according to sample size?
2. How do Type I error rates and power of MIMIC and LR DIF detection methods differ according to ability distributions of the groups?
3. How do Type I error rates and power of MIMIC and LR DIF detection methods differ according to percentage of items with DIF?

METHOD

Simulation Conditions and Data Generation

This study is a DIF detection research using MIMIC and logistic regression methods for dichotomous data based on various simulation conditions. In this simulation study, conditions different from those of previous studies in which the MIMIC model was used were investigated.

The conditions that were kept constant throughout the study

For all conditions, the ability parameters of the individuals in the reference group were generated based on the standard normal distribution, N(0, 1). Furthermore, 30 dichotomously scored (either 0 or 1) responses for each individual were produced. The change in the item difficulty parameters between the groups for the items with DIF was set to a constant value as 0.6 units against the focal group to form medium DIF. The ratio of the focal group to the reference group (1:1) is another condition that was kept constant.

The conditions that were varied throughout the study

One of the conditions that was varied in this study was the sample size. Two levels of large sample size were used: 2000 (R: 1000, F: 1000) and 4000 (R: 2000, F: 2000). Finch (2005) found in his study that MIMIC method produces type I error rates higher than .05 nominal alpha level for a shorter test (i.e., 20 items) responded by a sample of 1000 (R: 500, F: 500) individuals under 3PL model. Based on the findings of Finch (2005), for a test with 30 items under 3PL model considered in this study, larger sample sizes were taken into account. In addition to sample size, ability distribution of the focal group was also a condition that was varied. Two levels of ability distribution of focal group were used: N(0, 1) and N(-0.5, 1). For the first level of the ability distribution of focal group condition, the cases where the distribution of the reference group and the focal group is the same were considered. For the second level of the ability distribution of focal group condition, the cases where the distribution of the focal group is lower than the reference group were considered Another condition that was varied in this study was the percentages of items with DIF. Two levels were used for this condition: 10% (3 items) and 20% (6 items). Items with DIF were kept the same throughout the test. In 10% of items with DIF condition, DIF was formed for items 4, 15, and 27 and in 20% of items with DIF condition, it was formed for items 1, 4, 15, 18, 26, and 27. By crossing the levels of each condition, total of 8 simulation conditions were created.

For each simulation condition, the data were derived for dichotomously scored (0/1) items using a 3PLM via R 3.0.2 program (R Core Team, 2013). The derivation of the data was performed 100 times for each condition. The item parameters used in this study were selected randomly from the item parameters used in Finch's (2005) study. The selected parameters are shown in Table 1.

Data Analysis Procedures and Evaluation Criteria

In the DIF analyses of the data, Mplus 6.12 (Muthén & Muthén, 1998, 2010) program was used for the MIMIC method and SAS 9.1.3 (SAS Institute, 2007) program was used for the logistic regression method. The DIF analyses were conducted using a pairwise approach in which the groups are compared with each other (i.e., focal group compared with reference group) (Sari & Huggins, 2014).

In the study, the effects of sample size, ability distribution of focal group, and the percentage of items with DIF on Type I error rates and power were investigated. The level of significance (α level) was assumed to be .05 in detecting items with DIF. Type I error is defined as a misclassification of an item without DIF as an item with DIF. Under 10% of items with DIF condition, there were 27 non-DIF items whereas under 20% of items with DIF condition, there were 24 non-DIF items. The percentage of non-DIF items that were falsely detected as DIF items was calculated for Type I error rate. The concept of power, on the other hand, is correct classification of an item with DIF as an item with DIF. Under 10% of items with DIF condition, there were 3 DIF items whereas under 20% of items with DIF condition, there were 6 DIF items. The percentage of DIF items that were correctly detected as DIF items was calculated for power. Both Type I error and power are equally important for DIF researches (Vaughn & Wang, 2010). According to Cohen and Cohen (1983) when investigators need to set the power, it is reasonable for them to choose a value in the .70 - .90 range. In the current study, the desired value for power rate was considered as .70 and above.

Table 1. Item Parameter Values Used in Generation of Simulated Data

	Reference Group		
Item	a_i	b_i	c_i
1	1.10	-0.70	.20
2	0.70	-0.60	.20
3	1.40	0.10	.20
4	0.40	0.80	.20
5	1.40	-0.40	.20
6	1.60	-0.10	.16
7	1.20	0.50	.20
8	1.20	1.40	.11
9	1.80	1.40	.12
10	2.00	1.60	.16
11	1.00	1.60	.13
12	1.50	1.70	.09
13	0.70	-0.50	.20
14	1.20	-0.30	.20
15	0.90	0.20	.20
16	0.70	-0.40	.20
17	1.00	0.70	.15
18	1.60	1.10	.12
19	1.10	2.00	.06
20	1.10	2.40	.09
21	1.70	1.30	.17
22	0.90	1.00	.15
23	0.50	-0.60	.20
24	1.30	0.40	.18
25	1.30	1.40	.06
26	1.10	1.20	.05
27	0.90	0.80	.20
28	0.40	-0.40	.20
29	0.80	-0.70	.20
30	1.00	1.10	.13

RESULTS

Type I Error Rate

Type I error rates are calculated for each condition, namely sample size, ability distribution of focal group, and percentage of items with DIF and given in Table 2.

Table 2. Type I Error Rates According to Sample Size, Ability Distribution of Focal Group, and Percentage of Items with DIF

DIF %	Sample Size	Ability Distributions R/F	MIMIC	LR
10	2000	(0,1) / (0,1)	.121	.069
		(0,1) / (-0.5,1)	.120	.068
	4000	(0,1) / (0,1)	.065	.087
		(0,1) / (-0.5,1)	.090	.097
20	2000	(0,1) / (0,1)	.129	.122
		(0,1) / (-0.5,1)	.128	.129
	4000	(0,1) / (0,1)	.076	.244
		(0,1) / (-0.5,1)	.078	.189

Note. DIF % refers to the percentage of items with DIF; LR = Logistic Regression; MIMIC = Multiple Indicators, Multiple Causes Model.

The main finding of this study was that the sample size was an important factor in DIF analyses conducted with MIMIC and LR methods. As the sample size increased from 2000 to 4000, the type I error rates decreased for MIMIC method but increased for the LR method when other conditions of the study were equal. For the MIMIC method, while the lowest rate was calculated under the condition

where the sample size was 4000, percentage of items with DIF was 10%, and the ability distribution of both groups showed a standard normal distribution N(0, 1), the highest rate was calculated under the condition where the sample size was 2000, percentage of items with DIF was 20%, and the ability distribution of both groups showed a standard normal distribution N(0, 1). On the other hand for the LR method, while the lowest rate was calculated under the condition where the sample size was 2000, percentage of items with DIF was 10%, and ability distribution of the focal group was N(-0.5, 1), the highest rate was calculated under the condition where the sample size was 4000, percentage of items with DIF was 20%, and the ability distribution of both groups showed a standard normal distribution N(0, 1).

The second important finding was that the percentage of DIF items was an important factor that effected the type I error rates. As the percentage of DIF items increased from 10% to 20%, type I error rates were very similar in MIMIC method, however, increased in LR method when other conditions of the study were equal. According to the study results, in terms of type I error rates, the percentage of DIF items was more effective factor for the LR method.

The third finding was that the change in the ability distribution of focal group did not have an important effect on type I error rates for both methods.

Power

Table 3 presents the power values for the two DIF detection methods for all conditions included in the study. The acceptable power rate for this study was .70 and above. In general, both methods had power rates above acceptable levels for all conditions.

The power rate of the MIMIC method was quite high for conditions with a sample size of 4000 respondents. The power rate of the LR method, on the other hand, was quite high for conditions wherein the sample size was large and the ability distribution of both groups showed a standard normal distribution N(0, 1). The standard definition of power at a specified level of alpha is not meaningful in cases where Type I error rates are high (Finch, 2005). However, all power results were included in this study for comparison purposes. The power rates were shown in italics for cases where Type I error rate was higher than .10. Considering all conditions, both methods had power high enough and these results reached a higher value when sample size increased.

Table 3. Power Rates According to Sample Size, Ability Distributions, and Percentage of Items with DIF

DIF %	Sample Sizes	Ability Distributions R/F	MIMIC	LR
10	2000	(0,1) (0,1)	*.770*	.800
		(0,1) (-0.5,1)	*.750*	.700
	4000	(0,1) (0,1)	.933	.910
		(0,1) (-0.5,1)	.910	.817
20	2000	(0,1) (0,1)	*.852*	*.827*
		(0,1) (-0.5,1)	*.780*	*.772*
	4000	(0,1) (0,1)	.977	*.935*
		(0,1) (-0.5,1)	.943	*.872*

Note. DIF % refers to the percentage of items with DIF; LR = Logistic Regression; MIMIC = Multiple Indicators, Multiple Causes Model.

The condition in which the power was closest to perfect for the MIMIC method was the one in which the sample size was 4000 respondents, ability distributions of the reference and focal groups showed a standard normal distribution, and percentage of items with DIF was 20%. The power results of the MIMIC method were larger than those of the LR method, except for a single condition. This condition was the one in which the sample comprised 2000 respondents, ability distributions of the reference and focal groups showed a standard normal distribution, and percentage of items with DIF was 10%. The differentiation of the ability distributions for the focal group affected the power of the LR method

more than the power of the MIMIC method for almost all conditions. In addition, the change in the percentages of items with DIF did not substantially change the power of both methods.

DISCUSSION and CONCLUSION

In this study, the performances of MIMIC and LR methods were compared according to their type I error rate and power. It can be concluded in this study that the MIMIC method produced lower Type I error rates than the LR method in conditions where the sample size was larger (4000 respondents); the LR method produced lower Type I error rates than the MIMIC method in conditions where the percentage of items with DIF was lower (10%) with smaller sample size (2000 respondents). In general, the Type I error rates of the MIMIC method were observed to be lower than those of the LR method. However, for both methods, Type I error rates exceeded acceptable alpha level ($\alpha = .05$) in all conditions. Specifically, while the increase in the sample size substantially reduced the Type I error rate of the MIMIC method for all conditions, its effect on the type I error rate of the LR method changed according to the percentage of items with DIF. While the change in the sample size had a very small effect on the Type I error rate of the LR method for 10% DIF items conditions, it caused a substantial increase in the Type I error rate of this method for 20% DIF items conditions. In the study conducted by Finch and French (2007), Type I error rates of the LR and CFA methods in detecting items with nonuniform DIF were not substantially affected by the increase in the sample size. Based on this results, it can be concluded that similar results obtained from current study for the LR method with only the 10% DIF items conditions. As can be understood from this current research, in the conditions where the percentage of items with DIF is high the LR method is more sensitive to the sample size condition. But the MIMIC method is affected by the sample size in the same manner for all conditions. The difference based on CFA between current and Finch and French's (2007) study can be attributed to the type of DIF. In their study they focused on nonuniform DIF and emphasized the question of the usefulness of CFA method for identifying this type of DIF. MIMIC method is also based on CFA and it is capable of detecting uniform DIF as also stated by Woods (2009), and Woods et al. (2009).

On the other hand, in the current study the increase in the percentage of items with DIF did not affect the Type I error rate of the MIMIC method importantly but increased that of the LR method. It can be seen in Finch's (2005) results that for the MIMIC method, in the bigger test length condition the effect of percentage of items with DIF was reduced for both sample size conditions, 600 and 1000 respondents. In the current study for both sample size (2000 and 4000 examinees) the effect of percentage of items with DIF was already quite low but still the type one error rates were not small enough as they were desired. By combining the result of these two studies it can be concluded for the MIMIC method that, big sample sizes or relatively small sample sizes with bigger test lengths are needed to reduce the effect of percentage of items with DIF.

The other result obtained from this study is that, the difference in the ability distribution of the focal group did not substantially affect the Type I error rates of both methods. In conclusion, when these two methods were compared in terms of Type I error rates, the change in the sample sizes was more effective for the MIMIC method while the change in the percentages of items with DIF was more effective for the LR method.

When the results were examined in general, the power of both methods for all conditions was above the acceptable level (.70). For conditions where the sample size was higher, the power results of the MIMIC method were quite high. The power of the LR method, on the other hand, was quite high for conditions where the sample size was large and the ability distribution of both groups showed a standard normal distribution. The power results of the MIMIC method were higher than those of the LR method, except for a single condition. This condition was the one in which the sample comprised 2000 respondents, the ability distributions of the reference and focal groups showed a standard normal distribution, and the percentage of items with DIF was 10%.

The increase in the sample size increased the power for both methods. The fact that the ability distribution of the focal group differed from the ability distribution of the reference group decreased

the power of both methods. The amount of reduction that this change in the ability distribution caused was more for the LR method for almost every condition. The increase in the percentage of items with DIF increased the power of both methods to a small extent. As a result, considering the change in the power, the sample size was the most effective variable for both methods.

Specifically, the change in the sample size was very effective in changing the power of the MIMIC method. The power of the MIMIC method increased as the sample size increased. Finch (2005) concluded in his study that the power results of the MIMIC method for 2PLM were generally as high as the power results of the classical methods or even in some conditions higher than those of the SIBTEST and MH methods. Similar results were obtained in this study for 3PLM, the power results of the MIMIC method were higher than those of the LR method for almost all conditions.

In the study conducted by Finch and French (2007), the power results of the LR and CFA methods in detecting items with nonuniform DIF were below .70 for all conditions. In current study, the power results were over .70 for both methods for all conditions. Finch and French (2007) reported in their study that the power of the LR method increased as the sample size increased. But, according to their results the power of the CFA method decreased or stayed the same while the sample size increased. In current study, as the sample size increased, the power of both LR and MIMIC methods increased. These two studies support each other in terms of the increase in power of the LR method according to the sample size condition. However, the results differed in terms of the change in the power of the MIMIC method, which is a method based on CFA. As mentioned before this difference between two studies can be attributed to the difference of the type of DIF (uniform or nonuniform) used in these studies.

In this study, three main conditions and eight sub-conditions were considered, with two different sample sizes, two different ability distributions for the focal group, and two different percentages of items with DIF. The number of items in the test was kept constant for all conditions. In future studies, the number of items in the test can be increased to see how the results are affected in long tests. As seen in the comparison of recent and previous research, test length may have an important effect on MIMIC method.

It is an important issue how the MIMIC method performs in terms of DIF at different sample sizes. Two different sample sizes, 2000 and 4000 individuals, were used in the study. However, the desired Type I error rates could not be achieved even with a sample size of 4000 individuals. This points out an important issue. And hence, future studies can be conducted on larger sample sizes to investigate the ideal sample size for the MIMIC method.

In the study, the ratio between the reference and focal group sizes was taken as 1:1. However, during the actual examinations, there can be different situations regarding the proportions of sample size of these two groups. Therefore, studies can be done using different ratios. Furthermore, the study was conducted with 3PL model-based data. Similar work can be conducted with 2PL model-based data, and comparisons can be made between these studies.

It is thought that this study will be a reference to the studies on DIF detection through the MIMIC method and that it will make it easy for researchers to decide the appropriate DIF detection method according to sample size and ability distributions in the analysis of the actual test results.

The aim of this study is to provide a reliable source to researchers in selecting DIF detection techniques that are appropriate for the test to be used and the properties of the test group. Thus, with the help of more reliable DIF detection techniques, tests can be made fairer.

Based on the results obtained from this research, it can be suggested to choose the LR method in DIF analysis studies performed on small samples such as the one comprising 2000 respondents and with small amount of DIF items such as 10% of test items; and the MIMIC method in DIF analysis studies performed on samples as large as approximately 4000 respondents and higher. Subsequent to the detection of items with DIF using these methods, it is advisable to refer to expert's opinion to conduct a study to detect bias in these items.

REFERENCES

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Crane, P. K., Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*(2), 241-256. doi: 10.1002/sim.1713

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355-368.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295. doi: 10.1177/0146621605275728

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. doi: 10.1177/0013164406296975

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences, 57B*(5), 275-284.

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

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*(2), 131-144.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Lawrence Erlbaum Associates.

Muthén, L. K. & Muthén, B. O. (1998, 2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 5*(2), 107-124. doi: 10.1080/10705519809540095

R Core Team (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Sari, H. I. & Huggins, A. C. (2014). Differential item functioning detection across two methods of defining group comparisons: Pairwise and composite group comparisons. *Educational and Psychological Measurement, 75*(4), 648-676. doi: 10.1177/0013164414549764

SAS Institute Inc. (2007). *SAS® 9.1.3 qualification tools user's guide*. Cary, NC: SAS Institute Inc.

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

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

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Wang, W. C., & Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement, 34*(3), 166-180. doi: 10.1177/0146621609355279

Wang, W. C., Shih, C. L., & Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*(5), 713-731. doi: 10.1177/0013164409332228

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research,44*(1), 1-27. doi: 10.1080/00273170802620121

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. *J Psychopathol Behav Assess, 31*, 320-330. doi: 10.1007/s10862-008-9118-9

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

MIMIC ve Lojistik Regresyon Yöntemlerinin DMF Belirleme Performansları

Giriş

DMF (Değişen Madde Fonksiyonu), eşit yetenek düzeyinde ancak farklı gruplarda yer alan bireylerin belirli bir maddeye verdikleri cevapların doğru olma olasılığının birbirinden farklı olması durumunda ortaya çıkar (Crane, Belle & Larson, 2004; Mazor, Kanjee & Clauser, 1995). DMF'li maddeleri tespit etmek üzere çok sayıda DMF belirleme yöntemi geliştirilmiştir. Bu çok sayıdaki yöntem arasından MIMIC (Multiple Indicators, Multiple Causes) yöntem oldukça yenidir ve özellikle iki kategorili puanlanan test maddelerinde MIMIC yöntemin kullanıldığı araştırma sayısının eksikliği göze çarpmaktadır (Finch, 2005). Bu nedenle, MIMIC yöntemin DMF belirlemedeki performansının araştırılması gerekli görülmektedir.

Hem sürekli hem de kategorik birden çok sayıda gruplama değişkeni ile kullanılabilen MIMIC yöntemin, sadece tek bir kategorik değişkenle analiz yapmaya izin veren yöntemlere kıyasla daha esnek olduğunu ifade etmek mümkündür (Wang, Shih & Yang, 2009). IRT (Item Response Theory) kapsamında ele alınan DMF testlerinde ikiden fazla grup söz konusu olduğunda yöntemlerin oldukça karmaşıklaştığı görülmekte iken MIMIC yöntemin aynı anda çok sayıda değişkeni analize ekleyebilme avantajı söz konusudur (Fleishman, Spector & Altman, 2002). Birden fazla grubun eşzamanlı olarak tek bir aşamada analizine olanak sağladığı için MIMIC yöntemi oldukça kullanışlı bulunmaktadır (Muthen, 1988). DMF araştırmalarındaki avantajları göz önüne alındığında bu yöntem oldukça ilgi çekici bir araştırma konusu haline gelmektedir.

Yanlılık araştırmalarında kullanılan DMF belirleme yöntemleri kullanılan teste ve testin uygulandığı grubun özelliklerine uygun olmalıdır. Bu amaçla, bu araştırmada MIMIC yöntemin hangi koşullar altında daha doğru sonuçlar verdiği ortaya çıkarılmak istenmiş ve araştırma iki kategorili verilerle çeşitli koşullar kullanılarak yürütülmüştür. Çalışmada etkisi incelenen koşullar örneklem büyüklüğü, DMF'li madde yüzdesi ve gruplar arası yetenek dağılımlarıdır. Ayrıca, bu araştırmada tek biçimli (uniform) DMF'nin belirlenmesi üzerine odaklanılmıştır. Özetle bu araştırmada MIMIC ve LR (Logistic Regression) yöntemleri farklı örneklem büyüklüğü, grupların yetenek dağılımı farklılıkları ve DMF'li madde yüzdesinin değiştiği koşullarda Tip 1 hata ve güçlerine dayalı olarak karşılaştırılmıştır. Buna bağlı olarak araştırmanın problem cümlesine aşağıda yer verilmiştir:

MIMIC ve LR DMF belirleme yöntemlerinin Tip 1 hata ve güçleri örneklem büyüklüğü, grupların yetenek dağılımları ve DMF'li madde yüzdesine göre nasıl değişmektedir?

Yöntem

Bu çalışma iki kategorili puanlanan veriler için yürütülmüş, simülasyona dayalı bir DMF belirleme çalışmasıdır. Çalışmada kullanılan DMF belirleme yöntemleri MIMIC ve LR'dir. Çalışmanın verilerini üretmek üzere R 3.0.2, DMF belirleme analizleri içinse MPlus 6.12 ve SAS 9.3.1 programlarından yararlanılmıştır. Analizler her bir koşula ait veri setleri üzerinde 100 kez tekrarlanmıştır. Ayrıca araştırmanın verileri 3 parametreli lojistik modele (3PLM) uygun olacak şekilde üretilmiştir.

Çalışmada sabit tutulan koşullar şu şekildedir: referans grupta yer alan bireylerin yetenek parametrelerine ait dağılım [N(0,1)], test uzunluğu (30 madde), DMF'li maddeler için gruplara ait güçlük parametreleri farkı (0.6 birim), odak gruptaki bireylerin sayısının referans gruptakilere oranı (1:1). Çalışmanın değişen koşulları ise şu şekildedir: örneklem büyüklüğü (2000, 4000), odak grupta yer alan bireylere ait yetenek dağılımları [N(0,1), N(-0.5, 1)] ve DMF'li madde yüzdesi (%10, %20).

Sonuç ve Tartışma

Özetle bu araştırmada örneklem büyüklüğü, yetenek dağılımı ve DMF'li madde yüzdesinin MIMIC ve LR yöntemlerine ait Tip 1 hata ve güç üzerindeki etkileri incelenmiştir. Genel olarak bakıldığında MIMIC yöntemine ait Tip 1 hatanın LR yöntemininkilere göre daha düşük olduğu göze çarpmıştır. Ancak her iki yöntem için de tüm koşullarda Tip 1 hatalarının kabul edilebilir alfa düzeyinden (α = .05) yüksek çıktığı görülmüştür. Koşullar detaylı olarak incelenecek olursa, örneklem büyüklüğündeki artış tüm koşullar için MIMIC yöntemin Tip 1 hatasını önemli ölçüde düşürmüştür. Ancak LR yöntemin Tip 1 hatasındaki değişim DMF'li madde yüzdesine bağlı olarak değişmiştir. %10 DMF içeren koşullarda Tip 1 hata önemli ölçüde değişiklik göstermezken %20 DMF'li madde koşulunda hata önemli ölçüde artmıştır. Demek oluyor ki LR yöntemi DMF'li madde yüzdesi arttıkça örneklem büyüklüğüne duyarlı hale gelmiştir. Daha önce benzer şekilde LR ve DFA (Doğrulayıcı Faktör Analizi) yöntemleri ile yürütülen Finch ve French'in (2007) çalışma bulguları ise neredeyse her iki yöntem için de bu araştırmanın sonuçlarından farklılık göstermektedir ve bu farklılık MIMIC yöntem için daha belirgin çıkmıştır. Finch ve French'in (2007) bulguları LR ve DFA yöntemlerinin Tip 1 hatalarının örneklem büyüklüğünden önemli derecede etkilenmediklerini işaret etmiştir. MIMIC yöntemi DFA'ya dayalı bir yöntemdir. Bu iki çalışmanın sonuçları arasındaki farklılığın sebebinin bu açıdan düşünüldüğünde DMF türü olabileceği söylenebilir. Çünkü DFA yönteminin tek biçimli olmayan DMF'yi belirlemedeki kullanışlılığından şüphe duyulduğu Finch ve French'in (2007) araştırma sonuçları arasındadır. Ayrıca DFA'ya dayanan MIMIC yönteminin de tek biçimli DMF'yi belirleyebildiği, tek biçimli olmayan DMF'yi belirlemede yetersiz olduğu Woods (2009), Woods, Oltmanns ve Turkheimer'in (2009) araştırmalarında açıkça belirtilmiştir.

Çalışmanın bir başka sonucuna göre, hem 2000 hem de 4000 kişilik örneklem büyüklüklerinde DMF'li madde yüzdesindeki artışın MIMIC yöntemin Tip 1 hatasına etki etmediği ancak LR yöntemininkini arttırdığı görülmüştür. Finch'in (2005) yürüttüğü araştırmada 600 ve 1000 örneklem büyüklüklerinde test uzunluğunun artması ile DMF'li madde yüzdesinin MIMIC yöntem üzerindeki etkisinin azaldığı görülmüştür. Bu iki araştırmanın sonuçları birlikte düşünüldüğünde DMF'li madde yüzdesinin MIMIC yöntem üzerindeki etkisini azaltmak için 2000 ve 4000 gibi daha büyük örneklem büyüklüklerine ya da 600 veya 1000 gibi nispeten daha küçük örneklem büyüklükleri ile birlikte daha büyük test uzunluklarına ihtiyaç duyulmaktadır.

Araştırmanın bir başka sonucu ise odak grubun yetenek dağılımındaki farklılığın her iki yöntemin de Tip 1 hatalarını etkilemediği yönündedir. Özetle, iki yöntem Tip 1 hataları bakımından karşılaştırıldığında MIMIC yöntem için örneklem büyüklüğündeki değişim daha etkili iken, LR yöntem için DMF'li madde yüzdesindeki değişim daha etkili olmuştur.

Araştırma sonuçları yöntemlerin güçleri bakımından incelendiğinde, her iki yöntemin güç değerlerinin tüm koşullar için kabul edilebilir değerin (.70) üzerinde olduğu gözlemlenmiştir. Araştırma sonuçlarına göre her iki yöntem için de güç değerleri açısından, örneklem büyüklüğü en etkili değişken olmuştur. Ayrıca sonuçlar neredeyse tüm koşullarda MIMIC yöntemin güç değerlerinin LR yöntemininkilerden daha yüksek olduğunu işaret etmiştir. Benzer bir sonuca Finch'in (2005) araştırmasında rastlanmıştır. Bu araştırmada da MIMIC yöntemin güç değerlerinin klasik yöntemlerinki kadar yüksek olduğu vurgulanmış ve hatta bazı koşullarda SIBTEST ve MH yöntemlerine göre daha yüksek güç değerlerine sahip olduğu belirtilmiştir.

Bu araştırmada her iki yönteme ait güç değerlerinin tüm koşullar için .70 ve üzeri değerler verdiği tespit edilmiştir. Finch ve French'in (2007) araştırma sonuçlarına göre ise LR ve DFA yöntemlerinin güç değerlerinin neredeyse tüm koşullarda .70 değerinin altında olduğu görülmüştür. Ayrıca, örneklem büyüklüğü arttıkça LR yönteminin güç değerinin arttığı ancak, DFA yönteminin güç değerinin azaldığı ya da aynı kaldığı belirtilmiştir. Bu araştırmanın sonuçlarına göre ise örneklem büyüklüğü arttıkça LR ve MIMIC yöntemlerin güç değerlerinin arttığı gözlenmiştir. Bu bakımdan iki çalışma LR yöntemi sonuçlarına dayalı olarak birbirini destekler nitelikte iken MIMIC ve DFA yöntemleri sonuçları bakımından birbirini desteklememektedir. Daha önce de belirtildiği üzere MIMIC yöntem DFA'ya

dayalı bir yöntemdir ve bu iki araştırma sonucundaki farklılığın sebebinin DMF türüne (tek biçimli ve tek biçimli olmayan) dayandığı söylenebilir.

Bu araştırmada test uzunluğu sabit tutulmuştur. Ancak test uzunluğunun MIMIC yöntem üzerindeki etkisinin daha net ortaya konabilmesi için ileriki araştırmalarda araştırmacılara daha büyük test uzunluklarını kullanarak araştırmalar yürütmeleri önerilebilir. Ayrıca, MIMIC yöntemin farklı örneklem büyüklüklerinde nasıl sonuçlar verdiği önemli bir araştırma sorusudur. Bu araştırmada 2000 ve 4000 olmak üzere iki farklı örneklem büyüklüğü ele alınmıştır. Ancak, 4000 kişilik örneklem büyüklüğünde dahi istenen Tip 1 hata oranına ulaşılamamıştır. Bu nokta önemli bir soruna işaret etmektedir. İleriki araştırmalarda daha yüksek örneklem büyüklükleri kullanılarak MIMIC yöntemin yaklaşık hangi örneklem büyüklüğünde ideal sonuçlar verdiği tartışılmalıdır.

Bu araştırma ile, MIMIC yöntemin kullanılarak DMF'li maddelerin belirlenmeye çalışıldığı araştırmalara bir referans olması amaçlanmıştır. Böylece, kullanılan teste ve testi alan grubun özelliklerine uygun DMF belirleme yöntemlerinin seçiminde araştırmacılara güvenilir bir kaynak sağlanması umulmaktadır. Bununla birlikte, gerçek test sonuçlarının analizinde örneklem büyüklüğü ve yetenek dağılımlarına bağlı olarak uygun DMF belirleme yönteminin seçilmesinde araştırmacılara yardımcı olmak istenmiştir. Daha güvenilir yöntemlerin yardımıyla testler daha adil hale getirilebilir.

Bu araştırmadan elde edilen sonuçlara dayanılarak 2000 gibi küçük örneklem büyüklükleri ve %10 gibi küçük oranda DMF'li madde içeren çalışmalarda LR yönteminin, yaklaşık 4000 ya da daha yüksek örneklem büyüklükleri ile yürütülen çalışmalarda ise MIMIC yöntemin tercih edilmesi önerilebilir. DMF'li maddelerin belirlenmesinin ardından, bu maddelere yönelik yanlılık çalışması yapmak üzere uzman kanısına başvurulması da önerilmektedir.