Journal of Theoretical Educational Science, 13(3), 530-547, July 2020
Kuramsal Eğitimbilim Dergisi, 13(3), 530-547, Temmuz 2020
[Online]: http://dergipark.gov.tr/akukeg
DOI number: http://dx.doi.org/10.30831/akukeg.621581

The Effects of Distractors to Differential Item Functioning in Peabody Picture Vocabulary Test

Peabody Resim Kelime Testinde Çeldiricilerin Değişen Madde Fonksiyonuna Etkisi

Fatma Betül KURNAZ-ADIBATMAZ* **Hüseyin YILDIZ****

Received: 18 September 2019 **Research Article** **Accepted:** 19 May 2020

ABSTRACT: In this study logistic regression and Lord's Chi Square methods were used to research the items that have DIF. The study utilized Peabody Picture Vocabulary Test (PPVT). The original form of the PPVT includes four options. Three different forms (A, B and C) were formed by removing one of the distractors respectively. The original form of PPVT was implemented in a group of 970 preschool children who were aged between 3 to 6. 757 of them took one of the forms. In each implementation, the order to the implementation of the original form and the form (A, B or C) was changed. The applications were conducted 15 days apart. In the first application, the original form was applied, while one of the devised forms (A, B or C) was used in the following application. In this way, the effect of order of application on responses was investigated. The gender variable constituted the reference and focus group of the study. The Logistic Regression and Lord's Chi-square methods did not give compatible results in DIF analysis. DIF was found in 15 items in the original form according to the logistic regression method and in nine items according to the Lord's Chi-square method. The three-option and four-option applications of the test revealed that DIF was determined in five items in different forms. It was observed that there was no compliance in different applications and analyses in other items with DIF.

Keywords: Differential item functioning, logistic regression, Lord's chi square, Peabody picture vocabulary test.

ÖZ: Bu araştırmada değişen madde fonksiyonunun belirlenmesinde lojistik regresyon ve Lord'un Ki-kare yöntemleri karşılaştırılmıştır. Araştırmada Peabody Resim Kelime Testi (PRKT) kullanılmıştır. PRKT dört seçenekli maddelerden oluşmaktadır. Çeldiricilerin uygulamadaki etkisini görmek amacıyla farklı formlarda farklı bir çeldirici maddeden çıkarılarak üç seçenekli formlar oluşturulmuştur. PRKT 3-6 yaş arasında 970 çocuğa uygulanmış 757 uygulamadan elde edilen yanıtlar çözümlenmiştir. Uygulamalar 15 gün arayla gerçekleştirildi. Bir uygulamada önce original form uygulandı, diğer uygulamada oluşturulan formlardan biri (A, B veya C) uygulandı. Bu yolla yanıtlarda uygulama sırasının etkisi kontrol edildi. Cinsiyet değişkeni araştırmanın referans ve odak grubunu oluşturmuştur. DIF analizinde Lojistik Regreyon ve Lord'un Ki-kare yöntemi uyumlu sonuçlar vermedi. Araştırma bulgularına göre lojistik regresyon yöntemine göre orijinal formda 15, Lord'un Ki-kare yöntemine göre 9 maddede DIF belirlendi. Testin üç seçenekli ve dört seçenekli uygulamalarından elde edilen sonuçlarda farklı formlarda beş maddede uyumlu bir biçimde DIF belirlenmiştir. DIF belirlenen diğer maddelerde ise farklı uygulama ve analizlerde uyum olmadığı gözlenmiştir.

Anahtar kelimeler: Değişen madde fonksiyonu, lojistik regresyon, Lord'un ki-karesi, Peabody resim kelime testi.

* *Corresponding Author*: Asst. Prof. Dr., Karabuk University, Karabuk, Turkey, betulkurnaz@karabuk.edu.tr, https://orcid.org/0000-0002-7042-2159
** Expert, Bolu Assessment and Evaluation Center, Bolu, Turkey, huseyinyildiz35@gmail.com, https://orcid.org/0000-0003-2387-263X

Citation Information
Kurnaz-Adıbatmaz, F. B., & Yıldız, H. (2020). The effects of distractors to differential item functioning in Peabody picture vocabulary test. *Kuramsal Eğitimbilim Dergisi [Journal of Theoretical Educational Science], 13*(3), 530-547.

ISSN: 1308-1659

The cognitive or psychological construct to be measured with a test should be measured free of undesirable variables. This is related to the validity of the measuring tool. Measuring constructs other than the construct intended to be measured is a validity problem. There are many variables that can affect the validity of an assessment instrument. One of these variables is whether an instrument has bias that can change sample's possibility of giving the correct answer. The instruments that are used in developmental evaluations should be reported properly about their psychometric details considering the difference about culture and language (Alordiah & Agbajor, 2014; Washington, Kamhi, Pollock, & Harris, 1996). These details include precious findings about the validity of the instruments that are used.

In norm-referenced tests measuring cognitive tasks, item content can be an important validity problem. There are studies on how the differences in the ranking of the items (e.g. Hambleton & Traub, 1974) or the content of items lead to a difference in the total score (e.g. Zwick, 1991). Taking cultural and language differences in sub-groups into consideration in evaluation may prevent bias (Van de Vijver, 2018). Bias can stem from the structure of a test as well as from the items in that test. Item bias is the change in the possibility of responding correctly in one of the two groups with the same ability level (Osterlind, 1983).

Lord (1980) argues that when individuals with the same ability level have the same probability of answering an item correctly, then the test is fair. Holland and Wainer (1993) maintain that individuals with the same ability level should have equal chance of answering the item correctly, regardless of the group they are in.

In identifying item bias, whether there is differential item functioning (DIF) in the item or not may be researched. If there is a possibility of responding differently in one of the two groups that are at the same ability level, then, there is differential item functioning for that item (Gierl, Khaliq, & Boughton, 1999; Maller, 2001; Stump, Monahan, & McHorney, 2005). In cases where an item has bias, there is differential item functioning; however, the fact that there is differential item functioning in an item is not a decisive evidence for the existence of bias (Zumbo, 1999).

If a test contains DIF in one or more items, differential test functioning may occur. This seems to be more important than the presence of DIF in the item since items containing DIF are also used to obtain the total score (Chalmers, Counsell, & Flora, 2016). In the literature, there are studies whether items in a test had DIF or not (e.g. Adebule, 2013; Köse, 2015); there are also some studies that researched the effects of DIF identification techniques on item bias (e.g. Doğan & Öğretmen, 2010; Ikeda, 1995; Kalaycıoğlu & Kelecioğlu, 2001; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018; Yurdugül, 2003). In addition to these types of studies, some other studies focused on the comparison and contrast of DIF in contexts where different grading conditions were realized (e.g. Tunç & Kutlu, 2018). Thissen, Steinberg, and Fitzpatrick (1989) and Love (1997) stated that there is a relationship between choosing the incorrect option in a multiple-choice item and the level of ability. Sehmitt and Dorans (1990) argued that there is a relationship between choosing the incorrect option and ethnicity.

If an item includes DIF, it may be related to the difficulty level of the item. This difficulty may arise from the content of the item while it may also arise from the form (number of options, whether the items depend on the same shared root and so on) of the item. Ascalon, Meyers, Davis, and Smits (2007) applied items involving similar and

dissimilar distractors to 520 university students and conducted DIF analysis. They stated that factors such as the content and difficulty of distractors and their degree of proximity to the correct answer affect the difficulty level of items. They found that the probability of having DIF increased in the items involving distractors with a similar meaning.

Even if all the distractors in an item function equally, some subgroups may tend to choose particular distractors. Respondents may have little or no knowledge about some distractors (Banks, 2009). The time to respond to the item and the probability of answering the item correctly are related to the content of the distractors. Distractors may change the response behavior of subgroups (Meyer & Wise, 2006). Studies conducted on the structure and complexity of distracters revealed that distractors reduce test difficulty (Harasym, Leong, Violato, Brant, & Lorscheider, 1998), increase test difficulty (Hughes & Trimble, 1965), and do not affect test difficulty (Forsyth & Spratt, 1980). In an item devised for a test, the structure of the distractors as well as the correct response itself affects the psychometric property of the item (Suh & Talley, 2015).

The use of options in an item that can change the probability of a group responding correctly points to significant problems in terms of the psychometric properties of the test (Banks, 2009). There may be many external variables that can affect the ability of a distractor to function. Studies on distractors as causes of DIF in multiple choice items have been examined (e.g. Banks, 2009; Green, Crone, & Folk, 1989; Penfield, 2008, 2010; Suh & Talley, 2015; Terzi & Suh, 2015; Terzi & Yakar, 2018). In their study, Ascalon, Meyers, Davis, and Smits (2007) created distractors with similar and different content, and investigated whether DIF was found in the items. In the study, the items which aim to measure the same content and which involve similar and dissimilar distractors had significantly different Maentel Haenszel effect size.

Our study investigates with the DIF analysis whether the possibility of answering an item correctly changes in the subgroups when one of the distractors was removed from the item. This study is thought to be important as test developers may gain an insight into the function of distractors in a test. In the study, the main purpose of creating a new form by removing a distractor from the options for the same item and applying this form is to keep the options under control.

We used Logistic Regression and Lord's Chi Square methods to investigate the items that have DIF. One aim of this study is to compare logistic regression and Lord's Chi Square methods in determining DIF. Another aim is to determine whether the items with DIF are affected by the number of distractors.

Method

In this part of the study, the information regarding the sample, the data collection tool and data analysis were presented.

Participants

The original form of the PPVT which consisted of four options was applied to a total of 970 preschool children aged between 3 and 6, and 477 of them were boys while 493 of them were girls. All the children responded to the original form that has four options. Although the number of children who did the original four-choice form is 970, in the implementation of the three-choice forms, data loss is faced due to reasons such

as some children's absence, going onto holiday and so on. Consequently, the number of children responding to the three-choice forms is 757 in total. 242 of the students responded to the items in form A, and 254 and 261 students responded to the items in form B and form C, respectively. Students to receive the forms A, B and C were randomly selected. In addition, while the original form was applied to half of the sample randomly selected, the original form was applied to the other half after the application of forms A, B, C. In this way, the effects that may arise from the order of application of the forms were tried to be eliminated. The information regarding the gender of the children who did the original form was presented in Table 1.

Table 1

The Participant Information

City	Gender	N	%	City	Gender	N	%
Ankara	Girl	24	2.47	Samsun	Girl	25	2.58
	Boy	19	1.96		Boy	25	2.58
Amasya	Girl	25	2.58	Van	Girl	25	2.58
	Boy	25	2.58		Boy	25	2.58
Aydın	Girl	25	2.58	Zonguldak	Girl	25	2.58
	Boy	25	2.58		Boy	25	2.58
Bilecik	Girl	20	2.06	Konya	Girl	43	4.43
	Boy	24	2.47		Boy	47	4.85
Düzce	Girl	24	2.47	Mersin	Girl	25	2.58
	Boy	19	1.96		Boy	17	1.75
Elazığ	Girl	21	2.16	Osmaniye	Girl	25	2.58
	Boy	22	2.27		Boy	25	2.58
Gaziantep	Girl	25	2.58	Karabük	Girl	16	1.65
	Boy	25	2.58		Boy	19	1.96
Hatay	Girl	50	5.15	Kayseri	Girl	25	2.58
	Boy	40	4.12		Boy	25	2.58
İstanbul	Girl	24	2.47	Kocaeli	Girl	44	4.54
	Boy	26	2.68		Boy	46	4.74

The data belonging to the study sample were collected in regional destination of 18 city in Turkey. An equal number of girls and boys were ensured in implementation in each province. In this way, the goal was to have two equal groups in terms of gender. The number and percentage information of the children who answered the form was presented in Table 2.

Table 2

The Response Rates and Percentage Values for the Forms by the Children Who Made up the Sample

Forms	Gender	N	%	Total
Original	Girl	493	50.8	970
	Boy	477	49.2	
A	Girl	124	51.2	242
	Boy	118	48.8	
B	Girl	126	49.6	254
	Boy	128	50.4	
C	Girl	131	50.2	261
	Boy	130	49.8	

We ensure that that the number of girls and boys was equal in all applications. As a result, about half of the students were girls and boys.

The Data Collection Instrument

Peabody Picture Vocabulary Test (PPVT) was utilized to collect the data. PPVT which was developed by Dunn and Dunn (1959) was adapted to Turkish culture (Katz, Önen, Uzlukaya, Demir, & Uludağ, 1972). The adaptation study was realized in a sample group that consisted of 4200 children. The literature reported that it can be implemented on children between the ages of 2 and 12 (Özgüven, 1994). In the test, there are 100 questions that have options consisting of four different pictures. While responding to the items in the test, the child is asked to show the relevant picture or tell the number of the choice belonging to the picture. In the study that was undertaken to develop the test, the internal consistency was found to be between .71 and .81, and test re-test reliability was found to be between .52 and .82. The relationship of PPVT with Stanford-Binet Intelligence Scale was found to be between .82 and .86; while its relation with Wechsler Intelligence Scale for Children was found to be between .41 and .74 (Öner, 1997). Each correct answer in the test is worth 1 point and the sum of the correct responses a child provides makes up the raw score for that child (Temiz, 2002).

Due to existence of the studies (Kurnaz & Kelecioğlu, 2008; Washington & Craig, 1999) that showed that PPVT may be biased for sub-groups that have linguistic or cultural differences, it has the potential to include items with DIF and accordingly, in this study this assessment tool was chosen on purpose.

Procedures

In responding to the forms, if a child was firstly given the four-choice original form, the other child was given the three-choice form. In this way, the aim was to control the effect of the first implementation on the second one. The implementation was led by child development specialists. Before the implementations, the child development specialists were given an education on standard tests and the

implementation of the test. Ethical issues observed informed consent, voluntary participation, avoidance of plagiarism.

The aim of creating three-choice forms is to examine whether the results change in DIF analysis depending on the number of distractors. The chances of children choosing the correct answer may vary based on the distractor. It was decided to use three-option items, since chance factor increases when items have only two options, which would be an important limitation. If the presence of DIF in an item is due to the distractor itself, it will result in DIF in the item in at least two of the three forms and in the original form itself.

Data Analysis

First, descriptive data analyses were conducted as it was thought that descriptive analysis would provide information to interpret the obtained results. The scores were normally distributed in the subgroups according to the gender variable. In order to check whether the subgroups differed in total score in terms of gender variable, *t-test* was performed. The internal consistency of the data obtained from the applications was calculated with KR-20.

ANOVA, transformed item difficulty, Chi-square (χ^2), Item Characteristic Curve, Maentel-Haenszel, Logistic Regression, distractor response analysis methods can be used in identifying DIF (Gierl, Khaliq, & Boughton, 1999; Jensen, 1980; Osterlind, 1983). In this study, Logistic Regression and Lord's Chi-square methods were used to identify DIF.

Logistic Regression can be used for designating both uniform and also non-uniform DIF. It is a special regression model in which the dependent variable can have two values and the independent variable is a continuous variable (Gierl, Khaliq, & Boughton, 1999). The Logistic Regression model is analyzed using the $P(u=1)=\theta^2/1+\theta^2$ equation. Three sub-models are used to study DIF. These were presented below.

$$z=\beta_0+\beta_1X$$

$$z=\beta_0+\beta_1X+\beta_2G$$

$$z=\beta_0+\beta_1X+\beta_2G+\beta_2GX$$

Here X stands for the test score variable, G stands for group variable and GX stands for test score and group interaction variable. In the model when the variable X is significant, this shows that the model is valid; when the variable G is significant, this indicates a uniform differential item functioning and the significance of GX indicates a non-uniform differential item functioning (Yurdugül, 2003). In addition to the views suggesting that logistic regression analysis is affected by sample size (Tian, Pang & Boss, 1994), a consensus could not be achieved on how to identify and classify the items that have DIF (Hidalgo & Lopez-Pina, 2004; Jodoin & Gierl, 2001). In logistic regression model standardized regression coefficients (R^2) give the degree of the DIF and it is identified at three levels (Hidalgo & Lopez-Pina, 2004).

Lord (1980) proposed using the χ^2 method based on the item response model to determine uniform and non-uniform DMF (Maij-de Meij, Kelderman, & Van der Flier, 2010; Wiberg, 2007). This method is based on the comparison of item parameters in subgroups called reference and focus groups. With the help of the difference between

the item parameters calculated across subgroups and the variance-covariance matrix related to this difference, χ^2 statistics is calculated. In order to make comparisons between groups, the estimated parameters are brought to the same scale level. When the χ^2 statistical value exceeds the critical value, it is decided that the item includes DMF according to the relevant meaning level (Camilli, Shepard, & Shepard, 1994).

In DIF studies, the gender variable is widely investigated; thus, comparisons in this study were also made between different genders. In the DMF analysis of the data, the "*difR*" package in the R Studio 3.4.1 program was used. To detect DMF, the "*difLogistic*" function was used for DMF detection with the logistic regression method, and the "difLord" function was used in the Lord's chi square method. The difR package was written by David Magis et al in 2010.

Results

The study investigated whether the original form of the Peabody Picture Vocabulary Test and three different forms created by removing a distractor from the test contain DIF. Logistic regression and Lord's Chi-square methods were used to analyze the data. In Table 3, the descriptive test findings that were obtained from the original and the forms were presented.

Table 3

Descriptive Test Statistics according to Gender

Form	Group	*N*	*Min*	*Max*	$\bar{X}$	*S*	KR-20	*t*
Original	Girl	493	34	96	68.41	9.95	.84	-0.76*
	Boy	477	32	91	68.90	9.84		
A	Girl	124	35	93	72.30	9.06	.85	-0.85*
	Boy	118	35	92	73.39	10.69		
B	Girl	126	48	88	69.14	8.51	.79	-0.74*
	Boy	128	47	87	69.91	8.00		
C	Girl	131	41	93	71.70	8.82	.84	-0.37*
	Boy	130	35	92	71.75	10.45		

N: Frequencies, Min: Minimum Score, Max: Maximum score, $\bar{X}$:Arithmetic mean, *S*: Standard deviation, *p>.05

According to Table 3, the findings below may be deduced when the descriptive results obtained through various procedures were analyzed.

- The data obtained from the girls in the main application shows a broader range of score distribution. In the samples that were formed via random sampling, the range shrinks for both girls and also boys. In the group where form B was implemented, the range is narrower compared to all the other implementations.
- As scores showed normal distribution in each sub-group, whether there is a difference between the mean scores of the girls and boys was analyzed via t-

test and no significant difference was found between any of the groups (p>.05).

- The internal consistency of the data obtained from the implementations was calculated using KR-20 and it was found to vary between .79 and .85. Considering these findings, it can be argued that the data obtained through these implementations provide reliable results.

The logistic regression findings regarding the items with DIF were presented in Table 4.

Table 4

The Logistic Regression Analysis Results according to Gender Variable

Implementation	Item Number	$\chi 2$	p	DIF R^2	DIF Level
Original Form	11	6.190	0.045	0.018	A
	12	14.200	0.603	0.045	A
	18	6.924	0.031	0.013	A
	19	16.397	0.000	0.050	A
	20	7.313	0.025	0.027	A
	29	17.864	0.000	0.030	A
	32	19.452	0.000	0.024	A
	42	10.759	0.004	0.015	A
	44	9.977	0.006	0.043	A
	49	8.468	0.014	0.009	A
	51	12.519	0.001	0.017	A
	54	12.146	0.002	0.015	A
	65	11.636	0.003	0.013	A
	82	10.786	0.004	0.014	A
	96	9.507	0.008	0.017	A
FORM A	24	6.927	0.031	0.119	A
	25	7.493	0.023	0.034	A
	32	6.411	0.040	0.038	A
	34	7.326	0.025	0.157	B
	39	7.285	0.026	0.176	B
	40	6.883	0.032	0.038	A
	85	6.104	0.047	0.032	A
	93	7.366	0.025	0.041	A
FORM B	17	7.827	0.020	0.146	B
	19	7.883	0.019	0.148	B
	23	8.453	0.014	0.052	A
	37	10.436	0.005	0.139	B

	39	6.535	0.038	0.158	B
	59	10.374	0.005	0.051	A
	69	8.966	0.011	0.114	A
	76	6.943	0.031	0.031	A
	77	7.932	0.018	0.034	A
	88	11.611	0.003	0.065	A
FORM C	13	8.285	0.015	0.260	C
	22	10.369	0.005	0.128	A
	24	6.645	0.036	0.143	B
	29	9.015	0.011	0.051	A
	32	17.336	0.000	0.091	A
	38	7.205	0.027	0.058	A
	58	8.641	0.013	0.057	A
	78	9.970	0.006	0.051	A
	93	7.139	0.028	0.039	A

When the logistic regression results are analyzed, DIF was detected in 15 items in the original form (DIF level was A for all items), 8 items in form A (DIF level was A in six items and B in two items), 10 items in form B (DIF level was A in six items and B in four items), and 9 items in form C (DIF level was A in seven items, B in one item and C in one item).

The items with DIF in more than one application are as follows: Item 32 in the Original Form and Form A; item 19 in the Original Form and Form B; item 29 and item 32 in the Original Form and Form C; item 39 in Form A and Form B; item 24, item 32 and item 93 in Form A and Form C.

In the research, Lord's Chi Square test was also used for DIF detection. The findings obtained are given in Table 5.

Table 5

Lord's Chi-square Test Results according to Gender Variable

Implementation	Item Number	$\chi 2$	*p*	Delta Lord	DIF Level
Original Form	29	4.300	0.038	-2.362	C
	42	4.366	0.036	1.870	C
	47	4.468	0.034	1.929	C
	54	6.893	0.008	1.958	C
	58	5.989	0.014	2.257	C
	59	7.498	0.006	1.905	C
	79	6.281	0.012	1.781	C

	84	7.413	0.006	1.979	C
	98	3.852	0.049	1.514	C
Form A	25	10.247	0.001	-2.2866	C
	32	9.177	0.002	-2.830	C
	36	7.167	0.007	-2.419	C
	40	5.239	0.022	-2.189	C
	49	5.782	0.016	-1.910	C
	50	8.341	0.003	-2.350	C
	51	3.888	0.048	-1.692	C
	58	4.334	0.037	-2.196	C
	64	6.052	0.013	-1.998	C
	72	4.512	0.033	-1.483	C
	78	3.916	0.047	-1.510	C
	83	7.323	0.006	-1.897	C
	93	3.893	0.048	-1.531	C
Form B	23	6.478	0.010	2.309	C
	33	3.976	0.046	1.579	C
	42	3.980	0.046	1.580	C
	47	3.877	0.048	1.452	B
	54	6.198	0.012	2.124	C
	57	4.298	0.038	1.422	B
	59	11.402	0.001	2.263	C
	72	4.534	0.033	1.442	B
	85	6.265	0.012	1.776	C
	87	4.137	0.041	1.354	B
	88	4.460	0.034	1.577	C
	93	5.745	0.016	1.848	C
	94	8.363	0.003	2.042	C
	96	7.048	0.007	2.473	C
	97	7.923	0.004	2.606	C
Form C	19	3.972	0.046	-3.28	C
	32	14.192	0.000	-3.357	C
	58	6.357	0.011	-2.687	C
	82	3.849	0.049	-1.345	B

When Table 5 is analyzed, it was determined that there was DIF in nine items in the original form, 13 items in Form A, 15 items in Form B, and 4 items in Form C according to Lord's Chi Square test results. In the original form and form A, the DIF

level is C. In form B, the DIF level is C in eleven items and B in four items. In form C, the DIF level is C in three items and B in one item. In order to compare the results of the Logistic Regression and Lord's Chi-square methods, items containing DIF and item numbers are summarized in Table 6.

Table 6

The Items with DIF and the Number of Items that Have DIF in Different Implementations

	Logistic Regression Results		Lord's Chi Square Results	
Implementation	Item Number	Number of items	Item Number	Number of items
Original	11, 12, 18, 19, 20, 29, 32, 42, 44, 49, 51, 54, 65, 82, 96	15	29, 42, 47, 54, 58, 59, 79, 84, 98	9
A	24, 25, 32, 34, 39, 40, 85, 93	8	25, 32, 36, 40, 49, 50, 51, 58, 64, 72, 78, 83, 93	13
B	17, 19, 23, 37, 39, 59, 69, 76, 77, 88	10	23, 33, 42, 47, 54, 57, 59, 72, 85, 87, 88, 93, 94, 96, 97	15
C	13, 22, 24, 29, 32, 38, 58, 78, 93	9	19, 32, 58, 82	4

As seen in Table 6, the following findings were obtained when Lord's Chi Square results and logistic regression results were compared.

- In the original form, 15 items include DIF according to the logistic regression method. Nine items contain DIF according to the Lord's Chi Square method. Only three items (29, 42, and 54) identified as containing DIF were common in the two methods.
- In Form A, 8 items contain DIF according to the logistic regression method, 13 items include DIF according to the Lord's Chi Square method. Only four items (25, 32, 40, and 93) identified as containing DIF were common in the two methods.
- In Form B, 10 items include DIF according to the logistic regression method, and 15 items include DIF according to the Lord's Chi Square method. Only three items (23, 59, and 88) identified as containing DIF were common in the two methods.
- In Form C, 9 items include according to the logistic regression method, and 4 items include DIF according to the Lord's Chi Square method. Only two items (32 and 58) identified as containing DIF were common in the two methods.

In order to say that DIF is caused by a distractor (e.g. content) in an item, there must be a DIF in the item in both forms including that distractor. In Forms A, B and C, the correct response remained the same and one of the distractors was removed from the item and the item was applied with three options. In this context, if an item has DIF in

both form A and form C and if there is no DIF in form B, it may be due to the distractor in this item. In addition, the presence of the relevant distractor in the item may affect the psychometric feature of the item by changing responder behavior. When the findings were examined in this regard, DIF was determined in items 24, 32, and 93 in forms A and C, while no DIF was found in the same items in form B according to the logistic regression results. DIF was found in item 39 in forms A and B, while no DIF was determined in form C in the same item. According to the Lord's Chi Square results, DIF was found in item 72 in forms A and B, whereas DIF was not found in the same item in form C. Items 32 and 58 contain DIF in forms A and C; however, the same items do not contain DIF in form B. When these results were analyzed, it was seen that the Logistic Regression and Lord's Chi Square methods could not produce similar results except for one item.

Discussion and Conclusion

In this study, the Logistic Regression and Lord's Chi Square methods were compared by performing DIF analysis in four different applications. The two methods produced different results in different applications. When the literature is analyzed (e. g., Başusta, 2013; Erdem, 2015; Gierl, Khaliq, & Boughton, 1999; Gök, Kelecioğlu, & Doğan, 2010), there are findings that different methods used in determining DIF produce different results.

When the distribution of ability is not even in logistic regression analysis, this increases the possibility of type 1 errors (Narayanon & Swaminathan, 1996). French and Maller (2007) undertook a simulation study and used logistic regression analysis in DIF analysis. Yıldırım (2017) suggested that purification of total scores as a criterion for matching did not lead to coherent results and that it did not lead to changes in the number of items with DIF and the levels of DIF. Roznowski and Reith (1999) stated that the existence of biased items in a test did not significantly change the assessment quality. Tian, Pang, and Boss (1994) accordingly reported that the increase in sample size may change the results of logistic regression analysis.

In their study, Ryan and Chiu (2001) investigated the effect of the item content and ordering of the items according to difficulty level on DIF results. In this study, Form 1 questions were created through random ordering. Form 2 was created by ordering the items according to content and difficulty level. The results obtained from the application of the two forms revealed that DIF was not determined depending on the forms the items were in as far as gender variable is concerned. It can be stated that the magnitude of β values differed in different forms and male students were found to be more successful in the test with mixed order in terms of content and difficulty.

In the literature, there are studies which suggested that PPVT has bias towards socio-culturally disadvantaged groups (Washington & Craig, 1999). In the results of Kurnaz and Kelecioğlu (2008), which utilized logistic regression method that obtained data by implementing the test in another sample group, 14 items were found to have DIF in terms of gender. eight of the items (the 29th, 32th, 44th, 49th, 54th, 65th, 82th and 96th items) that were detected to have DIF in Kurnaz and Kelecioğlu (2008) according to the gender variable were also detected in this study. In this sense, the findings of the two studies are partially coherent. When the content of these items is examined, it is seen that they include words such as barbershop, parachute, spider web, hook, joy, and

stadium. The items with content such as barbershop and stadium may be more familiar to boys. This may cause items to contain DIF. In such items, the source of DIF may be the item root rather than the distractor.

Items 24 (content of the item: Insect), 32 (content of the item: parachute), 58 (content of the item: sailboat), and 93 (content of the item: law) include DIF in the original form, Form A and Form C; however, they do not contain DIF in Form B, which may be related to the distractors in these items. Item 72 (content of the item: evaluation) contains DIF in Forms A and B, but not in Form C. In this case, the distractors of this item may need to be re-examined. Future studies may investigate whether the meanings that children attach to the concepts change depending on gender by asking girls and boys their understanding of such concepts. Thus, in tests prepared for young children, the construct to be measured by the test can be explained as operational.

In the literature, there are some findings which suggest that not taking cultural differences into consideration in the tests that have been adapted to different cultures may lead to DIF (Allalouf, 2003; Petersen et al., 2003). In these studies, the significance of DIF analysis in adaptation studies in terms of identifying the psychometric properties of a test was emphasized. A similar suggestion may be made in this study.

A limitation of this study is that total score that were obtained from the implemented test were used to designate the talent criteria of focus and reference groups. In another study, in addition to the total score of the relevant test, other total score that were obtained from another test that assesses the same property can be used as a criterion for talents and abilities and, thus, the difference between the two cases may be discussed.

This study will contribute to the literature by setting an example which demonstrates the effects of different applications of the same test on DIF and also by including a different research design.

Statement of Responsibility

Fatma Betül Kurnaz Adıbatmaz; conceptualization, investigation, resources, software, formal analyses, writing-reviewing, editing, methodology, validation, visualization and supervision. Hüseyin Yıldız; conceptualization, software, formal analyses, writing-reviewing, editing, methodology, validation and visualization.

References

Adebule, S. O. (2013). A study of differential item functioning in Ekiti State Unified Mathematics Examination for senior secondary schools. *Journal of Education and Practice, 4*(17), 43-46.

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*(1), 55-73.

Alordiah, C. O., & Agbajor, H. T. (2014). Bias in test items and implication for national development. *Journal of Education and Practice, 5*(9), 10-13.

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education*, *20*(2), 153-170.

Banks, K. (2009). Using DDF in a post hoc analysis to understand sources of DIF. *Educational Assessment*, *14*(2), 103-118.

Başusta, N. B. (2013). *PISA 2006 fen başarı testinin madde yanlılığının kültür ve dil açısından incelenmesi* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

Camilli, G., Shepard, L. A., & Shepard, L. (1994). *Methods for identifying biased test items* (Vol. 4). Sage.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*(1), 114-140.

Doğan, N., & Öğretmen, T. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel, Ki-kare ve Lojistik Regresyon tekniklerinin karşılaştırılması [The comparision of Mantel–Haenszel, Chi-Square and Logistic Regression Techniques for identifying Differential item functioning]. *Eğitim ve Bilim, 33*(148), 100-112.

Dunn, L. M., & Dunn, L. M. (1959). *Manual for the Peabody Picture Vocabulary Test-revised*. Circle Pines, MN: American Guidance Service.

Erdem, B. (2015). *Ortaöğretime geçişte kullanılan ortak sınavların değişen madde fonksiyonu açısından kitapçık türlerine göre farklı yöntemlerle incelenmesi* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

Forsyth, R. A., & Spratt, K. F. (1980). Measuring problem solving ability in mathematics with multiple-choice items: The effect of item format on selected item and test characteristics. *Journal of Educational Measurement*, *17*(1), 31-43.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with Logistic Regression for differential item functioning detection. *Educational and Psychological Measurement, 67*(3), 373-393.

Gierl, M., Khaliq, S. N., & Boughton, K. (1999). *Gender differential item functioning in mathematics and science: Prevalence and policy implications.* Paper presented at the annual meeeting of the Canadian Society for the Study of Education, Sherbrooke, Quebec.

Gök, B., Kellecioğlu, H., & Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel–Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması

[The comparision of Maentel-Haenszel and Logistic Regression techniques in determinin differential item functioning]. *Eğitim ve Bilim*, *35*(156), 3-16.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement*, *26*(2), 147-160.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *The Journal of Experimental Education*, *43*(1), 40-46.

Harasym, P. H., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of" all of the above" on the reliability and validity of multiple-choice test items. *Evaluation & the health professions*, *21*(1), 120-133.

Hidalgo, M. D., & LÓPez-Pina, J. A. (2004). Differential Item Functioning detection and Effect Size: A Comparision between Logistic Regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Holland, P. W., & Wainer, H. eds. (1993). *Diferential item functioning.* Hillsdale, N.J.: Erlbaum.

Hughes, H. H., & Trimble, W. E. (1965). The use of complex alternatives in multiple choice items. *Educational and Psychological Measurement*, *25*(1), 117-126.

Ikeda, E. (1995). Raters differential item functioning and Mantel-Haenszel procedures Applied to an item analysis of write-answer type test. *Japanese Journal Of Educational Psychology, 43*(3), 343-350.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

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.

Kalaycıoğlu, D. B., & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi [Item bias analyses of the University Entrance Exam]. *Eğitim ve Bilim, 36*(161), 3-13.

Karakaya, İ., & Kutlu, Ö. (2012). Seviye belirleme sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi [An investigation of item bias n Turkish sub Tests in Level Determination Exam]. *Eğitim ve Bilim, 37*(165), 348-362.

Katz, J., Demir, N., Önen, F., Uzlukaya, A., & Uludağ, A. (1972). *Türkçe konuşan çocuklar için Peabody resim kelime testi resim dizisi (Peabody Picture-Vocabulary Test)*. Ankara Rehberlik ve Araştırma Merkezi. Ankara.

Köse, İ. A. (2015). PISA 2009 öğrenci anketi alt ölçeklerinde (Q32-Q33) bulunan maddelerin değişen madde fonksiyonu açısından incelenmesi [Investigation of items in PISA 2009 Student Questionnaire Subscales (Q32-Q33) in terms of differential item functioning]. *Kastamonu Eğitim Dergisi, 23*(1), 227-240.

Kurnaz, F. B., & Kelecioğlu, H. (2008). Investigation of Peabody Picture Vocabulary Test from the point of item bias. *World Applied Sciences Journal, 3*(2), 231-239.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Love, T. E. (1997). Distractor selection ratios. *Psychometrika*, *62*(1), 51-62.

Maij-de Meij, A. M., Kelderman, H., & Van Der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, *45*(6), 975-999.

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.

Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*(5), 793-817.

Meyer, J. P., & Wise, S. (2006). *Including item response time in a distractor analysis via Multivariate Kernel Smoothing*. National Council on Measurement in education (NCME).

Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*(3), 257-274.

Osterlind, S. J. (1983). *Test item bias* (Vol. 30). Sage.

Öner, N. (1997). *Türkiye'de kullanılan psikolojik testler: Bir başvuru kaynağı*. İstanbul: Boğaziçi Üniversitesi Yayınları.

Özgüven, İ. E. (1994). *Psikolojik testler*. Psikoloji Danışma Rehberlik ve Eğitim Merkezi (PDREM) Yayınları.

Penfield, R. D. (2008). An Odds Ratio approach for assessing differential distractor functioning effects under the Nominal Response Model. *Journal of Educational Measurement*, *45*(3), 247-269.

Penfield, R. D. (2010). Modeling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement*, *34*(3), 151-165.

Petersen, M. A., Groenvold, M., Bjorner, J. B., Aaronson, N., ... & Sullivan, M. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of Life Research, 12*(4), 373-385.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, *14*(1), 73-90.

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.

Sehmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, *27*(1), 67-81.

Stump, T. E., Monahan, P., & Mchorney, C. A. (2005). Differential item functioning in the short portable mental status questionnaire. *Research on Aging, 27*(3), 355-384.

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*(1), 48-67.

Temiz, G. (2002). *Okulöncesi eğitimin çocuğun dil gelişimine olan etkisi* (Yayımlanmamış doktora tezi). Selçuk Üniversitesi, Konya, Türkiye.

Terzi, R., & Suh, Y. (2015). An odds ratio approach for detecting DDF under the nested logit modeling framework. *Journal of Educational Measurement, 52*(4), 376-398.

Terzi, R., & Yakar, L. (2018). Differential item distractor functioning analyses on Turkish High School Entrance exam. *Journal of Measurement and Evaluation in Education and Pshchology, 9*(2), 136-149.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice Models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*(2), 161-176.

Tian, F., Pang, X. L., & Boss, M. W. (1994). *The effects of sample size and criterion variable on the identification of DIF by the Mantel-Haenszel and Logistic Regression procedures.* In Annual Meeting of the American Educational Research Association, New Orleans, LA.

Tunç, E. B., & Kutlu, Ö. (2018). İki ve çok kategorili puanlanan maddelerde değişen madde fonksiyonlarının karşılaştırılması [Comparision of differential item functioning for two-category scored and multi-category scored items]. *Başkent University Journal of Education*, *5*(1), 40-50.

Van de Vijver, F. J. R. (2018). Towards an integrated framework of bias in noncognitive assessment in international large-scale studies: Challenges and prospects. *Educational Measurements: Issues and Practice,37*(4), 49-56.

Washington, J. A., & Craig, H. K. (1999). Performances of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in Schools, 30*(1), 75-82.

Washington, J. A. (1996). Issues in assessing the language abilities of African American children. In A. G. Kamhi, K. E. Pollock, & J. Harris (Eds.), *Communication development and disorders in African American children: Research, assessment, and intervention* (pp. 35-54). Baltimore, MD: Paul H. Brookes.

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods*. Retrieved from: http://umu.diva-portal.org/smash/record.jsf?pid=diva2%3A146028&dswid=1318.

Yıldırım, A. (2017). *PISA 2009 okuma becerileri alanındaki maddelerin tek değişkenli ve çok değişkenli eşleştirme yöntemi ile değişen madde fonksiyonlarının incelenmesi* (Yayımlanmamış doktora tezi). Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Yıldırım, H., & Büyüköztürk, Ş. (2018). Using the Delphi Technique and focus-group interviews to determine item bias on the Mathematics Section of the Level Determination Exam for 2012. *Educational Sciences: Theory & Practice*, *18*(2), 447-470.

Yurdugül, H. (2003). *Ortaöğretim Kurumları Seçme ve Yerleştirme Sınavının madde yanlılığı açısından incelenmesi* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, *10*(3), 10-16.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters.