Eurasian Journal of Educational Research 72 (2017) 23-42

Eurasian Journal of Educational Research
www.ejer.com.tr

An Investigation of Ordering Test Items Differently Depending on Their Difficulty Level by Differential Item Functioning *

Ebru BALTA[1] Secil OMUR SUNBUL[2]

ARTICLE INFO

Article History:
Received: 29 Sep. 2017
Received in revised form: 19 Nov. 2017
Accepted: 21 Nov. 2017
DOI: 10.14689/ejer.2017.72.2

Keywords
Item Orderings, Mantel-Haenszel, Logistic Regression, Moodle

ABSTRACT

Purpose : Position effects may influence examinees' test performances in several ways and trigger other psychometric issues, such as Differential Item Functioning (DIF) .This study aims to supply test forms in which items in the test are ordered differently, depending on their difficulty level (from easy to difficult or difficult to easy), to determine whether the items in the test form result in DIF and whether a consistency exists between the methods for detecting DIF. **Research Methods:** Methods of Mantel Haenszel (MH) and Logistic Regression (LR) have been taken into consideration to identify whether the items in the tests involve DIF. The data of the work includes the answers of 300 students in the focal group and the reference group, who sat for three mathematics achievement tests. The data obtained from the tests have been statistically analyzed by using the R- 3.2.0. software program. **Findings:** Results of this study can be summarized with the following findings: "ordering the items differently, depending on their difficulty level, affects the probability of individuals in various groups answering the items correctly; also, LR and MH methods produce different results with respect to the items with DIF, which they have identified similar in terms of magnitude order in the amount of DIF. **Implications for Research and Practice:** In further test-developing studies, in order to identify if DIF emerges when giving the test form which has a different ordering of items, with regard to subjects and cognitive difficulty levels.

* This article has been produced from Ebru BALTA's master thesis.

[1] Corresponding author: Ankara University, Faculty of Educational Science, Department of Measurement and Evaluation, TURKEY, ebrubalta2@gmail.com, ORCID: orcid.org/0000-0002-2173-7189

[2] Mersin University, Faculty of Educational Science, Department of Measurement and Evaluation, TURKEY, secilomur@gmail.com, ORCID: orcid.org/0000-0001-9442-1516

Introduction

In the field of education and psychology, multiple-choice tests are primarily used to determine student performance. In testing situations, one strategy to deter cheating and to enhance test security in test administration is using alternate test forms, or forms constructed with the same items presented in different order. Scrambling, or the rearrangement of the same set of items to create additional test forms, is often used to discourage examinee copying. The assumption is that an examinee's response to a test item is independent of the context in which that item appears, an assumption that has always been a fundamental postulate underlying the derivations of classical test theory formulas and their applications in practical test analysis procedures (Lord & Novick, 1968). However, the responses of examinees who respond to the alternate forms are organized differently; therefore the scores taken from the test can change, and this situation can affect item and test statistics (Barciovski & Olsen, 1975; Kleinke, 1980). A *position effect* occurs when examinees' response behaviors are inadvertently influenced by the position of an item within a test (Kingston & Dorans, 1984; Leary & Dorans, 1985; Yen, 1980). Position effects may influence examinee test performance in several ways. Learning effects occur when items become easier when they are located at the end of the test. On the other hand, a fatigue effect occurs when items become more difficult when they are located later in the test. When examinees experience fatigue or practice effects on the test items, item difficulty estimation might be biased (Hohensinn et all., 2011). Thus, taking test items in different orders can possibly lower the reliability of the test, by causing test items to be perceived more difficult or easier (Leary & Dorans, 1985). Literature has shown that taking account of position effect is important to the test validity of an assessment (Hahne, 2008). Therefore, it is important to determine the location and order of an item within the test form when the test forms are being edited, in order to ensure that the test scores of individuals with the same ability level are controlled to eliminate differences due to one or more variability sources which are not related to the intended variable to be measured. In other words, it is important in terms of ensuring that test scores are not biased.

Bias is defined as the systematic errors of the measurement process; it is a condition that reduces the validity of the important psychometric properties of a test. Item bias occurs when people who have the same ability level but come from different groups and therefore have a different probability of a correct response (Holland & Wainer, 1993). Item bias involves processes that both investigate statistically the differences in responses given to the items and determine the source of the difference. Statistically differentiating the responses given to the items is called "Differential Item Functioning (DIF)" (Camilli & Shephard, 1994). DIF is a function that determines the situation of displaying differences in responding to an item correctly, depending on subgroups in every ability level or psychological structure targeted for measurement with the item. DIF detection methods can be examined in two groups: methods based on Classical Test Theory (CTT), and methods based on Item Response Theory (IRT). Some of the commonly-used approaches based in CTT, such as the Mantel-Haenszel and Logistic Regression, are powerful methods and are

used in dichotomously-scored items for detecting uniform DIF (Camilli & Shepard, 1994). In this study, methods were used that are based on the classical test theories Mantel-Haenszel (MH) and Logistic Regression (LR); as such, these methods will be discussed here briefly.

Mantel-Haenszel Method (MH)

When using the MH procedure based on a chi-square statistic, examinees are divided into levels according to their abilities, based on their total test scores, and a 2 x 2 contingency table is created for each ability level This table is created by cross-classifying each examinee as being either the Focal and Reference group and as having answered a particular item as right or wrong (Camilli & Shepard, 1994, p.105). The first step in the analysis is to calculate the common odds ratio, α_{MH}. Because the interpretation of these values (α_{MH}) is difficult, a logistic transformation is used. This measure is usually transformed into $\beta MH = \log_e(\propto MH)$. The common odds ratio is often transformed into the scale of differences in item difficulty used by the Educational Testing Service (ETS) by the Formula $\Delta_{MH} = -2{,}35\ \beta_{MH}$ (Holland &Wainer, 1993). ETS uses three categories to reflect the degree of DIF in items, labeling these A, B, and C. The categories are then defined by Zieky (1993) as follows: Type A items—negligible DIF: items with $|\Delta_{MH}| < 1$; Type B items—moderate DIF: items with $1 \leq |\Delta_{MH}| < 1.5$; Type C items—large DIF: items with $|\Delta_{MH}| \geq 1.5$.

Logistic Regression Method (LR)

The LR model, when applied to DIF detection, uses item response (0-1) as the dependent variable. Independent variables include group membership, ability, and group-by-ability interaction variables. The procedure for identifying DIF uses logistic regression and consists of fitting the models (Camilli & Shepard, 1994, p.126). A model comparison test can be used to simultaneously detect uniform and nonuniform DIF (Swaminathan & Rogers, 1990). With the Chi-squared (χ^2) test for logistic regression one can compute the statistical tests for DIF. In addition, the chi-squared value of each step is obtained, and the $R^2\Delta$ value is also calculated. Zumbo and Thomas (1996) proposed $R^2\Delta$ as a weighted least squares effect size measure for the LR DIF procedure, which could be used to quantify the magnitude of uniform or nonuniform DIF in items. Zumbo and Thomas (1996) suggested a negligible, moderate, and large classification method for $R^2\Delta$. They proposed $R^2\Delta$ values below 0.13 ($\Delta R^2 < 0.13$) for negligible or A-level DIF; between 0.13 and 0.26 ($0.13 \leq \Delta R^2 \leq 0.26$) for moderate or B-level DIF; and above 0.26 ($\Delta R^2 \geq 0.26$) for large or C-level DIF. We used the Zumbo and Thomas (1996) classification schemes in this study.

This study aims to supply test forms in which items in the test are ordered differently, depending on their difficulty level (from easy to difficult or difficult to easy), to determine whether the items in the test form result in DIF and to determine whether a consistency exists between the methods for detecting DIF. When the related literature is examined, studies are found—in Turkey and internationally—on item ordering in multiple choice tests, item and test statistics, test stress, test anxiety and test performance. Furthermore, some other studies have examined whether a difference exists between individuals from different groups and item ordering. These

studies have investigated item orderings with regard to various variables (e.g. gender, school type, etc.). All of this is to say that DIF has been examined in the relevant literature (Bulut, 2015; Chiu, 2012; Klimko, 1984; Miller, 1989; Ryan & Chiu, 2001). However, there is little research from abroad—and much less in Turkey—examining whether ordering items in respect of difficulty levels in a multiple-choice test creates DIF. Hence, this study is expected to contribute to the literature and to shed light on future studies. The study aims to answer the following questions with the results of the analyses with the following methods:

1. In the analysis performed with MH and LR methods

Is there any item indicating DIF

a. Under the conditions that the focal group takes the test form of item ordering with the difficulty level processing from easy to difficult, while the reference group takes the test form of item ordering with the difficulty level processing from difficult to easy?

b. Under the conditions that the focal group takes the test form of item ordering with the difficulty level processing from difficult to easy, while the reference group take the test form of item ordering with the difficulty level processing from easy to difficult?

c. In the analysis performed with MH and LR methods in both situations, are the items which indicate DIF are the same, or do they differ?

2. In the analysis performed with MH and LR methods

Are the items which indicate DIF in accordance with each other

a. Under the conditions that the focal group takes the test form of item ordering with the difficulty level processing from easy to difficult, while the reference group take the test form of item ordering with the difficulty level processing from difficult to easy?

b. Under the conditions that the focal group takes the test form of item ordering with the difficulty level processing from difficult to easy, while the reference group take the test form of item ordering with the difficulty level processing from easy to difficult?

Method

Research Design

This study aims to determine whether any DIF occurs, depending on: the different test forms in which items in the test are ordered differently; their difficulty level; and various analysis methods. Therefore, it may be considered as a baseline survey. Moreover, the study has a theoretical feature, in terms of giving information about the similarities and differences between the methods that are used in the study.

Research Sample

Participants were selected by purposive sampling method from among the students who study at Mersin University in Turkey, particularly students in the Erdemli and Social Sciences Vocational High School. Since the study has a repeated-measuring basis, the sample group of the study was assigned after some matching and data preview processes; it consisted of 300 students in total (focal group, 150 students, 50%; reference group, 150 students, 50%).

Research Instrument and Procedure

Three tests were used in the research, including two parallel tests on 'Square Roots and Operations with Square Roots', which is the subtitle of the Basic Mathematics Course topic of 'Numbers'. Test 1 was employed for constituting focal and reference groups and for measuring the students' competence levels, in terms of their knowledge and skills in basic mathematics course. Test 2 and Test 3, which are parallel tests, were employed for detecting whether DIF arises as a result of giving different test forms, ordered from difficult to easy and from easy to difficult. In order to make sure participants answered the items in the tests in the presented order, an open-source learning system via computer, called Moodle, was used. Table 1 demonstrates the test implementation design.

Table 1

The Design of Test Implementation

Group	Test 1	Before	After
Focal	✓	Test 2(ED Test Form)	Test 3
Reference	✓	Test 3(DE Test Form)	Test 2

According as the aim of the study, before starting the exam, participants partaking in the practice at the same time were equalized in terms of math knowledge and skills, with respect to their Test 1 scores; after that, they were divided into two groups, the focal group and the reference group. The tests were implemented in a balanced way, by ensuring that the focal group began with Test 2 while the reference group began with Test 3. At a one-week interval, the focal group (who had taken Test 2 in the previous application) was given Test 3, whereas the reference group (who had taken Test 3 in the previous application) was given Test 2. Thus, all participants took all test forms. Sequence effect has been eliminated by using a counter-balanced design.

Validity and Reliability

Before the trial test application, expert opinions were consulted to review the drawn items in terms of some criteria. The opinions were obtained from a group of 10 people consisting of experts in the fields of measurement and evaluation, and mathematics education, as well as teaching assistants of Vocational High Schools who give basic mathematics courses. Fleiss's Kappa coefficient was used to compute inter-rater consistency. Fleiss's Kappa Coefficients was 0.931 for Test 1, while the coefficients for Tests 2 and 3 were 0.930. In line with these results, a perfect

consistency can be considered as occurring the experts (raters). Additionally, experts were consulted about whether items of Test 2 and Test 3 were parallel; consequently, Fleiss's kappa coefficient was found to be 0.947 between Test 2 and Test 3. It can be said that there is a perfect consistency between experts about parallelism of the tests according to this kappa. The trial test form of Test 1 (which consisted of 40 multiple-choice items) was implemented to 365 students, whereas trial forms of parallel Tests 2 and Test 3 were implemented to 167 students repeatedly with a one-week interval. After implementation, the number of items in the tests was reduced to 20. Table 2 and Table 3 indicate test and item statistics about the final test form of Test 1, which consisted of 20 multiple-choice items used to measure students' competence levels in terms of their knowledge and skills in basic mathematics.

Table 2

Test Statistics of Test 1

Number of Items	20
Number of Participants	365
Mean	10.3
Variance	45.02
Standard Deviation	6.71
Skewness	0.22
Kurtosis	1.46
Median	8.00
KR 20	0.94

Table 3

Item Statistics of Test 1

Item Number	Item Difficulty Index	Item Discrimination Index	Item Standard Deviation
1	0.73	0.71	0.44
2	0.68	0.84	0.46
3	0.66	0.79	0.47
4	0.62	0.77	0.48
5	0.61	0.78	0.49
6	0.57	0.68	0.49
7	0.55	0.75	0.5
8	0.53	0.79	0.5
9	0.52	0.74	0.5
10	0.50	0.7	0.5
11	0.49	0.76	0.5
12	0.48	0.81	0.5
13	0.47	0.79	0.5
14	0.46	0.72	0.5
15	0.44	0.77	0.5
16	0.43	0.59	0.49
17	0.42	0.84	0.49
18	0.41	0.69	0.49
19	0.38	0.76	0.48
20	0.3	0.52	0.46

Table 4 and Table 5 indicate test and item statistics of the final test forms of the parallel tests (Test 2 and Test 3), which consisted of 20 multiple-choice items.

Table 4

Test Statistics of Test 2 and Test 3

	Test No.	
	2	3
Number of Items	20	20
Number of Participants	167	167
Mean	10.54	10.73
Variance	40.96	46.92
Standard Deviation	6.4	6.85
Skewness	1.54	1.38
Kurtosis	0.03	0.02
Median	11.00	10.00
KR 20	0.93	0.94

Table 5

Item Statistics of Test 2 and Test 3

Item Number	Item Difficulty Index	Item Discrimination Index	Item Standard Deviation	Item Number	Item Difficulty Index	Item Discrimination Index	Item Standard Deviation
1	0.77	0.46	0.42	20	0.75	0.77	0.43
2	0.71	0.54	0.45	19	0.71	0.86	0.45
3	0.7	0.76	0.46	18	0.7	0.86	0.46
4	0.67	0.66	0.47	17	0.65	0.66	0.48
5	0.66	0.67	0.48	16	0.6	0.79	0.49
6	0.6	0.78	0.49	15	0.56	0.79	0.5
7	0.58	0.72	0.49	14	0.55	0.78	0.5
8	0.57	0.73	0.49	13	0.56	0.81	0.5
9	0.56	0.77	0.5	12	0.55	0.78	0.5
10	0.54	0.7	0.5	11	0.55	0.71	0.5
11	0.53	0.78	0.5	10	0.51	0.8	0.5
12	0.51	0.57	0.5	9	0.5	0.63	0.5
13	0.47	0.64	0.5	8	0.49	0.77	0.5
14	0.45	0.75	0.5	7	0.48	0.75	0.5
15	0.41	0.78	0.49	6	0.47	0.84	0.5
16	0.4	0.74	0.49	5	0.45	0.73	0.5
17	0.38	0.77	0.49	4	0.39	0.72	0.49
18	0.34	0.74	0.48	3	0.37	0.71	0.48
19	0.32	0.56	0.47	2	0.34	0.58	0.47
20	0.21	0.64	0.41	1	0.29	0.53	0.46

Mean and median values of the test scores are close; kurtosis and skewness coefficient values are positive and close to zero; and reliability is observed as quite

high. Moreover, the correlation coefficient between Test 2 and Test 3 is calculated as 0.941. Once the test and item statistics are considered, it can be accepted that the tests are parallel. In addition to statistical parallelism, 8 experts studying in the measurement/evaluation and math education fields were consulted about parallelism of the tests. As a result of this consultation, a Fleiss Kappa consistency coefficient has been computed to check whether the tests are parallel in terms of the content as well. This Fleiss Kappa coefficient was 0.908 between Test 2 and Test 3. In conclusion, a perfect consistency among the experts has been asserted on the parallelism of the tests. In order to reveal the content validity, some experts were asked to evaluate selected items in the final tests, with regard to particular criteria; as a consequence, a Fleiss Kappa consistency coefficient was calculated for each test. This Fleiss Kappa consistency coefficient was found as 0.869 for Test 2 and Test 3. Thus, the perfect consistency among the experts is regarded as an indicator of the content validity.

Data Analysis

DIMTEST T statistic, which is a nonparametric multidimensionality, has been computed by using the Dimpack 1.0 packaged program in order to examine whether the data meet the assumption. According to the analysis results, regarding the unidimensionality of the tests, for Test 1, T=1.391 (p=.082); for Test 2, T=1.389 (p=.082); for Test 3, T=1.230 (p=.109). Therefore, the assumption of unidimensionality was not rejected for three tests. Descriptive statistics of Test 1 were calculated. Table 6 illustrates the descriptive statistics about Test 1.

Table 6

Descriptive Statistics of Test 1

	Group	
	Focal	Reference
Number of Participants	150	150
Mean	8.26	8.27
Median	8.00	8.00
Mode	7.00	7.00
Standard Deviation	4.38	4.39
Variance	19.21	19.34
Skewness	-0.14	-0.12
Kurtosis	0.52	0.53
Minimum	1.00	1.00
Maximum	19.00	19.00

It can be asserted that the focal and reference groups have similar features, and that similar statistical values have been obtained regarding group mean and homogeneity. Once kurtosis and skewness values are investigated, tiny deviations can be observed according to the normal distribution. A Mann-Whitney U test was performed to determine whether focal group and reference group participants significantly differed in their means of rank difference; with respect to these results,

there was no significant difference between the means of rank difference at 0.05 significance level (U= 11247,00, p>.05). Descriptive statistics can provide a view on whether a significant difference prevents DIF analysis for subgroups.

To analyze data for the first sub-problem, two separate MH and LR analyses were performed according to the first condition (focal group takes ED Test Form and reference group takes DE Test Form) and the second condition (focal group takes DE Test Form and reference group takes ED Test Form), after determining focal and reference groups. As a result of the first and second MH analyses, items demonstrating DIF were compared in terms of numbers and their levels. In the LR analysis method, two different analysis results were obtained in order to determine uniform and non-uniform DIF. Independent sample t-tests were performed to indicate the group to which DIF detected items providing an advantage. Items demonstrating DIF as a result of the first and second LR analyses were compared in terms of numbers and their levels.

To analyze data for the second sub-problem—whether the results regarding to DIF are concordant—DIF levels were compared with total number of DIF items, with respect to the findings of both MH and LR analyses in both conditions. Spearman's rank difference correlation coefficient was computed to determine the similarities between two methods regarding item ordering according to the amount of DIF they demonstrated. Analyses determining DIF were conducted with the R.3.0.1 packaged program and the "*difR*" package (Magis, Beland and Raiche, 2015), while the other analyses were performed using SPSS 20.0 and Microsoft Excel 2010.

Results

Findings Related to Items that Demonstrate DIF in Analyses Performed with MH and LR Methods

After the analysis performed with MH method for the first condition, 4 items were discovered to demonstrate moderate level (B) DIF, and 1 item demonstrated large level (C) DIF. One of the items showing B level DIF (item 15) was observed to have a medium difficulty level. One of the other items showing B level DIF (item 17) has a high difficulty level (difficult item); in addition, it is in support of the examinees given the ED test form (focal group). The other two items showing B level DIF (items 18 and 19) have a high difficulty level (difficult item) and are in favor of the examinees given the DE test form (reference group). Item 7, with a C level DIF, has a medium level difficulty and is support of the group that took the DE test form.

After the analysis performed for the second condition, 1 item was discovered to demonstrate moderate level (B) DIF, and 5 items demonstrated large level (C) DIF. Item 20 with a B level DIF has a high difficulty level (difficult item) and is in favor of the group that took the ED test form (reference group). Two of the 5 items demonstrate C level DIF: item 5, which has a low difficulty level (easy item); and item 13, which has a medium level difficulty. Both are in support of the group that took the DE test form (focal group). Two of the remaining 3 items (items 7 and 16) have a medium level difficulty, and the last item (item 19) has a high difficulty level

(difficult item); these are in favor of the group that took the ED test form (reference group). The graphs of the Δ_{MH} values of the test items are shown in Figure 1 and Figure 2.

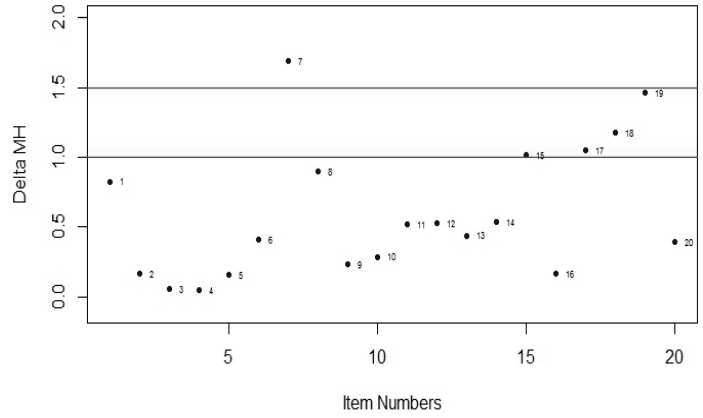

Figure 1. The Δ_{MH} values of items related to the MH analysis performed for the first circumstances

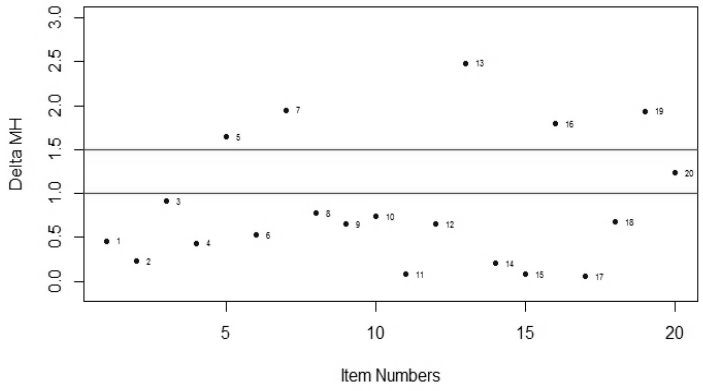

Figure 2. The Δ_{MH} values of items related to the MH analysis performed for the second circumstances

An analysis performed with the LR method was used to identify whether the items in the tests demonstrate both uniform and non-uniform DIF in the first condition; in these results, it appears that no item demonstrates moderate level (B) and/or large level (C) DIF. An analysis performed with the LR method was used to identify whether the items in the tests demonstrate uniform DIF in the second condition; in these results, it can be seen that 1 item has demonstrated moderate level (B) DIF. One of the items having B level DIF (item 13) has a medium difficulty level

and is in favor of the group that took the DE test form (focal group). One of the other items having B level DIF (item 17) has a medium difficulty level and is in favor of the group that took the ED test form (reference group). Another analysis was performed with the LR method to identify whether the items in the tests demonstrate non-uniform DIF in the second condition; in these results, it appears that no item demonstrates moderate level (B) and large level (C) DIF. The graphs of the $R^2\Delta$ values of the test items are shown in Figure 3, Figure 4, Figure 5, and Figure 6.

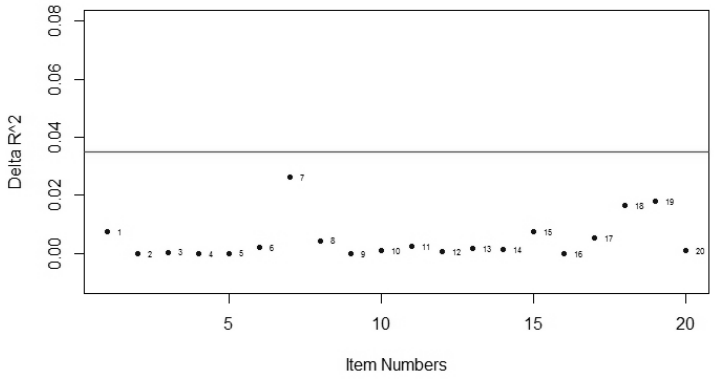

Figure 3. The $R^2\Delta$ values of items related to the LR analysis used to identify uniform DIF performed for the first circumstances

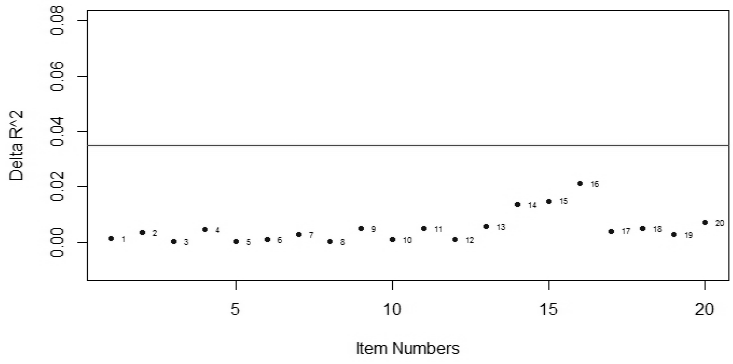

Figure 4. The $R^2\Delta$ values of items related to the LR analysis used to identify non-uniform DIF performed for the first circumstances

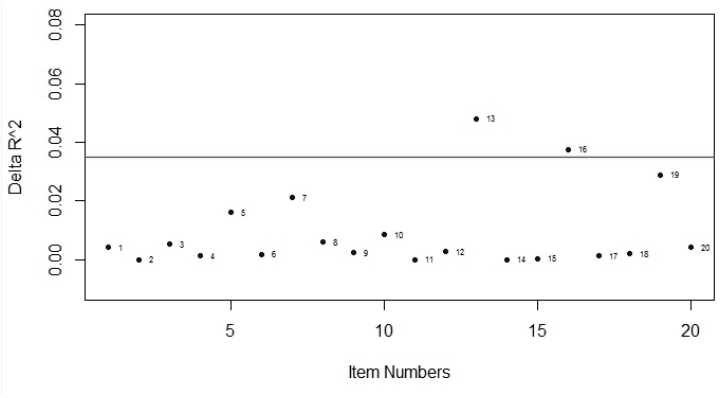

Figure 5. The $R^2\Delta$ values of items related to the LR analysis used to identify uniform DIF performed for the second circumstances

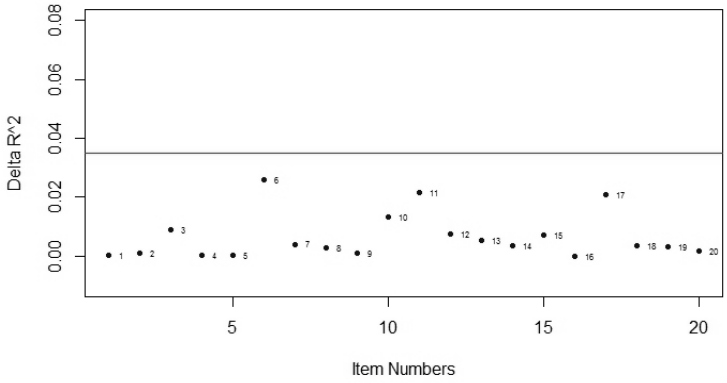

Figure 6. The $R^2\Delta$ values of items related to the LR analysis used to identify non-uniform DIF performed for the second circumstances

Findings whether Items Demonstrating DIF Correspond to Each Other in the Analyses Performed with MH and LR Methods

To determine similarities in magnitude order in the amount of DIF, Spearman's rank correlation coefficient was computed between chi-square values. These values were obtained by the LR method used to identify whether the items in the tests demonstrate uniform DIF and MH method. As a result of the calculations for both conditions, a statistically significant relationship can be seen between magnitude orders in the amount of DIF of the two methods ($r_1 = 0.90$, $r_2 = 0.92$, $p < .01$).

Discussion and Conclusion

Evaluation of MH Method-Analyze Results

When examinees from the focal group took the ED test form first, DIF emerged in favor of this group for the high difficulty level items (difficult items). On the other hand, DIF did not emerge in favor of this group for the low difficulty level items (easy items) when they took the DE test form later. As for examinees from the reference group, who took the ED test form second, it was observed that items with DIF at the moderate level increased, while DIF was highest at the high-difficulty level items. In light of these findings, the ordering of taking the test forms in other words, order effect can be said to affect the probability of answering the items correctly.

According to the findings of both analyses examined in terms of low difficulty level items (easy items), it can be argued that the group who took easy items later had a higher probability of answering the items correctly. Encountering items with medium difficulty after encountering easy or difficult items affects the probability of answering the items correctly. Similarly, it was found that it does not matter if difficult items are at the beginning or the end of test; their ordering affects the probability of answering the items correctly according to the findings of both two analyses. Therefore, it can be concluded that correct response probability is affected by encountering particularly difficult items both in the beginning and at the end of test or encountering items with medium difficulty after easy or difficult items. Additionally, analyses performed with the MH method revealed that the number of items that have DIF, and the items with DIF, differentiate according to the DIF level and the groups they support.

Therefore, it can be concluded from this study that ordering items differently depending on their difficulty level affects the probability of examinees in various groups answering the items correctly. Furthermore, it can be concluded that the placement of difficult items at the end of the test leads to an increased difference in the probability of the items being answered correctly. This finding is in agreement with learning effects, which exist in cases where items are put in order from easy to difficult—in other words, placing difficult items at the end of the test. In addition, this finding is consonant with other studies that have concluded that different item orderings (with respect to difficulty levels) make a significant difference in individuals' test performances (Barciovski & Olsen, 1974; Louisa, 2013).

Evaluation of LR Method-Analyze Results

The results of the analysis performed with the LR method (used to identify both uniform and nonuniform DIF in the first condition) indicated that there is no difference between the probabilities of examinees in both groups answering the items correctly; in other words, individuals from both the focal and the reference group have similar responses to the items. In the second condition, however, both the focal and the reference group showed higher performance on two item. In addition, examinees from different sub-groups differed in the probability of answering two items correctly. Thus, it can be concluded that encountering items with medium difficulty after easy or difficult items influences the probability of the items being answered correctly. Moreover, taking a test form first or second can affect the examinees' correct response probability, that is, order effect can influence the correct response probability.

Comparison of MH and LR Analyses Results

Concerning the findings, the LR and MH methods used in two analyses revealed similar consequences in terms of magnitude order in the amount of DIF; however, they produced different results with respect to the items with DIF. The MH method is more sensitive than the LR method with regard to the number of items containing DIF. This sensitivity might be explained, as the MH method estimates item parameters of the focal and reference groups at the same time, and thus the total sample size is larger than LR. From this point of view, the reason for why LR finds fewer items with DIF might be regarded as stemming from sample size (Penfield & Camilli, 2007). It has been stated that the LR method may reveal more sensitive results in larger samples (Jodoin & Gierl, 2001; Pang et al., 1994). Other studies have also failed to find an exact accordance between these two DIF determining methods (Betrand & Bouteau, 2003; Gomez, Benito & Navas Ara, 2000). Although several studies have argued that the MH method is more powerful in identifying DIF and gives more consistent results (Betrand & Bouteau, 2003; Narayanan & Swaminathan, 1994), other studies have argued that the LR method is one of the most effective and recommended methods in the literature (Clauser & Mazor, 1998; Wiberg, 2007). Despite this fact, similarity in terms of magnitude order in the amount of DIF, and difference in the criteria used for identifying the items with DIF, are considered to produce variation in DIF levels and the number of items with DIF. Some recommendations for future research are as follows:

Future studies may use other methods based on CTT or other methods based on IRT, in order to identify DIF. Results from these various studies may be compared.

Future studies may investigate whether DIF (in terms of lower skill levels) is caused by giving different test forms in which the items are encountered in different orders.

References

Barcikovski, R. S., & Olsen, H. (1975). Test item arrangement and adaptation level. *The Journal of Psychology, 90(1),* 87-93. doi: 10.1080/00223980.1975.9923929.

Bertrand, R., & Boiteau, N. (2003). *Comparing the stability of IRT-based and non IRT-based DIF methods in different cultural context using TIMSS data.* (EDRS Reports – Research -143 , ED 476 924, TM 034 975). Quebec, Canada: NA. (ERIC Document Reproduction Service No. ED476924).

Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. *European Journal of Research on Education*, *3(1),* 7-16. Retrieved from http://iassr2.org/rs/030102.pdf

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

Clauser, B. E., & Mazor, K. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement, Issues and Practice. 17(1),* 31–44. doi: 10.1111/j.1745-3992.1998.tb00619.x.

Chiu, P. (2012,April) . *The Effect of item position on state mathematics assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, Vancouver, Canada.

Gomez-Benito, J., & Navas-Ara, M. J. (2000). A comparison of Ki-kare, RFA and IRT based procedures in the detection of DIF. *Quality ve Quantity*, *34(1),*17–31. doi: 10.1023/A:1004703709442.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally Incomplete data. *Psychology Science Quarterly*, *50(3),* 379-390. Retrieved from http://journaldatabase.info/articles/analyzing_position_effects_within.html

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation*, *17(6)*, 497-509. doi: 10.1080/13803611.2011.632668.

Holland, P. W., & Wainer, H. E. (1993). Differential item functioning. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Jodoin, M. G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14(4),* 329-349. doi:10.1207/S15324818AME1404_2.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8(2),* 147–154. Retrieved from doi: 10.1177/014662168400800202.

Kleinke, D. J. (1980). Item order, response location, and examinee sex and handedness on performance on multiple-choice tests. *Journal of Educational Research, 73(4)*, 225–229. doi:10.1080/00220671.1980.10885240.

Klimko, I. P. (1984). Item arrangement, cognitive entry characteristics, sex, and test anxiety as predictors of achievement examination performance. *Journal of Experimental Education, 52(4),* 214-219. doi: 10.1080/00220973.1984.11011896.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A Historical Perspective on Immediate Concern. *Review of Educational Research, 55(33),* 387-413. doi: 10.3102/00346543055003387.

Lord, F. M.,& Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Welsley Publishing Company.

Louisa, N. (2013). Effect of item arrangement on test reliability coefficients: implications for testing. *Journal of Research in Education and Society*, *4(3),* 54-62. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/001316447303300224?journalCode=epma

Magis, D., Beland, S., & Raiche, G. (2015). *difR: Collection of methods to detect dichotomous differential item functioning (DIF).* [Computer software]. Available from http://CRAN.R-project.org/package=difR

Miller, S. K. (1989). *Interaction effects of gender and item arrangement on test and item performance (*Unpublished doctoral dissertation). University of Nebraska, Lincoln.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18(4),* 315-338. doi: 10.1177/014662169401800403.

Pang, X. L., & et all. (1994,April). *Performance of Mantel-Haenszel and Logistic Regression DIF procedures over replications using real data.* Paper presented at the Annual Meeting of the American Educational Research Association,New Orleans, LA.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics psychometrics* (Vol.26, pp. 125–167). Amsterdam: Elsevier.

Ryan, K. E. ,& Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14 (1),* 73–90. doi:10.1207/S15324818AME1401_06.

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: a theoretic comparison of methods* (EM No. 60). Umea, Sweden: Umea University, Department of Educational Measurement.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27(4)*, 361-370. doi: 10.1111/j.1745-3984.1990.tb00754.x.

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, Inc.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17(4)*, 297–311. doi: 10.1111/j.1745-3984.1980.tb00833.x.

Zumbo, B. D., & Thomas, D. R. (1996). *A measure of effect size for a model-based approach for studying DIF* (Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science). Prince George, Canada: University of Northern British Columbia.

Maddeleri Güçlüklerine Göre Farklı Sıralamanın Birey Tepkilerine Etkisinin Değişen Madde Fonksiyonuyla İncelenmesi

Atıf:

Balta, E. & Sunbul Omur, S. (2017). An investigation of ordering test items differently depending on their difficulty level by differential item functioning. *Eurasian Journal of Educational Research,* 72, 23-42, DOI: 10.14689/ejer.2017.72.2

Özet

Problem Durumu: Bireylerin maddelere verdiği tepki davranışlarının, maddenin, test içerisindeki sırasından beklenmedik şekilde etkilenmesi sıra etkisi (position effect) olarak tanımlanmaktadır. Sıra etkisi bireyin test performansını çeşitli şekillerde etkilemektedir. Madde güçlüğü açısından kolay ve zor maddelerin testin başında ya da sonunda yer almasına bağlı olarak öğrencilerin test boyunca motivasyonları artıp ya da azalmakta ve böylece test puanları etkilenmektedir. Ayrıca, maddelerin güçlük düzeylerine göre kolaydan zora doğru sıralandığı, yani madde güçlüğü açısından zor maddenin testin sonlarına doğru yer aldığı durumlarda pratik ya da öğrenme etkisi (learning effect), madde güçlüğü açısından kolay maddelerin testin sonlarına doğru yer aldığı durumlarda ise yorgunluk etkisi (fatique effect) gözlenmekte ve böylece maddelerin güçlük düzeyleri farklı değerler alabilmektedir. Literatür incelendiğinde madde sıra etkisinin göz önünde bulundurulması test geçerliğini değerlendirmede önemli olduğu görülmektedir.

Araştırmanın Amacı ve Önemi: Bu çalışmada, maddelerin güçlük düzeylerine göre test içerisinde farklı sıralarda (kolaydan zora ve zordan kolaya) yerleştirildiği farklı test formlarının verilmesinin, testte yer alan maddelerde DMF oluşturup

oluşturmadığının ve kullanılan DMF belirleme yöntemleri arasındaki uyumun belirlenmesi amaçlanmıştır. Çoktan seçmeli bir testte yer alan maddelerin güçlük düzeylerine göre sıralanmasının maddelerde Değişen Madde Fonksiyonu (DMF) yaratıp yaratmadığına ilişkin yurtdışında çok az çalışmaya rastlanmış olup yurtiçinde ise doğrudan bir çalışmaya rastlanamamıştır. Bu açıdan, bu çalışmanın alan yazına katkı sunacağı ve bu tarz çalışmalara ve geniş çapta yapılan sınavlara da ışık tutacağı düşünülmektedir.

Araştırmanın Yöntemi: Araştırmada, araştırmacı tarafından ikisi paralel olmak üzere toplamda üç adet Matematik Başarı Testi kullanılmıştır. Testlerden biri, odak ve referans gruplarının oluşturulması için, öğrencilerin Temel Matematik dersindeki bilgi ve becerileri açısından yetenek düzeylerinin belirlenmesinde ve paralel olan diğer iki test ise, maddelerin güçlük düzeylerine göre kolaydan-zora ve zordan-kolaya sıralanarak verilmesi durumunun DMF yaratıp yaratmadığının tespit edilmesinde kullanılmıştır. Öğrencilerin testlerdeki maddeleri, testlerde yer alan sıraya göre cevapladıklarından emin olmak için testler, bilgisayar ortamında Moodle açık kaynak kodlu uzaktan eğitim sistemi kullanılarak uygulanmıştır. Araştırmanın çalışma grubunu, amaçlı örnekleme yöntemiyle seçilen, araştırmanın tekrarlı ölçümlere dayanmasından kaynaklı olarak yapılan eşleştirme ve veri ön izleme süreçlerinin ardından belirlenen, toplamda 300 (odak grup (150 öğrenci) ve referans grup (150 öğrenci)) öğrenci oluşturmaktadır. Uygulamaya katılan öğrencilerin üç test formunu da alması sağlanmıştır. Karşıt dengelenmiş desen kullanılarak testlerdeki sıra etkisi ortadan kaldırılmıştır. Testlerde yer alan maddelerin DMF içerip içermediği Mantel-Haenszel (MH) ve Lojistik Regresyon(LR) yöntemleriyle odak grubunun KZ (maddelerin kolaydan zora doğru sıralandığı test formu), referans grubunun ZK (maddelerin zordan kolaya doğru sıralandığı test formu) test formunu alması (birinci durum) ve odak grubunun ZK, referans grubunun KZ test formunu alması durumuna (ikinci durum) göre belirlenmiştir. Bu testlerden elde edilen veriler R-3.2.0 ve "*difR"* paketi kullanılarak analiz edilmiştir.

Araştırmanın Bulguları: Birinci duruma göre, MH yöntemiyle yapılan analiz sonuçlarına göre DMF gösteren maddelerden, dört tanesinin orta düzeyde (B), bir tanesinin de yüksek düzeyde (C) DMF gösterdiği belirlenmiştir. B düzeyinde DMF gösteren maddelerden bir tanesinin orta güçlükte madde, bir tanesinin ise zor madde ve KZ test formunu alan öğrencilerin (odak grup) lehine olduğu ve diğer iki tanesinin ise zor madde ve ZK test formunu alan öğrencilerin (referans grup) lehine olduğu görülmektedir. C düzeyinde DMF içeren maddenin ise, orta güçlükte bir madde olduğu ve ZK test formunu alan öğrencilerin lehine olduğu görülmektedir. LR yöntemiyle hem TB DMF hem de TBO DMFyi belirlemek için yapılan analizlerde ise, orta düzeyde (B) ve yüksek düzeyde (C) DMF gösteren maddenin bulunmadığı görülmektedir. İkinci duruma göre, MH yöntemiyle yapılan analiz sonuçlarına göre, bir maddenin orta düzeyde (B), beş maddenin de yüksek düzeyde (C) DMF gösterdiği belirlenmiştir. B düzeyinde DMF içeren maddenin zor madde olduğu ve KZ test formunu alan öğrencilerin lehine olduğu, C düzeyinde DMF gösteren iki maddeden bir tanesinin kolay madde, bir tanesinin ise orta güçlükte madde ve ZK test formunu alan öğrencilerin lehine olduğu ve üç maddeden iki tanesinin orta

güçlükte madde ve bir tanesinin ise zor madde olduğu ve KZ test formunu alan öğrencilerin lehine olduğu görülmektedir. LR yöntemi ile TB DMF'yi belirlemek için yapılan analiz sonucuna göre iki maddenin orta düzeyde (B) DMF gösterdiği belirlenmiştir. Orta düzeyde (B) DMF gösterdiği belirlenen iki maddenin de orta güçlükte madde olduğu ve maddelerden bir tanesinin ZK test formunu alan öğrencilerin lehine diğerinin ise KZ test formunu alan öğrencilerin lehine işlediği görülmektedir. Yöntemlerin maddelerdeki DMF miktarlarının büyüklük sıralaması bakımından benzerliklerinin belirlenebilmesi için, test maddelerinin TB DMF gösterip göstermediğini belirleyebilmek için yapılan LR ve MH yöntemlerine göre elde edilen ki-kare değerleri arasında Spearman sıra farkları korelasyon katsayısı hesaplanmıştır. Hesaplamalar sonucunda her iki durum için de, iki yöntemin DMF büyüklük sıralamaları arasında istatistiksel olarak manidar bir ilişkinin bulunduğu görülmektedir ($r_1 = .90$, $r_2 = .92$; $p < .01$).

Araştırmanın Sonuçları ve Önerileri: Araştırmanın bulguları, güçlük düzeyi düşük olan maddeler açısından incelendiğinde, kolay maddeleri sonra alan grubun, maddeleri doğru cevaplama olasılıklarında artışların olduğu söylenebilir. Orta güçlükte yer alan maddelerin, her iki uygulamada da hem odak hem de referans grubunun lehine işlediği görülmektedir. Bu durumda, orta güçlükteki maddelerin kolay ya da zor maddeden sonra gelmesinin maddenin doğru cevaplama olasılığını etkilediği söylenebilir. Güçlük düzeyi yüksek olan maddeler (zor maddeler) açısından, her iki analize dair bulgular incelendiğinde, zor maddelerin hem testin başında yer aldığı durumda hem de testin sonunda yer aldığı durumda, maddelerin doğru cevaplandırılma olasılığını etkilediği söylenebilir. Böylece bu çalışmada, maddelerin güçlük düzeylerine göre farklı şekilde sıralanmasının farklı gruplarda yer alan bireylerin, maddelere, doğru cevap verme olasılıklarını etkilediği sonucuna ulaşılmıştır. Ayrıca, zor maddelerin test formunun sonunda yer alması, maddelerin cevaplanma olasılığındaki farklılığın artmasına neden olmaktadır. Ayrıca yapılan her iki analizde kullanılan LR ve MH yöntemlerinin, DMF miktarlarındaki büyüklük sıralamalarında benzer, DMF'li maddeler bakımından farklı sonuçlar ürettiği sonucuna ulaşılmıştır. DMF içeren madde sayısı bakımından, MH yönteminin LR yönteminden, daha duyarlı olduğu görülmektedir. Bu araştırma kapsamında, DMF'nin belirlenmesinde, Klasik Test Kuramı'na dayalı yöntemlerden MH ve LR yöntemleri kullanılmıştır. Daha sonraki yapılacak olan çalışmalarda, KTK'ya dayalı diğer yöntemler ve IRT'ye dayalı yöntemlerle DMF belirlenebilir. Farklı yöntemlerden elde edilecek sonuçlar karşılaştırılabilir.

Anahtar Kelimeler: Madde Sıralamaları, Mantel-Haenszel, Lojistik Regresyon, Moodle.