ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);136-149

Differential Item and Differential Distractor Functioning Analyses on Turkish High School Entrance Exam*

Seviye Belirleme Sınavında Değişen Madde ve Değişen Çeldirici Fonksiyonu Analizleri

Ragip TERZİ** Levent YAKAR***

Abstract
Test fairness is one of the most critical elements in creating assessments. Differential item functioning (DIF) and differential distractor functioning (DDF) analyses play complementary roles in justifying test fairness. This study aims to investigate the correct (i.e., DIF) and incorrect (i.e., DDF) response choices of students based on gender in a standardized high-stakes test administered in Turkey. Given the purpose of this study, the Math section of 2011 Turkish High School Entrance Exam was investigated. For DIF analyses, Mantel Haenszel and Logistic Regression methods were used. For DDF analyses, two odds ratio approaches under the two-parameter logistic-nested logit model and nominal response model were used. According to the findings, in 500 and 1,000 sample sizes, DIF was not detected, however, only a 2,000 sample size indicated significant DIF results. Five DIF items were observed among 20 items, where three out of those five DIF items also showed DDF.

Key Words: DIF, DDF, SBS

Öz
Testin adil olması başarı testi hazırlamadaki vazgeçilmez unsurlardan biridir. Değişen madde fonksiyonu (DMF) ve değişen çeldirici fonksiyonu (DÇF) analizleri test adilliğini değerlendirmede birbirlerini tamamlayan rollere sahiptirler. Bu çalışma Türkiye'de yapılan geniş ölçekli bir teste katılan öğrencilerin cinsiyetlerine göre doğru ve yanlış yanıtlarını ve çeldirici seçimlerini DMF ve DÇF yöntemleri ile incelemeyi amaçlamaktadır. Bu amaçla 2011 yılı Seviye Belirleme Sınavı Matematik bölümü DMF ve DÇF açısından analiz edilmiştir. DMF analizleri için, Mantel-Haenszel ve Lojistik Regresyon metotları kullanılmıştır. DÇF analizleri için, iki parametreli lojistik kümelenmiş logit ve nominal yanıt modellerinin altındaki iki olasılık oranı yaklaşımları kullanılmıştır. Sonuçlara göre 500 ve 1,000 örneklem büyüklüğünde DMF bulunmazken sadece 2,000 örneklem büyüklüğünde anlamlı DMF sonuçları bulunmuştur. 20 maddelik testin beş maddesinde DMF ve bu beş maddenin üçünde de DÇF görülmüştür.

Anahtar Kelimeler: DMF, DÇF, SBS

INTRODUCTION

The Turkish Ministry of National Education initiated a high school entrance exam (SBS) for sixth- and seventh-graders in 2008 and eighth-graders in 2009. The purpose of administering the SBS each year was to closely assess whether students of those grades have met the requirements of each academic year. However, note that the SBS administration is now carried out only at the end of 8th grades, called another name, passing to high school from elementary (TEOG) until 2017. Both exams mainly have the same purpose of high school registration, ordering students based on their test scores. The SBS was a nation-wide, high stakes test administered to more than one million

* An early draft of this paper was presented at the annual meeting of Northeastern Educational Research Association, Trumbull, CT., USA. (2016, October).
** Ph.D., Harran University, School of Education, Sanliurfa, Turkey, email: terziragip@gmail.com, ORCID ID: 0000-0003-3976-5054.
*** Ph.D., Kahramanmaras Sutcu Imam University, School of Education, Kahramanmaras, Turkey, email: l_yakar@hotmail.com, ORCID ID: 0000-0001-7856-6926.

DOI: 10.21031/epod.368081
Retrieved: 18.12.2017
Accepted: 26.03.2018

students in Turkey every year (Ministry of National Education, 2011). The results from SBS were used for summative assessment purposes.

Given the high number of students who have taken such a high-stakes test should be prepared carefully. Haladyna and Downing (2004) classified construct-irrelevant variance into 21 potential sources of systematic errors, including test development, item format, and item quality. Thus, it is crucial but difficult to thoroughly investigate each measurement component of the SBS. There is still much left to explore in order to have valid and reliable measures of their achievement in the SBS that complies with minimal construct-irrelevant variance, even after test was already examined. Investigating problematic and misused parts of previous tests can help us to create better tests in future.

Writing objective items requires a lot of effort due to many reasons. One aspect of writing objective items in a test form requires that members of different examinee groups at the same level of ability be not negatively affected by the test. Furthermore, test analysis, particularly item analysis, can help evaluate how the objective items have served to assessment purposes. In other words, item analysis is a crucial way to get feedback about the quality of items. Moreover, item analysis should be routinely carried out by psychometricians or measurement experts while especially developing an item pool.

In general, designing a test includes three steps: writing test items, assembling and administering a test, and analyzing test items. Starting from the beginning towards the end of the test development, the whole process must be carefully carried out. Test fairness is an inevitable component in designing not only achievement tests but also any type of assessments. The investigation of measurement invariance plays a key role for the fairness of the test scores across the observed groups at the same ability level (Camilli, 2006). Otherwise, the item is considered as differentially functioning if the assumption of the measurement invariance is violated (Dorans & Holland, 1993). At this point, differential item functioning (DIF) and differential distractor functioning (DDF) are important analyses to be conducted for the justification of test fairness.

DIF and DDF can give useful information about the measurement invariance of a test, which refers to the degree whether the test behaves in the same way for groups given the same ability level (Dorans & Holland, 1993; Zumbo, 2007). In this context, DIF is used to describe the situation in which one group answered an item correctly more often than the other group at the same ability level (Zumbo, 2007). In other words, DIF can be used to check the stability of item performance among equally knowledgeable groups (DeMars & Lau, 2011). In addition to DIF, this study also provides additional analyses for differential distractor functioning (DDF). The concept of DDF was extended from the notion of DIF by Green, Crone, and Folk (1989). Similar to DIF conditional on the ability level, DDF can occur if a group is disproportionately attracted to a distractor due to a biasing factor in the distractor (Suh & Bolt, 2011). In DIF analyses, all correct answers are compared against all incorrect answers; whereas, only incorrect responses are examined in DDF analyses (Green et al., 1989). Because all distractors are incorrect, the choice of a distractor does not have any impact on test scores, however, DDF analyses can be informative for different subgroups (Abedi, Leon, & Kao, 2007). Therefore, DDF should be investigated along with DIF because DDF can cause DIF in correct responses (e.g., Penfield, 2008; Suh & Bolt, 2011; Terzi & Suh, 2015).

Examining DIF and DDF based on gender is important because Marshall (1983) found a significant interaction between gender and a choice of distractor in a large majority of items administered to 6^{th} grade students. In analyzing multiple-choice items, DIF and DDF analyses are very important because they are typically used for a variety of purposes such as refining existing test items, developing new scales, and validating test score inferences (Zumbo, 2007). Item bias may exist where there is DIF, which is not relevant to item quality and thus test purposes (Zumbo, 1999). Moreover, the presence of item bias is a matter of concern for test fairness, especially for high-stakes tests (Camilli, 2006).

Purpose of the Study

It is quite imperative to provide all students with conceptually and psychometrically sound assessments regardless of their differences except for ability. As explained earlier, the SBS was a nationally administrated exam, conducted by the Turkish Ministry of National Education. More than one million each sixth-, seventh- and eighth-grade students in Turkey took this exam every year. The SBS is a high-stakes test that has lasting consequences on students' futures. Even though many studies have examined measurement issues related to DIF and DDF, the impact of DIF along with DDF specifically on the assessment of the SBS has not been explored. Results from analyses of this test may give a clue about TEOG. The purpose of this study, therefore, is to conduct the DIF and DDF analyses of the SBS in which equity and test fairness are paramount to more than one million students.

METHODOLOGY

Data Source

In our study, 27,952 students without missing responses were randomly chosen from more than one million eight-grade students who took the test in 2011. Only eighth-graders' responses were used because the new version of the exam is only available for the 8^{th}-grade students. Among 27,952 students, 55% of them are male with a mean of raw scores of 8.82 and standard deviation of 6.10 and 45% of them are female with a mean of raw scores of 9.60 and standard deviation of 6.44. Because the data include a large sample size, we sampled it based on three different sizes: 500, 1,000, and 2,000. To be able to generalize results to the whole population, 100 replications were implemented. As a cut-off point, significant 50 results out of 100 replications were considered DIF or DDF. In addition, for the impact of sample sizes, effect size results were considered in the DIF analysis.

Statistical Procedures

Two methods were used for DIF analyses. The first method was the Mantel Haenszel (MH), which is a chi-square based technique. The second method was the Logistic Regression (LR), which is a regression model with observed test scores and group membership (Swaminathan & Rogers, 1990). Nagelkerke's R^2 (Nagelkerke, 1991), which is the proportion of explained variance in a model, can be taken as a measure of effect size. Two definitions of DDF have been discussed in the literature (Suh & Talley, 2015): DDF in a "divide-by-total" framework (Thissen & Steinberg, 1986) and DDF in a "divide-by-distractors" framework (Suh & Bolt, 2011). For these frameworks, DDF was analyzed using two nonparametric odds ratio approaches under the nested logit model (ORA-NLM; Terzi & Suh, 2015) for divide-by-distractors and under the nominal response model (ORA-NRM; Penfield, 2008) for divide-by-total. The reasons we chose these nonparametric approaches are because (1) it is easy for practitioners and teachers to calculate with available user-friendly software (e.g., SPSS, Excel), which could be obtained upon request; (2) these two approaches can help us understand whether DIF can have a considerable impact on DDF by including or isolating the correct option in the analyses; and (3) these approaches require a relatively smaller sample size compared to parametric approaches.

Mantel Haenszel (MH) for DIF Analysis

The Mantel Haenszel (MH; Mantel & Haenszel, 1959) is a chi-square based technique based on a contingency table developed by Holland and Thayer (1986) for DIF detection purposes. This contingency table shown in Table 1 includes two columns and two rows; columns present the number of correct and incorrect answers, rows present the focal (F) and reference (R) group sizes.

Table 1. Contingency Table by Groups for an Item

Groups	Correct	Incorrect	Total
Focal Group	A	B	A+B
Reference Group	C	D	C+D
Total	A+C	B+D	T

$$MH = \frac{AD}{BC} \quad (1)$$

$$\Delta_{MH} = -\frac{4}{1.7}\ln MH = -2.35 \ln MH \quad (2)$$

Equation 1 shows the odds ratio of DIF for a particular item, which ranges from 0 to ∞. The expected optimum result is 1. That is, if the MH value exceeds 1, the item favors focal group, and if the value is smaller than 1, item favors reference group. For more interpretable results, Equation 1 was modified into Equation 2 by taking log of the MH statistic (Holland & Thayer, 1986). Equation 2 identifies the level of DIF. If $|\Delta_{MH}| < 1$, DIF level is A (ignorable); if $1 < |\Delta_{MH}| < 1.5$, DIF level is B (medium); and if $|\Delta_{MH}| \geq 1.5$, DIF level is C (high) (Zieky, 1993). One of the disadvantages of the MH method is that it cannot distinguish between uniform and non-uniform DIF.

Logistic Regression (LR) for DIF Analysis

The Logistic Regression (LR) method was used for DIF detection purposes based on the Logistic Regression analysis (Swaminathan & Rogers, 1990). Briefly, the LR model is established by using group membership as a dummy independent variable. If the group variable has an important role for the model, the item is flagged as DIF. While responses to a particular item become the dependent variable, intercept, the total score (*X*), group variable (*g*), and an interaction term between total scores and group membership are independent variables for the LR model.

$$\ln\frac{P}{1-P} = \beta_0 + \beta_1 x + \beta_2 g + \beta_3 xg.$$

If $\beta_3$is significant for the model, the item shows non-uniform DIF. If it is not significant, β_3 is excluded from the model. Then, if β_2 is significant for the new model, the item shows uniform DIF, otherwise the item does not show any DIF. Likelihood ratio statistic was used in the study to compute DIF statistic.

Differences between explained variance (R^2) by models can be considered as an effect size. Gierl, Khaliq, and Boughton (1999) and Zumbo and Thomas (1997) proposed two different cut-off points for effect size measures. According to Table 2, Zumbo and Thomas's criteria are more conservative than Gierl, Khaliq and Boughton's criteria. If a calculated effect size is 0.1 for Gierl, Khaliq, and Boughton, it is large, but it is negligible for Zumbo and Thomas.

Table 2. Effect Size Criterion for LR DIF Detection

Level of DIF	*Gierl, Khaliq and Boughton*	*Zumbo and Thomas*	*Meanings*
A	$R^2<0.035$	$R^2<0.13$	*Negligible effect*
B	$0.035<R^2<0.07$	$0.13<R^2<0.26$	*Moderate effect*
C	$0.07<R^2$	$0.26<R^2$	*Large effect*

2PL-NLM and 2PL-NRM

A two-parameter logistic-nested logit model (2PL-NLM; Suh & Bolt, 2010) was designed as an alternative to the nominal response model (NRM; Bock, 1972). The probability of an examinee at ability θ_j choosing the correct response under the 2PL-NLM can be modeled as a 2PL model:

$$P(u_{ij}=1|\theta_j)=\left[\frac{\exp(\beta_i+\alpha_i\theta_j)}{1+\exp(\beta_i+\alpha_i\theta_j)}\right],$$

where β_i is an intercept parameter and α_i is a slope parameter for item *i*.

Given an incorrect response, the probability of choosing a distractor *v* as the product of the probability of an incorrect response and the probability of selecting distractor *v* can be modeled as follows:

$$P(u_{ij}=1,d_{ijv}=1|\theta_j)=P(u_{ij}=0|\theta_j)P(d_{ijv}=1|u_{ij}=0,\theta_j)=$$

$$=\left[1-\frac{\exp(\beta_i+\alpha_i\theta_j)}{1+\exp(\beta_i+\alpha_i\theta_j)}\right]\left[\frac{\exp(z_{iv}(\theta_j))}{\sum_{k=1}^{m-1}\exp(z_{ik}(\theta_j))}\right], \quad (3)$$

where $z_{iv}\left(\theta_j\right)=\zeta_{iv}+\lambda_{iv}\left(\theta_j\right)$, which is multinomial logit for the propensity for each distractor; u_{ij} is the item response for item *i* by examinee *j*, where if an examinee *j* answers item *i* correctly, u_{ij} = 1, and 0 otherwise; and d_{ijv} represents an item by an examinee by a distractor array such that if an examinee *j* selects a distractor *v* (*v* = 1, 2, …, *m*–1) of item *i*, d_{ijv} = 1, and 0 otherwise (Suh & Bolt, 2010).

The NRM (Bock, 1972) has the same form as the second bracket term in Equation 3. An arbitrary linear restriction is imposed on distractor parameters as $\sum_{v=1}^{m-1}\lambda_{iv}$ and $\sum_{v=1}^{m-1}\zeta_{iv}$ in the NLM, whereas, this constraint is applied to all response categories, including the correct response in the NRM.

Odds Ratio Approach under the NRM (ORA-NRM) for DDF Analysis

Following the notion in Bock (1972) under the NRM, Penfield (2008) proposed an odds-ratio approach (ORA) based on the MH common odds ratio estimator (Mantel & Haenszel, 1959). In this approach, ability is divided into *k* ability levels, which are based on total raw scores. The conditional odds ratio for distractor *v* across *k* ability levels is shown as follows:

$$\hat{\alpha}_v=\frac{\sum_{k=1}^{K}R_{0k}F_{vk}/T_k}{\sum_{k=1}^{K}R_{vk}F_{0k}/T_k},$$

where R_{0k} is the number of examinees in the reference (R) group at the k^{th} ability level who have responded to the item correctly; R_{vk} is the number of examinees in the R group who have chosen distractor *v*; F_{0k} and F_{vk} represent the counterparts in the focal (F) group; and the summation of those numbers is represented by T_k (Mantel & Haenszel, 1959):

$$T_k = R_{0k} + R_{vk} + F_{0k} + F_{vk}.$$

The natural logarithm of $\hat{\alpha}_v$ is,

$$\hat{\lambda}_v = \ln(\hat{\alpha}_v). \quad (4)$$

Then, DDF in distractor *v* can be obtained by dividing $\hat{\lambda}_v$ by its standard error (SE). The SE of $\hat{\lambda}_v$ is as follows:

$$SE(\hat{\lambda}_v) = \sqrt{\frac{\sum_{k=1}^{K} T_k^{-2}(R_{0k}F_{vk} + \hat{\alpha}_v R_{vk}F_{0k})(R_{0k} + F_{vk} + \hat{\alpha}_v R_{vk} + \hat{\alpha}_v F_{0k})}{2\left(\sum_{k=1}^{K} \frac{R_{0k}F_{vk}}{T_k}\right)^2}}. \quad (5)$$

$z(\hat{\lambda}_v) = \hat{\lambda}_v / SE(\hat{\lambda}_v)$, which is calculated based on Equations 4 and 5, is approximately distributed as the standard normal (Hauck, 1979).

Odds Ratio Approach under the NLM (ORA-NLM) for DDF Analysis

The 2PL-NLM evaluates items with DDF independent of DIF because it separates distractor parameters from correct response parameters in multiple-choice (Terzi & Suh, 2015). Suh and Bolt (2011) made a noticeable investigation of whether distractors can be considered responsible for DIF. Briefly, in contrast to Penfield (2008)'s approach, the correct response is excluded in evaluating distractor *v* under the ORA-NLM. For more details, refer to Penfield (2008) for the ORA-NRM, and Terzi and Suh (2015) for ORA-NLM.

RESULTS

The number of DIF detection results based on gender was reported in Table 3. When *N* = 500 and 1,000, none of the items showed DIF more than 50% out of 100 replications under both MH and LR approaches. Increasing the sample size caused increases in the percentage of DIF items for almost all items as expected. The MH and LR analyses showed similar results, as reported in other studies (Doğan & Öğretmen, 2008; Gierl, Khaliq, & Boughton 1999; Mazor, Kanjee, & Clauser, 1995). The MH cannot distinguish non-uniform DIF, but the LR can, researchers used both uniform and non-uniform DIF detection, so the LR can detect more DIF than the MH (Rogers & Swaminathan, 1993; Hidalgo & LóPez-Pina, 2004). Similarly, the LR approach detected DIF in items 1, 8, and 18, which were not detected by the MH when *N*=2,000. These items indicated non-uniform DIF, so the LR was able to detect these DIF items.

When *N* = 2,000, items 1 and 8 showed DIF only under the LR method. Items 13 and 17 displayed DIF in both analyses; whereas, item 5 showed DIF only under the MH. However, DIF levels of all items had a negligible effect based on the LR, even for Gierl, Khaliq, and Boughton's liberal criteria. For the MH, item 5 was advantaged for male, which showed DIF 53 times in 100 replications, and 11 of 53 DIF detections had moderate effect sizes. Item 13 was advantaged for female, which showed DIF 74 times in 100 replications, and 13 of 74 DIF detections had moderate, two of 74 had large effect sizes. Moreover, item 17 was advantaged for male, which showed DIF 73 times in 100 replications, and 20 of 73 DIF detections had moderate effect sizes.

Table 3. Number of DIF Detection

Item	*N* = 500		*N* = 1,000		*N* = 2,000	
	LR	MH	LR	MH	LR	MH
1	22	5	37	13	64*	15
2	11	3	10	9	17	8
3	4	3	6	4	3	5
4	13	15	20	23	46	47
5	12	10	27	26	47	53*
6	13	7	17	12	19	13
7	2	3	8	5	3	3
8	13	3	31	9	62*	9
9	4	6	8	4	8	13
10	11	11	6	5	10	7
11	2	2	6	5	6	7
12	4	3	9	7	8	6
13	23	18	39	45	77*	74*
14	3	4	6	3	13	5
15	8	6	8	7	15	14
16	2	2	1	4	4	6
17	19	16	30	39	63*	73*
18	11	3	16	11	47	21
19	4	2	8	3	7	3
20	4	4	12	3	19	9

Note. * indicates percentage of observing significant DIF based on 50% cut-off criterion

Among responses to 20 items, in addition to five DIF items, item 4 showed DIF close to the cut off, 46% and 47%, using the LR and MH, respectively. Thus, six DIF items were further investigated for DDF. The other 14 DIF-free items were considered as anchor items where 15 ability levels were obtained, ranging from zero to 14. For convenience, the fourth option among the four response options was set as the correct response. The significance level was set to .05 with the Bonferroni correction based on the number of response options considered.

Results are reported as the average of significant DDF detections across 100 sampled data sets for the each of the three-sample sizes, as shown in Table 4. DDF was not observed when *N* = 500 under the two approaches. However, the distractor *c* showed 50% DDF for item 13 when *N* = 1,000 under the ORA-NRM. When *N* = 2,000, the distractor *b* in item 1 showed 54% DDF under the ORA-NLM, whereas, it was 53% under the ORA-NRM. As expected with a larger sample size, the distractor *c* showed 88% DDF for item 13, in addition to the distractor *b*, which displayed 72% DDF for item 17 under the ORA-NRM.

For reference, the items with DDF can be seen in appendices. Note that these distractors remained as is while the recoding was implemented because the correct responses were *a*, *a*, and *c* for items 1, 13, and 17, respectively, which were recoded into *d*. Then, the *d* option for these items was recoded into *a*, *a*, and *c*, respectively. The purpose of DDF investigation along with DIF was to check whether DIF in the correct response could be explained through DDF. The findings suggest that the distractors *c* and *b* in items 13 and 17, respectively, can be a considerable reason for DIF in the correct response due to the nature of the ORA-NRM. However, the distractor *b* in item 1 can be

ignored because not only it was detected with a large sample size but also DIF in item 1 had a negligible effect size.

Table 4. Number of DDF Detection

		ORA-NLM			ORA-NRM		
N	Items	a	b	c	a	b	c
500	1	5	16	5	5	17	0
	4	1	1	3	1	2	3
	5	1	3	3	4	2	7
	8	5	3	1	1	1	2
	13	1	8	6	6	2	20
	17	3	4	1	3	7	5
1,000	1	9	26	6	3	27	3
	4	2	6	3	4	6	10
	5	1	5	8	4	5	16
	8	10	2	6	2	5	10
	13	3	17	9	5	1	50*
	17	4	15	4	3	36	2
2,000	1	8	54*	18	7	53*	2
	4	3	2	2	10	1	23
	5	0	4	8	12	10	33
	8	15	2	10	1	5	10
	13	1	40	29	9	3	88*
	17	12	36	3	4	72*	16

Note. * indicates percentage of observing significant DDF based on 50% cut-off criterion.

SUMMARY and DISCUSSION

This study investigated one of the nationwide high-stakes tests (SBS) in Turkey, which is a previous version of a recent high school entrance exam (TEOG). The high school entrance exam was one of the most important high-stakes tests in Turkey taken by over one million students. For valid and fair results, these types of tests must be prepared and applied carefully. To evaluate this exam thoroughly in terms of DIF and DDF, results were obtained based on 100 replications for each condition with small sample sizes because single analyses with large sample sizes could give misleading results. Increase in sample sizes caused an increase in significant DIF rates. Various sample sizes are included as a condition similar to other DIF studies (Mazor, Clauser, & Hambleton, 1992; Scott et al, 2009; Zwick, 2012). If none of the items is flagged as DIF in small sample sizes, we cannot ensure that none of the items has DIF. Similarly, if most of the items are flagged as DIF in large sample sizes, it does not guarantee that those items have DIF. Therefore, to overcome this conflict, we used moderate sample sizes and reported effect sizes, as used in many studies.

When N = 500 and 1,000, none of the items showed DIF more than 50%. Because they were lower than the cut-off point, we accept that none of the items showed DIF, and therefore, we did not investigate effect sizes. When N = 2,000, the LR method detected more DIF items than the MH method because the MH method was not originally proposed for non-uniform DIF detection (Rogers & Swaminathan, 1993). But, all DIF detections had negligible effect sizes based on the LR, even for a liberal criterion (Gierl, Khaliq, & Boughton, 1999). The MH approach detected three DIF items, which were more than 50% over the replications. Because some of these DIF detections had moderate or large effect sizes, items 5, 13, and 17 are candidate for a possible bias investigation. While items 5 and 17 favored male, item 13 favored female. For reference, these items were demonstrated in appendices. If the approaches could detect DIF in small sample sizes, it would strengthen suspicion. Because sample size effects in DIF detection or differences between the approaches were not the main purpose of this research, we did not confront them.

While there are several DIF studies on SBS math exam with the MH approach, some of these studies also investigated item bias. In terms of DIF detection, Kelecioğlu, Karabay, and Karabay (2014) detected one math item with a moderate effect size in the 8th grade SBS exam administered in 2009. Another study used the same test detected another DIF item with a moderate effect (Arıkan, Uğurlu, & Atar, 2016). Results are not the same due to using different samples. Our dataset was used by Karakaya (2012), with N = 9,374, two items (2 and 6) had DIF with moderate effect sizes, none of them had DIF in our results. Another research, which used the same dataset with N = 121,137, detected items 5 and 7 as DIF with moderate effect sizes (Kan, Sünbül, & Ömür, 2013). Item 5 had DIF in our result, but not item 7. These results show that DIF detection can vary for different samples, and even for the same samples with different sample sizes. Therefore, by using different sample sizes and the number of replications, DIF information can be enriched.

Up until now, DDF analyses have not been investigated for the SBS exam. Therefore, this study aimed to carry out additional DDF analyses for the exam. Even though the choice of a distractor does not affect test scores, DDF analyses can give information about different subgroups (Abedi, Leon, & Kao, 2007). It is important to examine DIF and DDF together, because a significant interaction between gender and a choice of distractor was found (Marshall, 1983). It is also crucial that DIF and DDF analyses can be generally applied for different goals such as refining existing test items, developing new scales, and validating test score inferences (Zumbo, 2007).

Having said that, in this study DDF was investigated along with DIF to understand whether DDF can cause DIF in correct responses. Using 14 DIF-free items as anchor items, six items (i.e., 1, 4, 5, 8, 13, and 17) were further examined for DDF. When N = 500, the two approaches did not detect any DDF. However, when N = 1,000 under the ORA-NRM, the distractor *c* in item 13 showed DDF. Increasing the sample size to 2,000, the distractor *b* in item 1 showed DDF under both ORA-NLM and ORA-NRM. Again, when N = 2,000, DDF was detected for the distractor *c* in item 13 and *b* in item 17 under the ORA-NRM. Specifically, we could state that the distractor *c* in item 13 can be a serious cause for DIF in the correct response because it was detected twice with two sample sizes under the ORA-NRM. This distractor favors male. The distractor *b* in item 17 can also be considered a reason for DIF, which favors female. However, the distractor *b* in item 1 requires a special investigation because DIF in this item had a negligible effect size. Therefore, this distractor may or may not cause DIF because it was detected by both approaches, where the ORA-NLM isolated the distractor *b* from the correct response. In summary, based on the findings, we could conclude that those DIF items detected with DDF need to be revised and evaluated with caution for valid test inferences. Note that similar to effect size results in DIF analyses, large sample sizes showed DDF more than small sample sizes.

With high-stakes tests continuing to be a central part of the educational systems, the analyses of DIF and DDF thus remain important in obtaining measurement invariance for test items across different subgroups at the same ability level. We further suggest that DIF and DDF analyses of the remaining parts of the SBS and/or TEOG such as science and social studies should be investigated in order to obtain valid test results. There is also a limitation of this study that those items detected with DIF and DDF should be investigated in the context of gender, test bias, and mathematics. A deeper discussion of why these items exhibit DIF and DDF should be a topic of future studies from the perspective of mathematics educators.

REFERENCES

Abedi, J., Leon, S., & Kao, J. (2007). *Examining differential distractor functioning in reading assessments for students with disabilities*. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

Arikan, Ç. A., Uğurlu, S., & Atar, B. (2016) Mimic, Sibtest, Lojistik Regresyon ve Mantel-Haenszel yöntemleriyle gerçekleştirilen DMF ve yanlılık çalışması. *Hacettepe Eğitim Fakültesi Dergisi*, *31*: 34-52. doi:10.16986/huje.2015014226

Camilli, G. (2006). Test fairness. *Educational Measurement, 4*, 221-256.

DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: How accurately can we detect who is responding differentially? *Educational and Psychological Measurement, 71*, 597–616. doi.org/10.1177/0013164411404221

Doğan, N., & Öğretmen, T. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *33*(148), 100-112.

Doornik, J. A. (2002). *Object-Oriented matrix programming using ox* (3rd ed.). London: Timberlake Consultants Press and Oxford: www.nuff.ox.ac.uk/Users/Doornik.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Gierl, M., Khaliq, S. N., & Bougthon, K. (1999, June). *Gender differential item functioning in mathematics and science: Prevalence and policy implications*. Paper presented at the symposium entitled "Improving large–scale assessment in education" at the Annual Meeting of the Canadian Society for the Study of Education, Canada.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26*, 147–160.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*, 17–27.

Hidalgo, M. D., & LÓPez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure*. ETS Research Report No. 86-31. Princeton, NJ.

Kan, A., Sünbül, Ö., & Ömür, S. (2013). 6.-8. Sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *9*, 207-222.

Karakaya, İ. (2012). Seviye belirleme sınavındaki fen ve teknoloji ile matematik alt testlerinin madde yanlılığı açısından incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 12*, 222-229.

Kelecioğlu, H., Karabay, B., & Karabay, E. (2014). Seviye belirleme sınavı'nın madde yanlılığı açısından incelenmesi. *İlköğretim Online*, *13*, 934-953.

Nagelkerke, N. J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika*, *78*, 691-692.

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*, 847-862.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Marshall, S. P. (1983). Sex differences in mathematical errors: An analysis of distractor choices. *Journal for Research in Mathematics Educations*, *14*, 325–336.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, *52*, 443-451.

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*, 131–144.

Ministry of National Education. (2011). *2011 yılı 8. sınıflar seviye belirleme sınavı sayısal bilgiler*. Retrieved from http://www.meb.gov.tr/duyurular/duyurular2011/EGITEK/sbs2011BasinBulteni/01_2011SBS_8SayisalBilgiler.pdf

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, *45*, 247–269.

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

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., & EORTC Quality of Life Group. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, *62*, 288-295.

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement, 48*, 188–205. doi.org/10.1111/j.1745-3984.2011.00139.x

Suh, Y., & Talley, A. E. (2015). An empirical comparison of DDF detection methods for understanding the causes of DIF in multiple-choice items. *Applied Measurement in Education, 28*, 48–67. doi.org/10.1080/08957347.2014.973560

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

Terzi, R., & Suh, Y. (2015). An odds ratio approach for detecting DDF under the nested logit modeling framework. *Journal of Educational Measurement, 52*, 376–398. doi.org/10.1111/jedm.12091

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

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.* Ottawa: National Defense Headquarters.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*, 223-233.

GENİŞ ÖZET

Giriş

Bu araştırma Türkiye'de gerçekleştirilen geniş ölçekli bir sınavın değişen madde fonksiyonu (DMF) ve değişen çeldirici fonksiyonu (DÇF) açısından incelemeyi amaçlamaktadır. Bu amaçla 2011 yılı Seviye Belirleme Sınavı (SBS) analize tabi tutulmuştur. DMF ve DÇF çalışmaları adil sınav imkanının tüm gruplara eşit olarak sunulup sunulmadığını araştırmasından dolayı çok büyük bir önem arz etmektedir.

SBS, Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavının (OKS) yerine getirilen ilköğretimden orta öğretime geçişte öğrencileri seçme amacı taşıyan bir sınavdı. 2013-2014 eğitim öğretim yılında yerini Temel Eğitimden Ortaöğretime Geçiş (TEOG) sınavına bırakmıştır. Bu sınavların her birine yaklaşık 1 milyon öğrenci katılmaktadır. Sınava katılan öğrenci sayısı ve sınavın amacı sınavın önemini ortaya koymaktadır.

Objektif bir test geliştirmenin en önemli aşamalarından biri de kaliteli madde yazımıdır. Yazılan maddeler sınava katılan tüm alt grup bireyler için aynı işleve sahip olmalıdır. Bunun içinde soru yazımında azami gayret gösterilmelidir. Ancak yine de göz önünde bulundurulmayan bazı faktörler maddeleri alt gruplar için farklı anlam taşır hale getirebilmektedir. DMF araştırmaları ile maddelerin farklı alt gruptaki eşit yeteneğe sahip bireyler için aynı zorluğa sahip olup olmadığı sorgulanmaktadır.

SBS örneğinde olduğu gibi ülkemizde ve dünyanın pek çok yerinde kullanılan geniş katılımlı sınavlarda özellikle objektif değerlendirme olanağından faydalanmak adına çoktan seçmeli sorular kullanılmaktadır. Kolay ve adil değerlendirme çoktan seçmeli testlerde avantaj olarak ele alınırken madde hazırlamanın zorluğu dezavantaj olarak karşımıza çıkmaktadır. Madde içerik ve köklerinde bulunan kimi hatalar DMF'ye neden olabilmektedir. Çoktan seçmeli testlerde madde yazımı kaliteli ve objektif çeldirici yazımını da gerektirmektedir. Farklı gruplar için farklı anlam içeren çeldirici seçenekler, adil sınav ilkesini ihlal etmekte ve DÇF'ye neden olabilmektedir. DÇF'ler ise DMF sebep olabilmektedir.

DÇF araştırmaları DMF araştırmalarına göre sayıca daha azdır. Türkiye'de gerçekleştirilen bir sınav ile ilgili DÇF çalışması bulunmamaktadır. Araştırmanın Türkiye'de gerçekleştirilen bir sınav için ilk olması ve bu sınavın ismen sona ermiş olmasına rağmen bazı yapısal değişikliklerle devam eden geniş katılımlı önemli bir sınav olması araştırmanın önemini bir kat daha artırmaktadır.

Yöntem

Araştırmanın çalışma grubunu 2011 SBS'ye giren bir milyondan fazla öğrenciden, matematik testinin tüm sorularını yanıtlayan rasgele seçilmiş 27.952 öğrenci oluşturmaktadır. B kitapçığını alan öğrencilerin yanıtları, A kitapçığındaki sıraya göre düzenlenmiştir. DMF analizi için veriler cinsiyete göre 0 ve 1'e dönüştürülmüştür. Analizler cinsiyete göre farklılaşmayı araştırmak üzere gerçekleştirilmiştir.

DMF analizi için çok kullanılan yaklaşımlardan olan Mantel-Haenzsel (MH; Holland ve Thayer, 1986) ve Lojistik Regresyon (LR; Swaminathan ve Rogers, 1990) teknikleri seçilmiştir. DMF analizinde farklı yöntemler farklı sonuçlar sunabilmektedir. Bu nedenle eksik sonuçlar üretebilecek tek yönteme bağlı kalınmamış ve biri birini destekleyerek eksik yönlerini tamamlaması beklentisiyle iki yöntem kullanılmıştır.

DÇF analizleri de iki farklı yöntemle gerçekleştirilmiştir; iki parametreli lojistik iç-içe logit model (2PL-NLM) ve nominal yanıt modelinin (NRM) altındaki iki olasılık oranı yaklaşımı (ORAs) kullanılmıştır. DÇF'nin iki tanımı, "çeldiricilere bölünme" (Suh ve Bolt, 2011) çerçevesinde DÇF olmak üzere karşı "toplama bölünme" (Thissen ve Steinberg, 1986) çerçevesinde DÇP olmak üzere literatürde sunulmuştur (Suh ve Talley, 2015). Bu çerçevede DÇF, "toplama bölünme" için nominal yanıt modeli (ORA-NRM, Penfield, 2008) ve "çeldiricilere bölünme" için iç-içe geçmiş logit modeli (ORA-NLM; Terzi ve Suh, 2015) olmak üzere parametrik olmayan iki olasılık oranı yaklaşımları kullanılarak analiz edildi. Bu parametrik olmayan yaklaşımları seçmemizin sebepleri şunlardır: Birincisi, uygulayıcılar ve öğretmenler, mevcut kullanıcı dostu yazılım (örneğin, SPSS, Excel) ile hesaplamalarını kolayca yapabilirler; bunlar, talep üzerine elde edilebilir; İkincisi, bu iki yaklaşım, analizlerde doğru seçeneği ekleyerek veya izole ederek DMF'nin DÇF üzerinde önemli bir etkisi olup olmadığını anlamamıza yardımcı olabilir. Üçüncü olarak, bu yaklaşımlar parametrik yaklaşımlara kıyasla nispeten daha küçük bir örneklem boyutu gerektirir. Analizlerimizde kolaylık sağlaması için, doğru cevaplar D seçeneği olarak düzenlenmiştir; eğer sorunun doğru yanıtı D seçeneği ise, doğru yanıt olan D seçeneğinin yeri değiştirilmemiştir.

DMF analizlerinde etkin olan en önemli faktörlerden biri de örneklem büyüklüğüdür. Bu veri setinin tamamen kullanılması DMF gösteren madde sayısında çok büyük bir artışı getirebileceği göz önünde tutularak, 500, 1,000 ve 2,000 bireylik veri setleri üzerinden analiz yapılmasına karar verilmiştir. Tek bir örneklem yerine veri setinden 100 farklı örneklem seçerek, 100 tekrarla analiz gerçekleştirilmiştir. Böylece sonuçların daha genele hitap etmesi sağlanmıştır. Tekrarların en az %50'sinde DMF'nin anlamlı sonuç olarak görülmesi durumunda madde DMF'li olarak ele alınmıştır. DMF'li olarak görülen maddeler DÇF analizine tabi tutulmuştur. Böylece DMF'li maddelerin kaynağının DÇF olup olmadığı araştırılmıştır.

DMF analizleri R (R Core Team, 2015) yazılım uygulamasındaki difR (Magis, Beland, Tuerlinckx ve De Boeck, 2010) paketi aracığıyla gerçekleştirilmiştir. DÇF analizleri ise Ox (Doornik, 2002) yazılım programında yazılan kodlarla gerçekleştirilmiştir.

Bulgular ve Tartışma

Yapılan DMF analizleri sonucu 500 ve 1,000 bireylik örneklemlerde MH ve LR analizlerinin her ikisi de hiçbir maddede, tekrarların en az yarısında anlamlı sonuç üretmemiştir. Bu durum, bu örneklem büyüklükleri için DMF'ye rastlanılmadığı yönünde yorumlanmıştır. 2,000 bireylik örneklem büyülüğünde ise tekrarların en az yarısında, LR ve MH yöntemleri 13. ve 17. maddeleri DMF'li, LR yöntemi 1. ve 8. maddeleri, MH yöntemi ise 5. maddeyi DMF'li bulmuştur. Ancak tekrarların hiçbirinde, LR tarafından DMF'li olarak işaretlenen dört maddenin ihmal edilebilir düzeyin üstünde etki büyüklüğüne sahip olmadığı görülmüştür. MH tarafından tespit edilen DMF'lerde ise bazı tekrarlarda orta ve üst düzey DMF görülmüştür.

DÇF analizleri ise potansiyel DMF taşıyan yukarıda belirtilen beş madde, ve MH ve LR analizlerinde sınır olan 50'ye çok yakın DMF gösteren 4. madde için gerçekleştirilmiştir. Sonuçlara

göre ORA-NLM yöntemi 2,000 örneklem büyüklüğünde 1. maddenin C çeldiricisinde DÇF tespit etmiştir. ORA-NRM yöntemi ise 13. maddenin C çeldiricisini 1,000 ve 2,000 örneklem büyüklüklerinde, ve 1. ve 17. maddelerinin B çeldiricisini 2,000 bireylik örneklem büyüklüğünde DÇF'li olarak işaretlemiştir.

Yapılan analizler sonucu DMF tespit edilen maddelerin üçünde DÇF tespit edilmiştir. Geniş katılımlı önemli ve ciddi sınavda 20 maddenin 5'inin analizlerce farklı fonksiyona sahip olduğunun ortaya konması önemli bir bulgudur. Test hazırlayan ve uygulayanların DMF'ye yol açabilecek şüpheli ifadelerden kaçınmasına dikkat edilmelidir. DMF'nin kaynaklarından biri olan DÇF'nin farklı fonksiyona sahip maddelere neden olabileceği görülmektedir. Sınav uzmanlarının çeldirici hazırlamada da daha dikkatli olması gerektiği araştırma sonucunda görülmektedir.

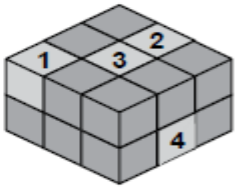
Bu çalışma diğer test alanlarında ve/veya diğer geniş ölçekli testlerde gerçekleştirilebilir. Yanlılık kaynağı araştırması için uzmanlar eşliğinde değerlendirmesi gelecek çalışmalar için önerilmektedir.

Appendices

1. $(-3)^{-2}$ sayısı aşağıdaki sayılardan hangisi ile çarpılırsa sonuç 3 olur?

A) 3^3 B) 3^{-1}
C) 3^2 D) $(-3)^{-3}$

13. Birim küplerden oluşan yandaki yapıda, numaralandırılmış küplerden hangisi çıkarıldığında yapının yüzey alanı <u>değişmez</u>?



A) 1 B) 2 C) 3 D) 4

17.

1. adım 2. adım 3. adım 4. adım

Yukarıda verilen örüntü, aynı kurala göre devam ettirildiğinde 19. adımdaki çemberin içine çizilen çokgenin kenar sayısı kaçtır?

A) 24 B) 33 C) 39 D) 42