ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(2);133-148

Performances Based on Ability Estimation of the Methods of Detecting Differential Item Functioning: A Simulation Study*

İbrahim UYSAL** Levent ERTUNA*** F. Güneş ERTAŞ****
Hülya KELECİOĞLU*****

Abstract
The aim of the study is to examine differential item functioning (DIF) detection methods—the simultaneous item bias test (SIBTEST), Item Response Theory likelihood ratio (IRT-LR), Lord chi square ($\chi 2$), and Raju area measures—based on ability estimates when purifying items with DIF from the test, considering conditions of ratio of the items with DIF, effect size of DIF, and type of DIF. This study is a simulation study and 50 replications were conducted for each condition. In order to compare DIF detection methods, error (RMSD) and coefficient of concordance (Pearson's correlation coefficient) were calculated according to estimated and initial abilities for the reference group. As a result of the study, the lowest error and the highest concordance were seen in the case of 10% uniform DIF in the test and the method of IRT-LR, considering all other conditions. Moreover, for the method of SIBTEST and IRT-LR in all conditions, it was found that the error obtained by purifying items with C level DIF is lower than the error obtained by purifying items with both B and C level DIF. Similarly, for the method of SIBTEST and IRT-LR in all conditions, it was seen that the concordance coefficient found by purifying C level DIF is higher than the coefficient by purifying items with both B and C level DIF.

Key Words: Differential item functioning, simulation, ratio of the items with DIF, type of DIF

INTRODUCTION

Tests which are used in education and psychology for various purposes should meet specific standards, such as validity, reliability, and practicality. According to Messick (1995) these characteristics are not only the fundamental principles of measurement, but also the social values used by decision-makers in addition to measurement. In this regard, items in the test should not provide advantages or disadvantages for any subgroup at the same ability level. Otherwise, the test will be biased for specific groups. Bias can be defined as a systematic error in test scores depending on a group of individuals (Camilli & Shepard, 1994). When viewed from this aspect, bias is a major threat for validity and objectivity of a test (Clauser & Mazor, 1998; Kristanjansonn, Aylesworth, McDowell, & Zumbo, 2005).

The process of investigating item bias starts with examining differential item functioning (DIF), which is based on more objective results and may be a measurement of item bias. DIF is defined as differentiation of the probability of correctly responding to an item if individuals are at the same ability level but from different groups (Hambleton, Swaminathan, & Rogers, 1991). It is mentioned in the literature that group differences can be caused by two reasons. One of these is real ability

* This study presented as an oral presentation at the International Meeting of the Psychometric Society (IMPS) in Asheville/USA in 2016.

** Dr., Abant İzzet Baysal University, Faculty of Education, Bolu-Turkey, e-mail: ibrahimuysal@ibu.edu.tr, ORCID ID: 0000-0002-6767-0362
*** Res. Assist., Sakarya University, Faculty of Education, Sakarya-Turkey, e-mail: leventertuna@sakarya.edu.tr, ORCID ID: 0000-0001-7810-1168
**** Res. Assist., Boğaziçi University, Faculty of Education, İstanbul-Turkey, e-mail: gunes.ertas@boun.edu.tr, ORCID ID: 0000-0001-8785-7768
***** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

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

Received: 01.03.2019
Accepted: 26.05.2019

difference between subgroups, which is also called item impact. Item impact refers to the fact that different level subgroups perform differently on items, and this difference does not mean that the item is biased. The other reason is item bias. Different performances can be observed in subgroups due to the item. This means that the item causes one or more of the parameters to be too high or too low, depending on the group (Camilli & Shepard, 1994; Zumbo, 1999).

DIF is classified as uniform and non-uniform functions in terms of its occurrence (Mellenbergh, 1982). The basis of this differentiation is that the ability level and group membership together influence the probability of correct response to an item. Accordingly, uniform DIF occurs when the probabilities of correct response to an item for two groups at the same ability level is constant across all ability levels. On the other hand, non-uniform DIF occurs when the probabilities of correct response to an item for two groups at the same ability level is incoherent at different ability levels (Camilli & Shepard, 1994; Penfield & Lam, 2000; Zumbo, 1999).

Methods of detecting DIF are basically classified according to Classical Test Theory (CTT) and Item Response Theory (IRT). According to CTT, methods of detecting DIF are analysis of variance, chi-square, converted item index, logistic regression, Mantel-Haenszel (MH), and the simultaneous item bias test (SIBTEST). IRT methods are Lord's chi square ($\chi 2$), Raju's area measure, and IRT-likelihood ratio (IRT-LR) (Camilli & Shepard, 1994; Oshima & Morris, 2008). In this study, SIBTEST, IRT-LR, Lord's $\chi 2$, and Raju's area measure are examined; the below provides a brief introduction to these tests.

SIBTEST: DIF in the SIBTEST method is based on the comparison of the response rate of the tested item in the focal group and reference group according to true score. This method tests the null hypothesis that the expected value of differences between specified ratios is equal to zero. In this regard, it can be decided whether or not DIF is present and the level of DIF (Roussos & Stout, 1996). Moreover, on a theoretical basis, this method uses regression-based corrections in order to reduce Type I error (Cheng, 2005).

IRT-LR: In this method, proposed by Thissen, Steinberg, and Wainer (1993), item parameters are estimated for the focal and reference groups. For the item parameters, constrained and extended models are generated. While in the constrained model it is assumed that item parameters are equal for both groups, in the extended model it is assumed that item parameters for each tested item are different for focal and reference groups and the same for all other items. The likelihood ratio is calculated for the constrained and extended models for each item, and the null hypotheses are tested for these values (Thissen, 2001).

Lord's $\chi 2$: In the Lord's $\chi 2$ method, variance and covariance of items are calculated for the focal and reference groups in order to detect DIF. These values calculated for the two groups are scaled for the purpose of comparison. These scaled values are calculated by using Lord's $\chi 2$. Then, the null hypothesis of no DIF is tested by comparing with critical values and it is decided whether DIF exists or not (Cromwell, 2002).

Raju's Area Measure: In this method, proposed by Raju (1990), item characteristic curves are considered while detecting DIF. In the calculation stages, item characteristic curves are drawn based on the probability of correct response to the item for focal and reference groups. If the probabilities of responding to the item are different for two groups, a specific area occurs between the curves, and this area is defined as the area index.

In a test, it is important not only to detect DIF, but also to decide what will be done after detecting items with DIF. It may be required to purify DIF items in order to provide unbiasedness. However, if the item is compulsory or essential for a latent trait or construct, it may not be appropriate to remove the item. Sometimes, editing a relevant item may result in removing DIF, although sometimes this solution may not be enough (Golia, 2015). When items with DIF exist in the test, it is known that these items will affect test statistics, results, and individual scores; however, it is not known what the effect will be (Li & Zumbo, 2009). If it is decided to purify the item from the test, the validity of the test may decrease, depending on the decreasing number of items of test. Moreover, the level at which purifying items with DIF will affect the ability estimation cannot be predicted. In this study, this is

the question to answer. Also, the effects of purifying items with middle level (B) DIF from the test are examined.

In the literature, studies exist about how test statistics change when items are discarded from the test in the case of dichotomous scoring (Lee & Zhang, 2017; Li & Zumbo, 2009; Roznowski & Reith, 1999; Rupp & Zumbo, 2003, 2006; Wells, Subkoviak & Serlin, 2002) and polytomous scoring (Golia, 2010, 2015; Tennant & Pallant, 2007). Some of these studies examined cases within the context of item parameter invariance (Roznowski & Reith, 1999; Rupp & Zumbo, 2003, 2006; Well, Subkoviak & Serlin, 2002), and some of these regard the cases as parameter invariances within the context of DIF as is the case in this current study (Golia, 2010, 2015; Lee & Zhang, 2017; Li & Zumbo, 2009; Tennant & Pallant, 2007). It can be stated that the studies in this direction are limited. Tennant and Pallant (2007) examined the effects of discarding items with uniform DIF from the test. The results of this study, which was conducted on five categorical items, found that discarding items in significant levels causes differences in individual and group levels. Li and Zumbo (2009) focused on the number of items with DIF and the size of DIF conditions in their study, which aimed to investigate the impacts of keeping and discarding items with uniform DIF. In the study, it was pointed out that when there are few items with DIF and a low size of DIF, even if the items in the test show DIF, the error and the effect size do not change significantly; when the size of DIF increases, discarding items with DIF from the test increases the error. Golia (2010) considered the effects of keeping and discarding three items with uniform DIF in different sizes and found that if there are few items with DIF, keeping them in the test does not affect ability estimations negatively; on the contrary, discarding them from the test has a negative impact on ability estimations. Golia (2015) also studied the effects of having items with DIF in a 15-item test and indicates that when there are three items with DIF or the size of DIF is large, the ability estimation is affected by these conditions. Lee and Zhang (2017) studied uniform DIF and investigated the conditions of the ratio of items with DIF and the existence of items with B and C levels. They also determined items with DIF by using MH methods in their study and they found that when the ratio of items with DIF increased, the ability estimations differed in individual and group levels. Moreover, the study shows that if the items with DIF are in C level, then the ability differences between reference and focal groups will be larger. Similar to this current study, several studies have compared DIF detection methods in the literature. Finch (2005) has compared the methods of MH, SIBTEST, IRT-LR, and MIMIC by considering the ratio of items with DIF. This study indicated that the method of IRT-LR was affected more than other methods when the ratio of items with DIF increased. Finch and French (2007) studied non-uniform DIF and compared the methods of logistic regression, SIBTEST, IRT-LR, and confirmatory factor analysis with the variables of DIF size, sample size, ability distribution, and IRT model. The study, which was conducted on 30 dichotomous items, showed that SIBTEST was the best in terms of Type 1 error and power, but factors that were manipulated did not have significant impact on the methods in terms of Type 1 error. Atalay Kabasakal, Arsan, Gök, and Kelecioğlu (2014) compared the methods of MH, SIBTEST, and IRT-LR in a simulation study conducted on uniform DIF. In this study, the ratio of items with DIF was studied and effect size of DIF was fixed at B level. The results of the study, conducted on dichotomously scored items, indicated that the largest Type 1 error was in SIBTEST method and the smallest Type 1 error was in the IRT-LR method. It also showed that when the ratio of items with DIF was increased, the error increased in IRT-LR and SIBTEST methods, with a larger increase in the SIBTEST method.

This study is different from the other simulation studies (Golia, 2015; Lee & Zhang, 2017; Li & Zumbo, 2009) in terms of the method used to detect DIF, number of items in the test, and number of response categories; from this point of view, it aims to evaluate the conditions. This has not been previously covered in the literature. This research also differs from other studies in the literature in terms of purifying the DIF items identified in the methods.

Purpose of the Study

In this study, the aim is to investigate how the errors will change depending on the ability estimates for the DIF detection methods -SIBTEST, IRT-LR, Lord's $\chi 2$ and Raju's area measures- when the items with DIF are purified from the test under the ratio of the number of items with DIF, effect size of DIF, and type of DIF.

METHOD

Research Design

Because the performances of different DIF detection methods are examined under specific conditions and based on the ability estimation obtained by purifying items with DIF from the test, this study was conducted as a Monte Carlo simulation study.

Simulation Conditions

The study investigates DIF detection methods—SIBTEST, IRT-LR, Lord's $\chi 2$ and Raju's area measures—through purifying items with DIF according to ratio of items with DIF, effect size of DIF (for SIBTEST and IRT-LR), and the type of DIF. The reason for choosing these four methods in the research is that they are frequently preferred in DIF researches and they are curious about the performance of these methods when item purifying applied. Atalay Kabasakal et al. (2014), Finch (2005), Finch and French (2007), and Lopez's (2012) studies investigated DIF according to IRT and even though SIBTEST is a CTT-based and a non-parametric method they have used SIBTEST method in their studies. For this reason SIBTEST was included in the current study. Hence, Finch (2005) compared the IRT-based IRT-LR method and the SIBTEST method in his study and pointed out that the SIBTEST provided effective results for the short tests. Also, researchers have included the SIBTEST method in a DIF study based on IRT and CAT (Lei, Chen, & Yu, 2006).

In the current study, sample size, test length, ability distribution, item type, and type of IRT model are constant. In the first place, Item type, test length, and IRT model are determined as simulation conditions. Thirty dichotomous items (1-0) were generated according to 3PLM (the three parameter logistic model), which considers the case of responding correctly by chance. Thirty-item tests were selected because the number of items is close to the number of items in high stakes tests in Turkey. Moreover, Downing and Haladyna (2004) indicate that usually a minimum of 30 items are used in achievement tests in order to be representative for the investigating area. Glas and Meijer (2003) used 30 items for the short test form in their simulation study conducted with item response theory. Suh (2016) also created a 30-item test form in their study about multidimensional IRT and DIF.

Secondly ability distribution and sample size are decided as simulation conditions. Ability parameters consisting of 1000 people were generated using normal distribution. Shepard, Camilli, and Averill (1981) stated that it is required to use at least 1000 people in order to obtain stable results.

In this study, the first condition tested for impact was the ratio of the items with DIF. The ratio of the items with DIF was determined to be 10% and 20%. Narayanan and Swaminathan (1994) stated that a 20% DIF item ratio is the worst scenario. In their research, Jodoin and Gierl (2001) studied the 10% and 20% items with DIF ratios. Thus, in 30-item tests, three and six items were made with DIF. The second condition tested for impact was the effect size of DIF. The effect sizes were examined in two ways as C level and B & C level for the methods of IRT-LR and SIBTEST. B & C and C levels were included in the study in order to evaluate the effect of items with middle level (B level) DIF on the ability estimation. The types of DIF were examined through the determination of uniform DIF, non-uniform DIF, and both uniform and non-uniform DIF. The simulation conditions are summarized in Table 1.

Table 1. Simulation Conditions

	Rates of items with DIF	SIBTEST		IRT-LR		Lord χ^2	Raju Area Measure
		B Level	B & C Level	B Level	B & C Level		
Non-uniform	10%	√	√	√	√	√	√
	20%	√	√	√	√	√	√
Uniform	10%	√	√	√	√	√	√
	20%	√	√	√	√	√	√
Non-uniform and uniform	10%	√	√	√	√	√	√
	20%	√	√	√	√	√	√

Data Generation

Firstly, item parameters were generated. In accordance with 3PLM, item parameters were obtained through the software WINGEN 3 (Han, 2007). While generating parameters, the item parameters that are usually encountered in real test applications were used. From the item parameters, a discrimination parameter was generated using lognormal distribution with a mean of 0 and a standard deviation of 0.2; the difficulty parameter was generated by normal distribution with a mean of 0 and standard deviation of 1; the guessing parameter was generated by beta distribution with an a-value of 8 and a b-value of 32. Kim and Lee (2004) also used similar distributions and values while obtaining test forms in their simulation study. The generated test form is shown in Table 2.

Table 2. Item Parameters in the Test Form

Item No	Model	Number of Cathogory	a	b	c	Item No	Model	Number of Cathogory	a	b	c
1	3PLM	2	1.130	-.727	.216	16	3PLM	2	1.114	-1.353	.322
2	3PLM	2	.791	-1.606	.241	17	3PLM	2	1.384	-1.817	.125
3	3PLM	2	1.491	0.928	.197	18	3PLM	2	1.118	.361	.222
4	3PLM	2	1.252	.348	.173	19	3PLM	2	.911	.276	.273
5	3PLM	2	1.236	1.488	.177	20	3PLM	2	1.723	-.044	.208
6	3PLM	2	.913	-2.291	.151	21	3PLM	2	.993	.525	.336
7	3PLM	2	.824	-.840	.122	22	3PLM	2	1.045	.207	.239
8	3PLM	2	.680	-1.333	.178	23	3PLM	2	.785	.591	.159
9	3PLM	2	1.008	-.669	.088	24	3PLM	2	.963	.064	.213
10	3PLM	2	1.128	-.253	.201	25	3PLM	2	1.259	.047	.116
11	3PLM	2	.781	1.036	.145	26	3PLM	2	.933	-1.285	.267
12	3PLM	2	.994	1.524	.162	27	3PLM	2	1.109	.984	.148
13	3PLM	2	.822	.464	.261	28	3PLM	2	1.077	-.296	.171
14	3PLM	2	.957	1.879	.146	29	3PLM	2	.952	-.462	.164
15	3PLM	2	1.106	-.267	.195	30	3PLM	2	.949	.947	.219

After generating item parameters, ability parameters were generated by normal distribution with a mean of 0 and standard deviation of 1. For the tests consisting of uniform and non-uniform or both types of DIF items, the ability parameters were obtained similarly. Mazor, Clauser, and Hambleton (1993) examined non-uniform DIF and generated abilities for a reference group with a similar distribution and values. In order to make sure that the results are stable, this was repeated 50 times in the study. Harwell, Stone, Hsu, and Kirisci (1996) reported that this should be repeated at least 25 times in Monte Carlo simulation studies. Finally, 1-0 data were created by applying the items to the individuals.

The obtained 1-0 data were rescaled using the software PARSCALE 4.1 (Muraki & Bock, 2003). This process was done to obtain 50 ability parameters by using items without DIF and to fix abilities for each condition. The a-parameter was increased by .75 for displaying some items in the test to display non-uniform DIF. A similar rate was used in the study of Mazor, Clauser and Hambleton (1993). They stated that by considering the b-parameter, the difference in a-parameter over a value

of .50 increased the rate of detection. Furthermore, the b-parameter was increased by .60 for displaying items in the test uniform DIF. Because the rate of DIF item conditions were being examined, in the first case, this process was applied to three items (Items 7, 12, and 26) and in the second case it was applied to six items (Items 6, 9, 12, 17, 21, and 29).

For displaying both uniform and non-uniform DIF items in the test, in the case of three items, DIF b-parameters of two items were increased by .60 and the a-parameter of one item was increased by .75; in the case of six items, DIF b-parameters of four items were increased by .60 and a-parameters of two items were increased by .75. DIF was randomly assigned to the items. Items with DIF were applied to an individual by using WINGEN; thus, 1-0 data were obtained for focal and reference groups. Simulation conditions were checked by comparing the parameters obtained from focal and reference groups.

Data Analysis

Binary data of focal and reference groups were analyzed using SIBTEST (Li & Stout 1994), IRTLRDIF (Thissen, 2001), and the difR package in R software (Magis, Beland, Tuerlinckx, & De Boeck, 2010; Magis, Beland, & Raiche 2013). For each condition in the SIBTEST and IRTLRDIF software, items with C level DIF and then items with B & C level DIF were removed from the response matrix and estimated using PARSCALE 4.1 software. Using the difR package, items that demonstrated significant DIF according to Lord $\chi 2$ and Raju's area measures were removed from the response matrix and estimated similarly with PARSCALE 4.1 software. In order to compare the methods, root mean squared difference (RMSD) and the coefficient of concordance (Pearson correlation coefficient) were calculated from estimated and initial abilities. Below, the criteria used are explained in detail.

RMSD (root mean squared difference)

To calculate RMSD, first the square of the difference between estimated and real ability values were found and summed. After that, this value was divided by the frequency of ability level and the square root of the result was calculated. The following is the equation of the RMSD:

θ: Real ability level
θ^*: Estimated ability level
f: Frequency of ability level

$$\text{RMSD} = \sqrt{\frac{\sum_i f_i(\theta^* - \theta)^2}{\sum_i f_i}} \quad (1)$$

Coefficient of concordance

The coefficient of concordance was calculated depending on the mean of Pearson correlation coefficients between estimated and real abilities of an individual.

In order to determine the effectiveness of DIF detecting methods, all RMSD values and coefficients of concordance that were obtained as a result of repetition according to simulation conditions were examined with the significance tests. For this, firstly the normality of data according to DIF detecting methods were examined and, if the normality conditions were not met, the methods were compared using a Kruskal-Wallis H test. Group comparisons were made by nonparametric multiple comparison test. The $\eta 2$ value was calculated to determine the effect of DIF detecting methods on RMSD and coefficient of concordance coefficients. The size of the eta square of .01, .06 and .14 respectively shows small, medium and large effect size (Green & Salkind, 2005). The following is the equation of the $\eta 2$:

$\chi 2$: Chi square value

N: Sample size

$$\eta 2 = \chi 2 / (N-1)$$

RESULTS

The research results were examined within the framework of the research question and the DIF detecting methods were compared using the error (RMSD) and coefficient of concordance.

The results, obtained from detecting items with DIF and removing them with the different methods according to 10% and 20% item rates and uniform, non-uniform, and both uniform and non-uniform DIF types, are shown in Table 3.

Table 3. The Coefficients of Error and Concordance for DIF Conditions

	DIF	SIBTEST		IRT-LR		Lord χ^2		Raju Area Measure	
	Rates	RMSD	Pearson	RMSD	Pearson	RMSD	Pearson	RMSD	Pearson
Non-Uniform	%10	.581435	.751599	.584374	.748612	.586027	.746705	.610559	.714193
	%20	.585285	.747123	.598210	.734050	.598814	.733239	.599150	.732580
Uniform	%10	.579508	.753530	.511010	.781589	.583243	.749699	.586162	.746380
	%20	.590214	.742381	.565683	.753397	.589310	.742946	.587441	.744388
Non-Uniform	%10	.578935	.753578	.521621	.777584	.578815	.753490	.579444	.752800
&Uniform	%20	.587103	.745318	.602092	.726336	.592590	.739431	.593482	.738539

Table 3 illustrates that when the rate of DIF items increases, removing DIF items increases the error. Only when using Raju's area measures for the non-uniform DIF type, removing DIF items decreased the error when the rate of DIF items increased. As a result of removing items with DIF in all conditions, the method of IRT-LR showed the minimum error in the 10% rate of DIF and uniform DIF type. If the coefficients of concordance were examined, after removing DIF items, the method of IRT-LR showed the maximum correlation in the 10% rate of DIF and uniform DIF type. Furthermore, it is possible to state that, generally, for all types of DIF, correlation coefficients calculated by removing DIF items decrease when the rate of DIF increases. Only in the condition of non-uniform DIF does the coefficient of concordance calculated as a result of removing DIF items increase according to the rate of DIF for the Raju method. Table 4 shows whether the RMSD and the coefficients of concordance have a significant difference according to the DIF detection method.

Table 4. The Results of Kruskal-Wallis H Test of RMSD and Coefficients of Concordance According to DIF Detecting Methods

	DIF detection method	N	Mean Rank	df	χ^2	p	Difference
RMSD	SIBTEST	300	549.53	3	10.584	.014	SIBTEST - Raju Area Measure
	IRT-LR	300	595.53				
	Lord χ^2	300	623.47				
	Raju Area Measure	300	633.47				
Pearson	SIBTEST	300	653.77	3	11.684	.009	SIBTEST - Lord χ^2 SIBTEST - Raju Area Measure
	IRT-LR	300	606.14				
	Lord χ^2	300	577.12				
	Raju Area Measure	300	564.98				

Table 4 shows that there is a significant difference between coefficients of RMSD obtained from the simulation conditions according to DIF detecting methods [$\chi 2$=10.584, p=.014]. The nonparametric multiple comparisons which were conducted to investigate which groups this difference occurs between indicate that the difference in RMSD coefficients are between the methods of SIBTEST and Raju's area measures. Therefore, it can be stated that the mean rank of SIBTEST (549.53) is lower than the mean rank of Raju area measure (633.47). In addition, the median of SIBTEST (.585) is

lower than the median of Raju area measure (.588). This means that the error value (RMSD) of SIBTEST is lower than Raju area measure. The η2 value was calculated to determine the effect of DIF detecting methods on RMSD coefficients. Consequently, the effect size (η2=.01) was found to be low (Green & Salkind, 2005). Similarly, it can be seen that there is a significant difference between coefficients of concordance obtained from the simulation conditions according to DIF detecting methods [χ^2=11.684, p=.009]. The nonparametric multiple comparisons, which were conducted to investigate which groups this difference occurs between, indicate that the difference in concordance coefficients are between the methods of SIBTEST and Lord χ2, as well as SIBTEST and Raju's area measures. Therefore, it can be stated that the mean rank of SIBTEST (653.77) is higher than the mean ranks of Raju area measure (564.98) and Lord χ2 (577.12). In addition, the median of SIBTEST (.749) is higher than the medians of Raju area measure (.745) and Lord χ2 (.744). This means that the coefficient of concordance of SIBTEST is higher than Raju area measure. The η2 value was calculated to determine the effect of DIF detecting methods on concordance coefficients; thus, the effect size (η2=.01) was found to be low level (Green & Salkind, 2005).

In order to assess the effect of purifying items with B level DIF from the test on ability estimation, firstly items with C level DIF and then items with B & C level DIF in the methods of SIBTEST and IRT-LR were extracted from test; the abilities were estimated later. The error and coefficient of concordance values calculated from the ability levels which were obtained in both cases are shown in Table 5.

Table 5. The Effect of Extracting B-Level DIF Items on the Error and Concordance Coefficients

			SIBTEST		IRT-LR	
		DIF Effect Level	RMSD	PEARSON r	RMSD	PEARSON r
Non-uniform	10 %	C	.576762	.756118	.380603	.839445
		B & C	.581435	.751599	.584374	.748612
	20 %	C	.576233	.756162	.583750	.744341
		B & C	.585285	.747123	.598210	.734050
Uniform	10 %	C	.574934	.757978	.000000	1.00000
		B & C	.579508	.753530	.511010	.781589
	20 %	C	.570526	.761617	.000000	1.00000
		B & C	.590214	.742381	.565683	.753397
Non-uniform and uniform	10 %	C	.572760	.759623	.046230	.980451
		B & C	.578935	.753578	.521621	.777584
	20 %	C	.569988	.762370	.081300	.966065
		B & C	.587103	.745318	.602092	.726336

Table 5 shows that in the methods of SIBTEST and IRT-LR the error values obtained from purifying C level DIF items are lower than the errors obtained from purifying B & C level DIF items when the rate of DIF items are 10% and 20% and when the type of DIF changes. Both methods at the rate of 10% and 20% DIF showed that the correlation coefficients calculated by purifying C level DIF items in all DIF type conditions were higher than the correlation coefficients calculated by purifying B & C level DIF items.

DISCUSSION and CONCLUSION

This study aims to investigate the effect of purifying DIF items from a test by using different DIF detection methods on individuals' ability estimates. For this purpose, a simulation study was conducted and firstly item parameters and depending on this the ability parameters were generated. In the fifty-replication study, the data set were generated according to 1000 participants' responses to 30 items and the ability estimates were rescaled after purifying items with DIF.

The abilities determined and scaled through items without DIF are accepted as real abilities. The cases of 10% and 20% DIF items rates in the uniform, non-uniform and both uniform and non-uniform DIF types were examined. Different methods to detect DIF (SIBTEST, IRT-LR, Lord's χ2,

Raju's area measures) were used and discussed the effects of these methods on ability estimations. For two conditions in three items with DIF and six items with DIF, the abilities were estimated again after purifying DIF items determined by the methods and then the concordance and error coefficients were calculated according to each method. For the methods of SIBTEST and IRT-LR, purifying only C-level DIF items the ability estimates were calculated and then purifying B & C level DIF items the abilities were estimated. Since there is no such distinction for the methods of Lord's $\chi 2$ and Raju's area measures, the values were compared by purifying DIF items at one time.

DIF is caused by the fact that the probability to respond an item correctly of a group is more or less relative to other group depends on not the ability level but the group (Osterlind, 1983; Zumbo 1999). Therefore, the existence of DIF items in the test can cause bias and error in individuals' ability estimations (Camilli, 1993). In other words, DIF is an indicator of systematic error of measurement (Camilli & Shepard, 1994; Kelecioğlu, Karabay & Karabay, 2014). Although DIF items are threats for the validity, since DIF items will cause a bias in ability estimation (Golia, 2015) purifying items may be seen as an appropriate solution to estimate abilities accurately. Lee and Zhang (2017) have found differences in estimations of ability when the ratio of items with DIF increased. Golia (2015) examined how the ability estimations would change in instrument that belonged the polytomously scored items with DIF. If the test belonged more than one items with DIF, there was a significant bias in estimations of ability. Golia (2010) investigated the effects of keeping and purifying three items with uniform DIF in 15 items tests and found that the goodness of ability estimations was not influenced by this condition when the test belonged a few number of items with DIF. Li and Zumbo (2009) studied on the number of items with DIF and the size of DIF by conducting a simulation study. They pointed out that if there was quite a little number of items with DIF or there was a small number of items with DIF and the size of DIF was small, then there was no bias in ability estimations. They also observed that when the number of items with DIF and the size of DIF increased then the errors changed. The studies indicated that if the size of DIF and the ratio of DIF increase, this increase causes the bias in ability estimations. Therefore, in the current conducted study the effects of purifying items with DIF which are determined by the DIF detecting methods were examined when the ratio of items with DIF 10% and 20%. In this way, not only the effects of purifying items with DIF from the test were observed but also the DIF detecting methods were compared. Concordance and error between ability estimation after purifying item which is detected as with DIF through methods, and true abilities in the case of no items with DIF. Thus, the results state that the error which shows the ability estimation differences, increases when the ratio of items with DIF even if these items are discarded. Tennant and Pallant (2007) indicated that there may be differences in individual ability estimations after purifying items with DIF. Similarly, Golia (2010) studied on polytomously (6) scored 15 items and pointed out that purifying 3 items with DIF from the test negatively affected ability estimations.

According to findings, purifying items with DIF determined by the method of IRT-LR yielded the most concordant and the least inaccurate results with the real abilities. The highest error and the lowest concordance were obtained in the estimation through excluding items with DIF determined by the method of Raju's area measure. When the number of items with DIF increases, errors generally increase but in the method of Raju's area measure the error may decrease. Atalay Kabasakal, Arsan, Gök and Kelecioğlu (2014) compared DIF detecting methods (MH, SIBTEST, IRT-LR) in a simulation study and found that IRT-LR method had the smallest error. In this study which compares methods according to ability estimations, the similar relationship was found in RMSD and Pearson Correlation concordance index. On the other hand, Finch (2005) compared the methods of MH, SIBTEST, IRT-LR and MIMIC and stated that the increase in the number of items with DIF was more effective on IRT-LR method. However, in some different studies under the different conditions different results were obtained according to methods. Therefore, it will be more appropriate to discuss which method under which conditions gave results with the highest concordance and the lowest error. Considering the error and concordance in the nonparametric comparisons based on ability estimations under the conditions of this study, SIBTEST & Lord's $\chi 2$ and SIBTEST & Raju's area measure produced different results. Finch and French (2007) conducted

a study on nonuniform DIF and compared the methods LR, SIBTEST, IRT-LR and confirmatory factor analysis. They indicated that DIF size, sample size, ability distributions and IRT model had no significant impact on methods when the error was considered. In the current study, it was found that the manipulated factors did not cause a significant difference for the methods of IRT-LR and SIBTEST.

The methods of Lord's $\chi 2$ and Raju's area measures are based on the parameter estimations. Therefore, while determining DIF these methods may be affected by the algorithms used in item parameter estimations (Cohen & Kim, 1993). As a result of this, it is thought that the concordance coefficients of these methods may be lower than the others. Furthermore, in the method of Raju's area measure the situation of when the number of items with DIF increases the error decreases may be caused by the characteristics that the methods are based on.

In this study, for only the methods of SIBTEST and IRT-LR, both the cases of excluding C-level DIF items and the case of excluding B & C level DIF items were examined and compared. In the methods of SIBTEST and IRT-LR under the conditions of 10% and 20% DIF items ratio, when only C-level DIF items were extracted, the error ratio was found to be lower and the concordance index were found to be higher. Lee and Zhang (2017) remark that when the items with DIF is in C level instead of B level, the difference in ability estimations will be larger. The results support this finding. Since items in B level do not affect ability estimations negatively as in C level, keeping B level items in test may decrease the error of ability estimations. Furthermore, purifying items in B and C level decreases the number of items in test. This situation may cause finding the larger error after purifying items in B and C level. In this situation, for SIBTEST and IRT-LR under this condition, it can be said that the error of ability estimation increases when items with B-level DIF are extracted from the test. Therefore, for the conditions in this study it may be suggested that items with B-level DIF should not be excluded from the test in the methods of SIBTEST and IRT-LR.

In the scope of this study, for the investigation of the effect of purifying DIF items from the test on the ability estimations, different methods were compared according to uniform, non-uniform, both uniform and non-uniform DIF types under the 10% and 20% DIF item ratios. There were differences between the methods in terms of the error and concordance coefficients. Further studies may repeat this under similar conditions by using different IRT estimation methods. Moreover, when the conditions and methods change the obtained results will be different. Therefore, the effect of purifying items with DIF on ability estimations may be examined under different conditions and using different methods.

REFERENCES

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

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 397–418). New York: Routledge.

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

Cheng, C. M. (2005). *A study on Differential Item Functioning of the basic mathematical competence test for junior high schools in Taiwan* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3189625).

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-47. doi: 10.1111/j.1745-3992.1998.tb00619.x

Cohen, A. S., & Kim, S. (1993). A comparison of Lord's Chi Square and Raju's Area Measures in detection of DIF. *Applied Psychological Measurement, 17*(1), 39–52. doi: 10.1177/014662169301700109

Cromwell, S. (2002, February). *A primer on ways to explore item bias*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX.

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327-333. https://doi.org/10.1046/j.1365-2923.2004.01777.x

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

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in Item Response Theory models. *Applied Psychological Measurement, 27*(3), 217-233. doi: 10.1177/0146621603252216

Golia S. (2010). The assessment of DIF on Rasch measures with an application to job satisfaction. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation, 1*(1) 16–25. doi: 10.1285/i2037-3627v1n1p16

Golia S. (2015). Assessing the impact of uniform and nonuniform differential item functioning items on Rasch measure: The polytomous case. *Computational Statistics, 30*, 441–461. doi: 10.1007/s00180-014-0542-x

Green, S., & Salkind, N. (2005). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (4th Ed). Upper Saddle River, NJ: Pearson.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. California: Sage.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459. doi: 10.1177/0146621607299271

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in Item Response Theory. *Applied Psychological Measurement, 20*(2), 101-125. doi: 10.1177/014662169602000201

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. doi: 10.1207/S15324818AME1404_2

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

Kim, S., & Lee, W. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report 2004-5). Iowa City, IA: Act, Inc.

Kristanjansonn, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response model. *Educational and Psychological Measurement, 65*(6), 935-953. doi: 10.1177/0013164405275668

Lee, Y. H., & Zhang, J. (2017). Effects of differential item functioning on examinees' test performance and reliability of test. *International Journal of Testing, 17*(1), 23-54. https://doi.org/10.1080/15305058.2016.1224888

Lei, P-W., Chen, S-Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*(3), 245-264. https://doi.org/10.1111/j.1745-3984.2006.00015.x

Li, H. H., & Stout, W. (1994). SIBTEST: A fortran V program for computing the simultaneous item bias DIF statistics. Department of Statistics, University of Illinois, Urbana Champaign.

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

Lopez, G. E. (2012). *Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-Likelihood Ratio test, Crossing-SIBTEST, and Logistic Regression procedures* (Doctoral dissertation). Retrieved from http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=5327&context=etd

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous Differential Item Functioning. *Behavior Research Methods, 42*(3), 847–862. doi: 10.3758/BRM.42.3.847

Magis, D., Beland, S., & Raiche, G. (2013). difR: Collection of methods to detect dichotomous Differential Item Functioning (DIF) in psychometrics. R package version 5.0. http: //www.CRAN.R-project.org/package=difR

Mazor, K. M., Clauser, R. E., & Hambleton, R. K. (1993, March). *Identification of nonuniform Differential Item Functioning using a variation of the Mantel-Haenszel procedure.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105–118. doi: 10.3102/10769986007002105

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037/0003-066X.50.9.741

Muraki, E., & Bock, R. D. (2003). PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales [Computer software]. Skokie, IL: Scientific Software International, Inc.

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

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

Osterlind, S. J. (1983). *Test item bias.* Newbury Park, California: Sage.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5–15. doi: 10.1111/j.1745-3992.2000.tb00033.x

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207. doi: 10.1177/014662169001400208

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*(2), 215-230. Retrieved from http://www.jstor.org/stable/1435184

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*(2), 248-269. doi: 10.1177/00131649921969839

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research, 49*, 264-276.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84. doi: 10.1177/0013164404273942

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*(4), 317–375. doi: 10.2307/1164616

Suh, Y. (2016). Effect size measures for differential item functioning in a multidimensional IRT model. *Journal of Educational Measurement, 53*(4), 403-430. https://doi.org/10.1111/jedm.12123

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if Differential Item Functioning makes a difference. *Rasch Measurement Transactions, 20*(4), 1082-1084.

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio tests for Differential Item Functioning. L.L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, NC.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*(1), 77–87. doi: 10.1177/0146621602261005

Zumbo, B. D. (1999). *Handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and likert-type item scores.* Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Değişen Madde Fonksiyonu Belirlemede Yöntemlerin Yetenek Kestirimine Dayalı Performansları: Bir Benzetim Çalışması

Giriş

Madde yanlılığın incelenme süreci daha nesnel sonuçlara dayanan ve madde yanlılığının bir ölçüsü olabilecek değişen madde fonksiyonunun (DMF) incelenmesi ile başlar. DMF aynı yetenek

düzeyinde fakat farklı gruplardaki kişilerin bir maddeyi doğru yanıtlama olasılıklarının birbirinden farklılaşması olarak tanımlanmaktadır (Hambleton, Swaminathan ve Rogers, 1991). DMF ortaya çıkışı açısından tek biçimli (uniform) ve tek biçimli olmayan (non-uniform) fonksiyonlar şekilde sınıflandırılır (Mellenbergh, 1982). Bu farklılaşmanın temelinde yatan gerçek ise yetenek düzeyi ile grup üyeliğinin birlikte maddeyi doğru yanıtlama olasılığını etkilemesidir. Buna göre tek biçimli DMF, aynı yetenek düzeyindeki iki grubun bir maddeye doğru yanıt verme olasılıklarının tüm yetenek düzeyleri için sabit bir değer olması durumunda meydana gelir. Buna karşın tek biçimli olmayan DMF ise aynı yetenekteki iki grubun maddeye doğru yanıt verme olasılıkları farklı yetenek düzeylerinde tutarsız olduğu durumda oluşur (Camilli ve Shepard, 1994; Penfield ve Lam, 2000; Zumbo, 1999).

DMF belirleme teknikleri temelde Klasik Test Kuramı (KTK) ve Madde Tepki Kuramına (MTK) göre sınıflandırılmaktadır. KTK'ya göre DMF belirleme yöntemleri varyans analizi, ki-kare, dönüştürülmüş madde indeksi, lojistik regresyon, Mantel-Haenszel (MH) ve SIBTEST'tir. MTK yöntemleri ise Lord'un $\chi 2$'si, Raju'nun alan ölçüsü ve MTK-olabilirlik oranı (MTK-OO)'dır (Camilli ve Shepard, 1994; Oshima ve Morris, 2008).

Bir testte DMF'nin belirlenmesin yanında DMF gösteren madde bulunduğunda ona ne yapılacağına karar verilmesi önemlidir. Yansızlığı sağlamak adına ilgili maddenin testten çıkarılması gerekebilir. Buna karşın ilgili madde ölçülen örtük özellik ya da yapının önemli ya da zorunlu maddesiyse maddenin atılması uygun olmayabilir. Bazen ilgili maddenin yeniden ifade edilmesi DMF'nin ortadan kalkmasını sağlayabilirken bazen bu çözüm yeterli olmayabilir (Golia, 2015). Testte DMF'li maddeler bulunduğunda bu maddelerin test istatistiklerini, sonuçları, bireylere ait puanları etkileyeceği bilinmekte fakat bu etkinin nasıl olacağı bilinmemektedir (Li ve Zumbo, 2009). Eğer maddenin testten çıkarılmasına karar verilirse, testteki madde sayısının azalmasına bağlı olarak testin geçerliliği düşürebilir. Bununla birlikte DMF'li maddelerin testten çıkarılmasının yetenek kestirimini hangi düzeyde etkileyeceği kestirilememektedir. Bu çalışmada bu soruya yanıt aramaktadır. Bununla birlikte orta (B) düzeydeki DMF'li maddelerin testten çıkarılmasının etkileri de incelenmektedir.

Alanyazında maddelerin ikili puanlandığı (Lee ve Zhang, 2017; Li ve Zumbo, 2009; Roznowski ve Reith, 1999; Rupp ve Zumbo, 2003, 2006; Wells, Subkoviak ve Serlin, 2002) ve çoklu puanlandığı (Golia, 2010, 2015; Tennant ve Pallant, 2007) durumlarda testten madde çıkarılmasının teste ilişkin istatistikleri nasıl değiştiğine dair çalışmalar bulunmaktadır. Bu çalışmaların bir kısmı madde parametreleri değişmezliği kapsamında bu durumu incelerken (Roznowski ve Reith 1999; Rupp ve Zumbo, 2003, 2006; Well, Subkoviak ve Serlin, 2002), bazıları ise ilgili durumu bu çalışmada olduğu gibi DMF kapsamında parametre değişmezliği olarak ele almıştır (Golia, 2010, 2015; Lee ve Zhang, 2017; Li ve Zumbo, 2009; Tennant ve Pallant, 2007).

Bu araştırmada DMF belirleme yöntemlerinden SIBTEST, MTK-OO, Lord'un $\chi 2$'si ve Raju'nun alan ölçüsünün DMF'li madde oranı ve DMF etki büyüklüğü altında DMF'li maddelerin testten çıkarılması durumunda yetenek kestirimine dayalı olarak hataların nasıl değiştiğinin incelenmesi amaçlanmaktadır.

Yöntem

Araştırmada farklı DMF belirleme yöntemlerinin performansları, belirli koşullar altında DMF'li maddelerin testten çıkarılmasıyla elde edilen yetenek kestirimine dayalı olarak incelendiğinden bir Monte Carlo benzetim çalışması yürütülmüştür.

Araştırma SIBTEST, MTK-OO, Lord $\chi 2$, Raju'nun alan ölçüleri DMF belirleme yöntemlerini DMF'li madde oranları, DMF etki büyüklüğü (SIBTEST ve MTK-OO için) ve DMF türüne göre, tespit edilen DMF'li maddelerin testten çıkarılmasıyla incelemektedir. Bu araştırmada sıklıkla kullanılan DMF yöntemleri seçilmiştir. Bunun sebebi sıklıkla kullanılan bu yöntemlerin maddelerin testten çıkarılması durumundaki performanslarını belirlemektir. SIBTEST KTK'ya dayalı olması ve parametrik olmayan bir yöntem olmasına rağmen araştırmaya dahil edilmiştir. Bunun sebebi SIBTEST yönteminin Atalay Kabasakal vd. (2014), Finch (2005), Finch ve French (2007), Lopez

(2012) gibi araştırmacılar tarafından madde tepki kuramında gerçekleştirilen DMF çalışmalarına dahil edilmesidir. Nitekim Finch (2005) araştırmasında bir MTK yöntemi olan IRTLR ile SIBTEST yöntemini karşılaştırmış ve kısa testlerde SIBTEST'in etkili sonuçlar verdiğini belirlemiştir. CAT temelinde ve MTK'ya dayalı olarak gerçekleştirilen bir DMF araştırmasında (Lei, Chen ve Yu, 2006) da SIBTEST'e yer verildiği görülmektedir.

Örneklem büyüklüğü, test uzunluğu, yetenek dağılımı, madde türü, MTK model türü koşulları araştırmada sabit tutulmuştur. Araştırmada belirlenen koşullardan ilki madde türü, test uzunluğu ve MTK modelidir. Araştırmada ikili puanlanan (1-0) 30 madde şansla doğru cevaplama olasılığını da dikkate alan (Baker, 2001) 3PLM'ye göre oluşturulmuştur. 30 maddelik testler Türkiye'de geniş ölçekli testlerde karşılaşılan madde sayısına yakın olduğu için seçilmiştir. İkinci koşul yetenek dağılımı ve örneklem büyüklüğüdür. 1000 kişiden oluşan yetenek parametreleri normal dağılım kullanılarak oluşturulmuştur. Shepard, Camilli ve Averill (1981) kararlı sonuçlar elde edebilmek için en az 1000 bireyden oluşan örneklemler kullanılması gerektiğini belirtmiştir.

Araştırmada etkisi test edilen koşullardan ilki DMF'li madde oranıdır. DMF'li madde oranı %10 ve %20 olarak belirlenmiştir. Narayanan ve Swaminathan (1994) %20 DMF madde oranının testlerdeki en kötü senaryo olduğunu belirtmiştir. Böylece 30 maddelik testlerde 3 ve 6 madde DMF'li hale getirilmiştir. Etkisi test edilen ikinci koşul DMF etki büyüklüğüdür. MTK-OO ve SIBTEST yöntemleri için etki büyüklükleri C düzeyinde, B ve C düzeyinde olmak üzere iki durum altında incelenmiştir. C, B ve C düzeyleri orta düzeydeki (B düzeyi) DMF'li maddelerin yetenek kestiriminde bulunmasının etkisini değerlendirmek amacıyla araştırmaya dâhil edilmiştir. DMF türü tek biçimli, tek biçimli olmayan, hem tek biçimli hem tek biçimli olmayan DMF'nin tespiti üzerinden incelenmiştir.

Verilerin türetilmesi aşamasında öncelikle madde parametreleri 3PLM'e uygun olarak WINGEN 3 (Han, 2007) programıyla elde edilmiştir. Parametreler elde edilirken gerçek test uygulamalarında genellikle karşılaşılan madde parametreleri kullanılmıştır. Madde parametrelerinden ayırıcılık parametresi ortalaması 0, standart sapması ,2 olan lognormal dağılımla, güçlük parametresi ortalaması 0 standart sapması 1 olan normal dağılımla, şans parametresi ise a değeri 8, b değeri 32 olan beta dağılımıyla oluşturulmuştur.

Madde parametrelerinin türetilmesinin ardından ortalaması 0 standart sapması 1 olan normal dağılımla yetenek parametreleri türetilmiştir. Tek biçimli, tek biçimli olmayan ya da her iki DMF türündeki maddelerin bir arada yer aldığı testler için yetenek parametreleri benzer dağılımlarla elde edilmiştir. Sonuçların kararlılığından emin olmak amacıyla araştırmada 50 tekrar yapılmıştır. Harwell, Stone, Hsu ve Kirisci (1996) Monte Carlo benzetim çalışmalarında en az 25 tekrar kullanılması gerektiğini belirtmiştir. Son olarak bireylere maddeler uygulanarak 1-0 verilerinin elde edilmesi sağlanmıştır.

Elde edilen 1-0 verileri PARSCALE 4.1 (Muraki ve Bock, 2001) programıyla tekrar ölçeklenmiştir. Bu işlem 50 yetenek parametresinin DMF'siz maddeler üzerinden elde edilmesi ve her bir koşul için yeteneklerin sabitlenmesi için gerçekleştirilmiştir. Bazı maddelerin tek biçimli olmayan DMF göstermesi için a parametresi ,75 arttırılmıştır. Benzer oran Mazor, Clauser ve Hambleton (1993)'ın çalışmasında kullanılmıştır. Mazor, Clauser ve Hambleton (1993) b parametresi de dikkate alınarak a parametresinin ,50 üzerindeki farkının tespit oranını yükselttiği belirtilmiştir. Bunun yanında testteki maddelerin tek biçimli DMF göstermesi için b parametresine ,60 oranında arttırım uygulanmıştır. Bu işlem; DMF'li madde oranı koşulları incelendiği için ilk durumda 3 maddeye (7, 12 ve 26. maddeler), ikinci durumda ise 6 maddeye (6, 9, 12, 17, 21 ve 29. maddeler) uygulanmıştır. Testteki maddelerin hem tek biçimli hem de tek biçimli olmayan DMF göstermesi için ise 3 maddenin DMF'li olduğu durumda 2 maddenin b parametresine ,60 oranında, 1 maddenin a parametresine ,75 oranında; 6 maddenin DMF'li olduğu durumda 4 maddenin b parametresine ,60 oranında, 2 maddenin a parametresine ,75 oranında arttırım uygulanmıştır. DMF, maddelere seçkisiz olarak atanmıştır. DMF'li maddeler WINGEN programıyla bireylere uygulanmış ve böylece odak ve referans grupları için 1-0 verileri elde edilmiştir.

Odak ve referans gruplarına ait ikili puanlan veriler SIBTEST (Li ve Stout 1994), IRTLRDIF (Thissen, 2001) ve R programında yer alan difR (Magis, Beland, Tuerlinckx ve De Boeck, 2010; Magis, Beland ve Raiche 2013) paketi kullanılarak analiz edilmiştir. SIBTEST ve IRTLRDIF programlarında her koşul için öncelikle C ve sonrasında B ve C düzeyinde DMF'li bulunan maddeler cevap matrisinden çıkarılarak PARSCALE 4.1 programıyla kestirim yapılmıştır. difR paketi ile Lord $\chi 2$, Raju'nun alan ölçülerine göre anlamlı DMF gösteren maddeler cevap matrisinden çıkarılarak PARSCALE 4.1 programıyla benzer şekilde kestirim yapılmıştır. Yöntemleri karşılaştırabilmek için referans gruplar için kestirilen yetenekler ve ilk yetenekler üzerinden hata (RMSD) ve uyum katsayısı (Pearson korelasyon katsayısı) hesaplanmıştır.

DMF belirleme yöntemlerinin etkililiğini belirlemek amacıyla benzetim koşullarına göre yapılan tekrarlar sonucunda elde edilen tüm RMSD ve uyum katsayıları anlamlılık testleriyle incelenmiştir. Bunun için öncelikle verilerin DMF belirleme yöntemlerine göre normalliği incelenmiş ve normallik koşulları sağlanmadığından Kruskal-Wallis H testi ile yöntemler karşılaştırılmıştır. Yöntemler arasında ortaya çıkan farklılığın hangi yöntemlerden kaynaklandığını belirlemek üzere nonparametric çoklu karşılaştırma testi kullanılmıştır. $\eta 2$ değeri aracılığıyla ortaya çıkan farka ilişkin etki büyüklükleri hesaplanmıştır.

Sonuç ve Tartışma

Bu araştırma, farklı DMF belirleme yöntemleri kullanılarak bir testte DMF'li maddelerin çıkarılma durumlarının bireylerin yetenek kestirimine olan etkisini incelemeyi amaçlamaktadır. Araştırmanın sonuçlarına göre MTK-OO yöntemiyle belirlenen DMF'li maddelerin testten çıkarılması gerçek yeteneklerle en uyumlu ve en az hatalı sonucu vermiştir. En yüksek hata ve en düşük uyum ise Raju'nun alan ölçeği yöntemi ile belirlenen DMF'li maddelerin testten çıkarılmasıyla yapılan kestirimde görülmüştür. DMF'li madde sayısı arttığında hatalar genel olarak artarken Raju'nun alan ölçüleri yönteminde hata miktarı azalabilmektedir. Atalay Kabasakal, Arsan, Gök ve Kelecioğlu (2014) DMF belirleme yöntemlerini karşılaştırdıkları benzetim çalışmasında MTK-OO yönteminin Tip 1 hata dikkate alındığında en düşük hatayı verdiğini bulmuştur. Aynı çalışmada SIBTEST yöntemi güç açısından MTK-OO yönteminden daha üstün bulunmuştur. Yetenek kestirimleri üzerinden yöntemlerin karşılaştırıldığı bu çalışmada da benzer bir ilişki RMSD hata ve Pearson korelasyonu uyum indeksi açısından bulunmuştur. Diğer bir yandan Finch (2005), MH, SIBTEST, MTK-OO ve MIMIC yöntemlerini karşılaştırmış ve DMF'li madde sayısının arttığında MTK-OO'nun daha etkili olduğunu belirtmiştir. Ancak birçok farklı çalışmada farklı koşullar altında yöntemlere ilişkin farklı sonuçlar elde edilmektedir. Bu yüzden hangi yöntemin hangi koşullar altında en uyumlu ve en az hatalı sonuçlar verdiğini tartışmak daha doğru olacaktır. Bu çalışmanın koşulları altında yetenek kestirimleri üzerinden yapılan nonparametrik karşılaştırmalarda hata ve uyum dikkate alındığında SIBTEST ve Lord'un $\chi 2$'si ile SIBTEST ve Raju alan ölçüleri yöntemlerinin birbirlerinden farklı sonuçlar verdiği görülmektedir. Finch ve French (2007) çalışmalarında tek biçimli olmayan DMF'li maddeler üzerinde lojistik regresyon, SIBTEST, MTK-OO ve doğrulayıcı faktör analizi yöntemlerini karşılatırmış ve DMF büyüklüğü, örneklem büyüklüğü, yetenek dağılımı ve MTK modelinin hata açısından anlamlı bir etkisinin olmadığını belirtmiştir. Bu çalışmada da, manipüle edilen faktörlerin MTK-OO ve SIBTEST yöntemlerinde anlamlı bir farklılığa sebep olmadığı bulunmuştur.

SIBTEST ve MTK-OO yöntemleri için sadece C düzeyinde belirlenmiş maddeler atıldığı, B ve C düzeyinde belirlenmiş maddelerin birlikte atıldığı durumlar araştırmada incelemiş ve karşılaştırılmıştır. SIBTEST ve MTK-OO yönteminde hem %10 hem de %20 DMF'li madde oranı koşullarında sadece C düzeyinde madde atıldığı durumda hata oranı daha düşük ve uyum indeksi daha yüksek bulunmuştur. Bu durumda SIBTEST ve MTK-OO için bu çalışma koşulları altında B düzeyinde belirlenen DMF'lerin testten çıkarılması durumunda yetenek kestirimdeki hataların arttığı söylenebilir. Bu nedenle araştırmada yer alan koşullarda SIBTEST ve MTK-OO yöntemlerinde B düzeyindeki maddelerin testten çıkarılmaması önerilebilir. Lee ve Zhang (2017) araştırmasında

DMF'li maddelerin C düzeyinin altında olmasının testlerde daha düşük etki yaratacağını belirtmektedir.

DMF'li maddelerin çıkarıldığı testlerin bireylerin yetenek kestirimine olan etkilerinin araştırılmasında bu çalışma kapsamında tek biçimli, tek biçimli olmayan, hem tek biçimli hem tek biçimli olmayan DMF türünde %10 ve %20 DMF'li madde barındıran koşullarda farklı yöntemler karşılaştırılmıştır. Hata ve uyum katsayıları açısından yöntemler arasında farklılıklar bulunmuştur. Bundan sonraki çalışmalar benzer koşullarda farklı MTK kestirim yöntemleri kullanılarak tekrarlanabilir. Ayrıca koşullar ve yöntemler değiştikçe elde edilen sonuçlar farklılaşmaktadır. Bu yönde farklı koşullar ve yöntemler kullanılarak DMF'li maddelerin testten çıkarılmasının yetenek kestirimine etkisi incelenebilir.