Eurasian Journal of Educational Research 83 (2019) 167-182

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

Comparison of Different Forms of a Test with or without Items that Exhibit DIF

Onder Kamil TULEK[1], Ibrahim Alper KOSE[2]

ARTICLE INFO

Article History:
Received: 07 Jun.2019
Received in revised form: 07 Aug. 2019
Accepted: 16 Sept. 2019
DOI: 10.14689/ejer.2019.83.8

Keywords
purification, the estimate of ability, DIF

ABSTRACT

Purpose: This research investigates Tests that include DIF items and which are purified from DIF items. While doing this, the ability estimations and purified DIF items are compared to understand whether there is a correlation between the estimations.

Method: The researcher used to R 3.4.1 in order to compare the items and after this situation; according to manipulated factors, we carried out the data production under different circumstances with the help of simulation study. The manipulated factors were determined levels of sample size (1000, 2000), test length (40, 60) and percentage of DIF (%5, %10). By using the new data each condition's DIF items' ability estimations were carried out. Afterward, DIF items purified from the tests and later the abilities were estimated. The correlation between the ability parameters was calculated by using the Spearman's Rank Correlation Coefficient and these parameters were calculated separately according to the eight conditions.

Findings: After calculations, all of the coefficients of correlations (rs)' values were almost zero ($p<0.01$). In other words the test length 40 and 60, sample size 1000 and 2000, percentage of DIF %5 and %10, when we crossed these parameters in different eight conditions, there was no familiar correlation between the tests that include DIF items and tests of that purified from DIF items. Besides, there was no correlation between the tests thinking the ability estimations; if we exclude DIF items from the tests, the individuals' test ranking changes, too.

Implication for Research and Practice: This study showed that tests that include DIF items affect the ability estimation of individuals. In the frame of this result, teachers, administrators, and policymakers should bear in mind tests DIF potential. Also, this study may be carried out by using various conditions.

© 2019 Ani Publishing Ltd. All rights reserved

[1] Milli Egitim Bakanlığı, Düzce.TURKEY,E-Mail: onderkamilt@gmail.com, ORCID: https://orcid.org/0000-0002-9343-5868,
[2] Abant İzzet Baysal Üniversitesi, TURKEY, E-Mail: i.alper.kose@gmail.com, ORCID: https://orcid.org/0000-0003-0842-1929,

Introduction

Measurement and evaluation studies, which are an indispensable part of the education system, make it possible to observe whether or not the targeted characteristics are acquired by individuals during the education and training process or to note the extent of their acquisitions and based on these observation results, several decisions can be made for these individuals. However, these decisions, which may be of vital importance to individuals at times, should be based on structurally acceptable foundations which underline the necessity of objective evaluations. Objective evaluations in education and psychology are only possible through the use of measurement tools which should have distinct properties so that results obtained from the evaluation process can be used in line with the purpose of evaluation. These are validity, reliability, and usability.

In the classical sense, the validity of an assessment tool is defined as the ability of the tool to measure the quality of the desired trait without the interference of any other trait. However, in recent years, current definitions exist for validity and a new formation is discussed. Validity can be defined in a broader and more contemporary manner as the extent of support for the interpretations made on test scores based on the purpose of test both by theoretical means and by the evidence collected (AERA, APA, and NCME, 2014; cited in Kelecioğlu and Şahin, 2014).

Many factors threaten the validity of the measurement tool. The scope of the assessment tool, the reliability of the scores, the length of the assessment tool, the average difficulty, inadequate examination periods and cheating, etc. are the factors that can pose a threat to validity (Turgut and Baykul, 2015). Bias is one of the factors that threaten the validity of the measurement tool (Clauser and Mazor, 1998). Bias is the advantage provided by the test, based on the conditions not covered by the purpose of the test or based on the properties of test items, to one of the groups at the same ability level but is included in different subgroups (Zumbo, 1999).

The purpose of studying item bias is to determine whether the difference between the subgroups of individuals at the same ability level originates from an actual difference in the measured property or from the assessment process. The first thing that should be done in bias studies is to determine whether there are any differences between the response structures of subgroups in responding to items. Determining whether there are differences between the response structures of subgroups is possible via differential item functioning (DIF) analysis. DIF is different from the concept of item bias. Hambleton, Swaminathan, and Rogers (1991) reported that an item exhibits DIF if individuals having the same ability level, but from different subgroups based on gender, race, etc., do not have the same probability of getting the item right. As it is seen, the common idea in both DIF definitions is that individuals with the same ability level are expected to respond in a similar manner to items. If responses differ, it can be argued that DIF is present in that item. In order to be declared biased, the item must first exhibit DIF. However, the presence of DIF in an item does not mean that the item is definitively biased. In other words, each biased item exhibits DIF; but not all items exhibiting DIF are biased. In order to determine the bias of an item with

DIF, the possible causes of DIF should be determined and the expert opinion should be consulted on whether the item is advantageous for one of the subgroups other than the structure it intends to measure (Camilli and Shepard, 1994; Zumbo 1999). Within the scope of this study, only DIF studies were carried out on the items exhibiting DIF in the data set. No expert opinion was sought as to whether these items demonstrated possible bias. In other words, studies on DIF, the first step of bias determination studies, were carried out by the researcher and bias determination study which is a continuation of this step were not carried out.

According to Dorans and Holland (1993), individuals from different subgroups in DIF analyzes should be similar in terms of the properties that the test aims to assess, i.e. they should be matched at the same ability level. As a matter of fact, DIF investigations are based on the assumption that the likelihood of responding to an item is similar for the groups which are similar in terms of the properties the test wants to measure. In DIF analyzes, the total scores of individuals, especially from the tests based on binary scoring, are used as matching criteria. Similarly, Clauser, Mazor, and Hambleton (1993) stated that reference and focus groups could be matched by the use of valid subtest scores. However, the degree of purification of the relevant sub-test scores of the variable to be used as the matching criterion is an issue that needs attention and care.

The total scores used as the matching criterion which are calculated by using the responses to the items in a test are the total scores obtained by taking into account the items exhibiting DIF if the test contains items that exhibit DIF. The process of calculating the total score by subtracting the mentioned DIF items from this test is called purification. Briefly, purifying the matching criterion means the removal of items exhibiting DIF from the test while calculating total scores; thus, it is ensured that only DIF-free items are used for the necessary analyzes (Lee and Geisinger, 2016).

Tests with specific properties are used to measure psychological characteristics such as ability and achievement. Based on the results, it is necessary to prove the validity of these tests which are used to make important decisions about individuals. Studies exist which demonstrate the presence of items that exhibit DIF, which is a significant threat to the validity of the test items used in national-level large scale tests which require ranking (Bakan Kalaycioglu and Kelecioglu, 2011; Basusta, 2013;Cepni 2011; Demir, 2013; Dogan and Ogretmen, 2008; Erdem, 2015; Gok, Kelecioglu and Dogan 2010; Ogretmen 2006; Yildirim, 2017). However, studies that investigate the changes that will occur in achievement ranking in test areas based on recalculations after the removal of the items that exhibit DIF are not conducted often. It is believed that the presence of items that provide advantage to a certain group in the test may cause inequality and injustice among individuals in such examinations where vital decisions are taken about individuals. Therefore, tests should be purified from these items. This study aimed to compare the ability estimates predicted from test forms that contained items with or without DIF based on different number of items, different sample size and different DIF ratio conditions.

Method

Research Model

This study, which aimed to compare the ability estimates predicted from a test form that contained items with or without DIF based on a different number of items, different sample sizes and different DIF ratio conditions, utilized relational screening model.

Simulation Conditions

In this study, the comparison of the predictive estimates of a test form that contained items with or without DIF under various conditions was carried out by a simulation study. The conditions that were constant and manipulated in data generation for this simulation study are described below.

Constant conditions

The simulation data were generated in accordance with the items that were scored based on the two-category structure in the study. The uniformity of the items that exhibit DIF was another constant condition of the study. In addition, in all conditions, ability parameters of individuals were obtained according to a standard normal distribution with a mean of 0 and a standard error of 1. The generation of data fit for the Item Response Theory (IRT) model was based on a three-parameter logistic model. For this model, the mean and standard deviation or minimum and maximum values of a, b and c parameters were determined and data were generated between these values.

Manipulated conditions

The literature on DIF studies shows many variables, such as test length, sample size, and the proportion of items exhibiting DIF, have an effect on DIF (Clauser, Mazor and Hambleton, 1993; Narayanan and Swaminathan 1996; French and Maller 2007; Atar and Kamata, 2011). In this study, these conditions were manipulated based on the determined levels of the related conditions for data generation.

The number of items (k): For this condition, two levels were determined as k=40 and k=60. Standardized achievement and ability tests generally have between 35-80 items (Narayanan and Swaminathan 1996; French and Maller 2007). Sample size (n): Two levels were determined for the sample size of the study as n=1000 and n=2000. In simulation studies conducted on IRT based DIF determination methods, it was found that the minimum sample size for each group was 200-250 and 600 people in total (Narayanan and Swaminathan 1996; French and Maller 2007; Atar and Kamata, 2011).

The proportion of items exhibiting DIF: There were two levels in the proportion of items exhibiting DIF as d=5% ve d=10% in this study since according to Jodoin and Gierl (2001), higher proportions of items exhibiting DIF would threaten test validity. In addition, it was found that tests that were investigated in DIF studies included more than one item that exhibited DIF.

Data Generation

In this simulation study, data generation was performed by writing codes to R 3.4.1 program based on a three-parameter logistics model. 50 replications were performed for each condition considered.

Table 1 displays the study plan for the simulated data generation performed according to the levels of each of the manipulated conditions such as the number of items, sample size and the proportion of items exhibiting DIF.

Table 1.

Simulative Data Generation Plan

K	Number of Items	Sample Size	The proportion of Items Exhibiting DIF
1	40	1000 (R:500/O:500)	5%
2	60	1000 (R:500/O:500)	5%
3	40	2000 (R:1000/O:1000)	5%
4	60	2000 (R:1000/O:1000)	5%
5	40	1000 (R:500/O:500)	10%
6	60	1000 (R:500/O:500)	10%
7	40	2000 (R:1000/O:1000)	10%
8	60	2000 (R:1000/O:1000)	10%

K: Condition, R: Reference group, O: Focus group

Data based on the planned conditions were obtained from the normal distribution in which the mean *parameter a* was 0.8 and standard deviation was 0.04 under all conditions. The minimum and maximum values of *parameter b* were identified to be -2 and +2. Finally, the value range of *parameter c* was determined to be between 0.2- 0.3 and then the data were generated. 0.75 was added as the amount of DIF to the b parameters of the respective items in accordance with the number of items required for the production of items that exhibit DIF with respect to the determined levels of the proportion of items exhibiting DIF as manipulated condition.

Data Analysis

For the purpose of the study and in accordance with the conditions described above, the "difR" package was used in the R program to generate data sets that contained items with DIF. "ltm" package of the R program based on the three-parameter logistic model of IRT was used to conduct ability estimations of individuals based on their test responses in the data tests iteratively generated according to each condition. Individuals' abilities were re-estimated by removing the items that exhibited DIF from the same test under each condition. Ability estimations of individuals for the test containing items that exhibited DIF, i.e., the θ (theta) values of individuals, were determined to be θ_1 while ability estimations of individuals for the test with no DIF items were determined to be θ_2.

The relationship between θ_1 and θ_2 for each iteration was examined by SPSS 22.0 program, by using Spearman *Rho Correlation* Analysis. Mean correlation coefficients (r_s) obtained by Spearman *Rho Correlation* Analysis in each iteration at the same condition were calculated. Fisher-Z transformation proposed by Corey, Dunlap, and Burke (1998) was performed to obtain more clear results in calculating mean correlation coefficients. For this purpose, each r_s coefficient was converted to z value with Fisher-Z transformation, then the mean z values of the transformed values were calculated and the obtained mean z value was re-converted to r_s by Fisher-Z transformation. In this way, a relationship existed between θ_1 and θ_2 values obtained with 50 iterations for each condition was observed by finding a mean correlation coefficient. This process was performed for 8 different conditions in investigating the relationship level.

Results

For a total of eight conditions, the findings obtained by Spearman *Rho Correlation* Analysis for the relationship level between the ability estimations of individuals for the test with items that exhibited DIF (θ_1) and ability estimations of individuals for the test with no DIF items (θ_2) were first interpreted generally and later the findings that were presented separately were interpreted according to sob problems based on manipulated conditions of number of items, sample size and proportion of items exhibiting DIF.

The following table demonstrates the correlation coefficient values between θ_1 and θ_2 variables for all conditions.

Table 2.

Correlation Coefficients Obtained Under All Conditions

K	Sample size	The proportion of items exhibiting DIF	Number of items	r_s
1	1000	%5	40	0.199
2	1000	%10	40	0.069
3	2000	%5	40	0.058
4	2000	%10	40	-0.143
5	1000	%5	60	0.048
6	1000	%10	60	0.024
7	2000	%5	60	0.172
8	2000	%10	60	-0.027

Note 1: K refers to the condition for each crossing

Note 2: r_s indicates the Spearman rho correlation coefficient value ($p<0.01$).

Figure 1 is presented below in addition to Table 2 since it is thought to be useful to examine the level of the relationship according to all the conditions that were crossed, in a more detailed and clear manner in which all conditions are demonstrated interactively.

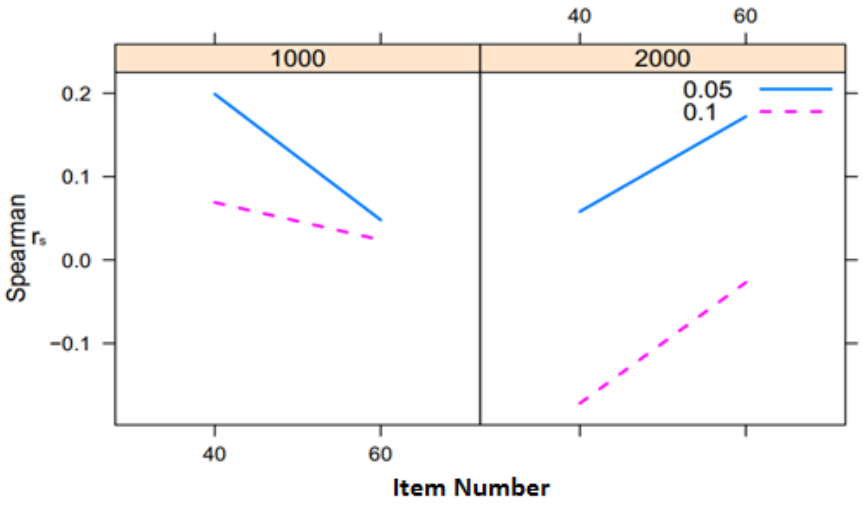

Figure 1. *Correlation coefficients of all conditions*

When Figure 1. was examined in general, it was observed that all r_s values presented in Table 1 changed between 0.2 and 0.15 according to the levels of the researcher's conditions. In other words, it was found that for all the conditions including the conditions where the relationship at the vertical axis was maximum or minimum, each was between either smaller than positive 0.3 or bigger than negative 0.3. These findings also indicate the lack of a general relationship between the predicted ability estimates, regardless of any of the determined conditions. Specifically, when the items that exhibited DIF were removed from the test and re-estimates were made, the rankings of the individuals who took the test changed. This situation caused a low level of relationship.

Since the validity of a test will be affected by the presence of items that exhibit DIF, the exactness and soundness of the scores obtained from the test become controversial. That is, it is not desirable to have items that exhibit DIF in a test. In fact, change in individual rankings when individuals are re-scored after eliminating the items that exhibit DIF shows that actual scores are not really obtained. In this case, individuals will need to question the vital decisions taken from the examinations given for ranking or selection purposes in many national and international areas when actual scores are not definitively obtained.

Bakan Kalaycioglu and Kelecioglu (2011), Basusta (2013), Cepni (2011), Demir (2013), Dogan and Ogretmen (2008), Erdem (2015), Gok, Kelecioglu and Dogan (2010), Ogretmen (2006), Yildırim (2017) found that national or international achievement tests include items that exhibit DIF. The presence of DIF items in these tests constitutes a significant threat to the validity of these tests.

The current study found that changes occurred in the achievement ranking of the individuals as a result of the estimations made by the purification of DIF items. Therefore, it is thought that tests in which items that exhibited DIF were identified in literature would give different results in achievement rankings of individuals if new analyses were to be conducted by purification of these items from the test. This argument points to a significant effect of the purification of items with DIF. In this sense, the findings of this study are parallel to studies that pointed to the need to purify tests from items that exhibited DIF for DIF studies (French and Maller, 2007; Holland and Thayer, 1988, Lee and Geisinger, 2016, Zumbo 1999).

When Table 2 and Figure 1 were re-examined based on the number of items, it was seen that Spearman correlation coefficient value (r_s) decreased and got closer to 0 when the number of items was increased from 40 to 60 in two different graphics where the proportion of items exhibiting DIF was 5% and 10% and the sample size was 1000. The fact that the relationship between ability estimations decreased and became closer to 0 as the number of items increased demonstrated no relationship between individuals' rankings in a test with or without items that exhibited DIF and hence, these rankings change to a great extent. Although the level of relationship between ability estimations increased when the number of items was increased from 40 to 60 in two different when proportion of items exhibiting DIF was 5% and 10% and the sample size was 2000, it was still not sufficient to mention the existence of any relationship because the

correlation coefficients were not differentiated to allow the existence of the relationship for both 40 and 60 number of items. The lack of a significant relationship also indicates that purification of the test changed the rankings of individuals in the test.

When Table 2 and Figure 1 were re-examined based on the differentiation status of sample size, it was seen that Spearman correlation coefficient value (r_s) somewhat decreased when the sample size was increased from 1000 to 2000 while the number of items was identified to be 40 in two different graphics where the proportion of items exhibiting DIF was 5% and 10%. However, this decrease was not significant enough to point to a relationship. It is also seen that increasing sample size under this condition in the graphic where the proportion of items exhibiting DIF was 10% generated a decrease in the correlation coefficient but this change was not significant to eliminate the finding in regards to lack of relationship. In other words, there was no significant relationship between ability estimations for the test with or without DIF for conditions such as the same number of items, the same proportion of items exhibiting DIF and different sample sizes (n=1000 and n=2000). The lack of a significant relationship also indicates that purification of the test changed the rankings of individuals in the test.

When Table 2 and Figure 1 were re-examined based on the differentiation status of the proportion of items exhibiting DIF, the existing lack of relationship still decreased when the proportion of items exhibiting DIF was increased from 5% to 10%. In other words, as the proportion of items exhibiting DIF increased, Spearman correlation coefficient value (r_s) got closer to 0. This was due to the increase in the number of items marked with DIF resulting from the increase in the proportion of items exhibiting DIF and the fact that an abundance of number of items that exhibited DIF generated less relationships among variables. As a result, it was found that increasing the proportion of items exhibiting DIF from 5% to 10% under all conditions with respect to the level of relationship did not have a significant effect on the non-correlation between the test forms that included items that exhibited DIF and items that exhibited no DIF. Therefore, it was found that the purification of the test from DIF items under conditions where the proportion of items exhibiting DIF was 5% and 10% caused changes in the rankings of the individuals.

Spearman correlation coefficient value (r_s) somewhat increased when the sample size was increased from 1000 to 2000 in the graphic where the proportion of items exhibiting DIF was the 5%, the number of items was 60. However, this increase was not significant. In other words, there was no significant relationship between ability estimations for the test with or without DIF for conditions such as the same sample size, same number of items and different proportions of items exhibiting DIF (d=%5 and d=%10). The lack of a significant relationship also indicates that purification of the test changed the rankings of individuals in the test.

Discussion, Conclusion and Recommendations

The differentiation of the rankings of the individuals who take the test when the test is purified from items that exhibit DIF can make the validity of the problematic. In fact, while the presence of DIF in the test constitutes a significant threat to the validity of the test, the removal of these items from the test changes individuals' ranking, therefore, it appears that the purification process has a significant effect. In this case, the vital decisions are taken from the examinations given for ranking or selection purposes in many national and international areas become questionable.

The measurement tool used in a test should not provide any advantages to any group taking the test. In some cases, other variables may be mixed with the properties we want to measure. These variables include gender, type of school, socio-economic level, ethnic origin, etc. (Atalay Kabasakal, 2014). The construct that the test wants to assess and the effect of unrelated variables on the test scores generate a threat on validity and lead to the bias of test scores (Camilli and Shephard, 1994). The first step in determining bias is the DIF analyses developed for this purpose with a large number of methods.

Undesirable results can be obtained if items in any test exhibit DIF, even partially. One of these undesirable results is the fact that DIF directly affects parameter estimation (Han, 2008). Another unintended consequence is the incorrect estimation of ability parameters (Atalay Kabasakal, 2014). As a result of the erroneous estimation of item and ability parameters, the results of many statistical studies based on these parameters become suspect. The literature presents studies investigating the negative effects of the presence of items that exhibit DIF in tests on statistical processes. Some of these studies examined the effects items that exhibit DIF on test equalization process (Atalay Kabasakal, 2014; Chu, 2002; Chu and Kamata, 2005; Turhan, 2006) and their effect on computer-adapted tests (Miller, 1992; Zwick, Thayer and Wingersky, 1995; Zwick, 2000).

According to the results obtained from studies, incorporation of items that exhibit DIF in a test can affect item and ability parameters directly and the statistical studies performed with these parameters indirectly. In this study, it is concluded that there was no relation between the ability estimations predicted with the help of a test form with or without items that exhibited DIF, in other words, the achievement ranking of individuals changed when the test was purified. Thus, the study presented the importance of purification of a test from the items that exhibit DIF in a practical manner

Recommendations

This research was conducted as a simulation study. Considering that use of simulation studies on DIF with real data can help obtain more reliable results, a similar study can be performed with a simulation study supported by real data. In this study, manipulated variables included the number of items, sample size and proportion of items exhibiting DIF. A similar study can be conducted by manipulating different variables (such as reference-focus group ratio).

This study investigated the effect of purification of tests from items that exhibited DIF on the estimation of the ability parameters. A similar study can be performed via item parameters estimation. The results of the study demonstrated that the purification of tests from items that exhibited DIF changed the rankings of individuals. According to this result, it can be suggested that the practitioners should first detect the item that exhibit DIF in a test and recover the test results by purification according to appropriate conditions.

References

Atalay Kabasakal, K. (2014*). The effect of differential item functioning on test equating* (Unpublished doctoral dissertation). Hacettepe University, Ankara.Turkey

Atar, B. & Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. *Hacettepe University Journal of Education, 41*, 36-47.

Bakan Kalaycıoğlu, D. & Kelecioğlu, H. (2011). Item Bias Analysis of the University Entrance Examination *Education and Science, 36* (161), 3-12.

Basusta, N. B. (2013). *An investigation of item bias in PISA 2006 Science Test in terms of the language and culture* (Unpublished mastery dissertation). Hacettepe University, Ankara.Turkey

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

Cepni, Z. (2011). *Differential item functioning analysis using SIBTEST, Mantel Haenszel, logistic regression and item Response Theory Methods* (Unpublished doctoral dissertation). Hacettepe University, Ankara.Turkey

Chu, K. L. (2002). *Equivalent group test equating with the presence of differential item functioning* (Unpublished doctorate dissertation). The Florida State University.

Chu, K. L., ve Kamata, A. (2005). Test equating in the presence of dif items. Journal of Applied Measurement. *Special Issue: The Multilevel Measurement Model, 6* (3), 342-354.

Clauser, E. B., Mazor, K., ve Hambleton, K. R. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*(4), 269–279.

Clauser, B. & Mazor, K. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issue and Practice, 17*, 31-44.

Corey, D. M., Dunlap P. W. ve Burke, M. J. (1998). Averaging Correlations: Expected values and bias in combined Pearson rs and Fisher's z transformations. *The Journal of General Psychology, 125*(3), 245-261, doi: 10.1080/00221309809595548

Demir, S. (2013). *An analysis of the differential item function for the items available in the PISA 2009 mathematics literacy sub-test through Mantel-Haenszel, SIBTEST and logistic regression methods* (Unpublished mastery dissertation). Abant İzzet Baysal University, Bolu.Turkey

Dogan, N. & Ogretmen, T. (2008). The Comparison of Mantel – Haenszel, Chi-Square and Logistic Regression Techniques For Identifying Differential Item *Education and Science,, 33*, 100-112.

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

Erdem, B. (2015). *Investigation of common exams used in transition to high schools in terms of differential item functioning regarding booklet types with different methods* (Unpublished mastery dissertation). Hacettepe University, Ankara.Turkey

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.

Gok, B., Kelecioglu, H. & Dogan N. (2010). *The Comparison of Mantel-Haenszel and Logistic Regression Techniques in Determining the Differential Item Functioning∗ Education and Science, 35*(156).

Hambleton, R. K., Swaminathan, H. ve Rogers, H. J. (1991). *Fundamentals of item response theory.* USA, California: Sage.

Han, K. T. (2008). *Impact of item parameter drift on test equating and proficiency estimates* (Unpublished Doctorate Dissertation). University of Massachusetts, Amherst.

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

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

Kelecioglu, H. & Gocer Sakin, S. (2014). Validity from Past to Present. *Journal of measurement and Evaluation in Education and Psychology. 5*(2), 1-11.

Lee, H. & Geisinger, K. F. (2016). The matching criterion purification for differential item functioning analyses in a large-scale assessment. *Educational and Psychological Measurement*, *76*(1), 141-163.

Miller, T. R. (1992). *Practical considerations for conducting studies of differential item functioning in a CAT environment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show non uniform DIF. *Applied Psychological Measurement, 20*(3), 257–274. doi: https://doi.org/10.1177%2F014662169602000306

Ogretmen, T. (2006). *The investigation of psychometric properties of the test of progress in international reading literacy (PIRLS) 2001: The model of Turkey-United States of America* (Unpublished doctorate dissertation). Hacettepe University, Ankara.Turkey

Turhan, A. (2006). *Multilevel 2PL item response model vertical equating with the presence of differential item functioning*. Unpublished doctorate dissertation, The Florida State University.

Turgut, M. F. & Baykul, Y. (2015). *Measurement and Evaluation in Education* (7. Baskı). Ankara: Pegem Akademi.

Yildırim, A. (2017). *Investigation of differential item functioning of the items in PISA 2009 reading literacy test through univariate and multivariate matching dif* (Unpublished dostoral dissertation). Ankara Üniversity, Ankara, Turkey

Yurdugul, H. (2003). *The Investigation of the student selection and placement examination for secondary education in terms of item bias* (Unpublished dostoral dissertation). Hacettepe University, Ankara.Turkey

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistics regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement*, *32*, 341–363.

Zwick R. (2000). *The assessment of differential item functioning in comput adaptive tests*. In van der Linden W. J., Glas G.A. (eds) Computerized Adaptive Testing: Theory and Practice. Springer, Dordrecht.

Bir Testin DMF'li Madde İçeren ve DMF'li Maddeden Arındırılmış Formlarının Karşılaştırılması

Atıf:

Tulek, O.K., & Kose, I.A. (2019). Comparison of different forms of a test with or without ıtems that exhibit DIF. *Eurasian Journal of Educational Research*, 83, 167-182, DOI: 10.14689/ejer.2019.83.8

Özet

***Problem Durumu*:** Bir ölçme aracında bulunması gereken yapısal niteliklerden en önemlisi olarak kabul edilen geçerlik, klasik anlamıyla bir ölçme aracının ölçmek istediği özelliği başka özelliklerle karıştırmadan ölçebilmesi olarak açıklanabilir. Ancak bir testten elde edilen puanların test ile ölçülmek istenen özellik dışında farklı değişkenlerden de etkilenmesi her ne kadar istenmeyen bir durum da olsa pratikte bu durum kaçınılmazdır. Testi alan bireylerin bulunduğu alt grupların da bu değişkenlerden ne derece etkilendiği önemlidir. Değişkenlerin alt grupları farklı biçimlerde etkilemesi ise madde yanlılığına sebep olabilmektedir. Yanlılığının ilk koşulu olan Değişen Madde Fonksiyonunun (DMF'nin) bir maddede bulunması o maddenin, maddeyi yanıtlayan farklı alt gruplardan herhangi birine ya da birkaçına avantaj sağlamasına neden olmaktadır. Bir testin madde ya da maddelerinde DMF'nin bulunabilme ihtimali özellikle sonuçlarına bakarak bireyler hakkında çeşitli kararların alındığı geniş ölçekli sınavlar için ayrıca dikkat edilmesini zorunlu hâle getirmiştir. Öyle ki eğitimin birçok alanında, sıralama ya da seçme amaçlı uygulanan sınavlarda alınan kararlar bireyler için hayati olabilmekte ve bu sınavların niteliği alınan kararların doğruluğuna, isabetli ve yerinde olmasına direkt olarak etki etmektedir. Peki bahsi geçen yanlı maddelerin testten arındırılması bireyler hakkında verilen hayati kararları değiştirmekte midir? Yanlılık üzerine yapılan birçok çalışmada, SBS, TEOG, ÖSS, PISA, ALES, KPSS gibi geniş ölçekli sınavlarda DMF içeren maddeler tespit edilmiştir Ancak geniş ölçekli bu sınavlarda DMF içeren maddelerin testten çıkarılmasının sonuçlar üzerinde nasıl bir etki oluşturduğuna dair; başka bir ifadeyle DMF'li maddelerin testten çıkarılmasıyla yeniden belirlenen sonuçlara göre bireylerin sınavdaki başarı sıralamalarının etkilenip etkilenmediğine dair çalışmalar sınırlı sayıdadır.

Araştırmanın Amacı: Bireyler hakkında hayati kararların alındığı sınavlarda belirli bir gruba avantaj sağlayan maddelerin testte bulunmasının bireyler arasında eşitsizliğe ve adaletsizliğe neden olabileceği düşünülmektedir. Bu nedenle bu maddelerin testten arındırılması gerekli olabilmektedir. Bu düşünceyle gerçekleştirilen araştırmanın amacı bir testin DMF'li madde içeren ve DMF'li maddeden arındırılmış formlarından kestirilen yetenek kestirimlerinin farklı madde sayısı, farklı örneklem büyüklüğü ve farklı DMF oranı koşulları altında karşılaştırmaktır.

Araştırmanın Yöntemi: Araştırma kapsamında araştırmacı tarafından R 3.4.1 paket programı kullanılarak manipüle edilen değişkenlere göre farklı koşullar altında simülasyon çalışmasıyla veri üretimi gerçekleştirilmiştir. Manipüle edilen değişkenler düzeylerine göre örneklem büyüklüğü (n=1000 ve n=2000), madde sayısı (k=40 ve k=60) ve DMF oranı (d=%5 ve d=%10) olarak belirlenmiştir. Değişkenlerin çaprazlanması sonucunda sekiz koşulun her birine uygun olacak şekilde DMF'li madde içeren veriler üretilmiştir. Çeşitli düzeylerde DMF'li maddeler içerecek şekilde verilerinin üretildiği bir testin öncelikle DMF'li maddeler içeriyorken yetenek kestirimleri gerçekleştirilmiştir. Testin DMF'li maddeler içeren hâliyle kestirilen yetenek kestirimlerine θ_1 ismi verilerek veriler saklı tutulmuştur. Ardından bu testte yer alan DMF'li maddeler testten arındırılarak aynı şekilde yetenekler kestirilmiştir. Testin DMF'li maddeler içermeyen hâliyle kestirilen yetenek kestirimleri ise θ_2 şeklinde saklanmıştır. Son olarak da aynı testin θ_1 ve θ_2 adıyla elde edilmiş olan bu kestirimleri arasındaki ilişkiye bakılmıştır. Bu yetenek kestirimleri ilişkisine göre bireylerin sıralamalarının farklılaşıp farklılaşmadığını tespit etmek amaçlandığı için spearman sıra farkları korelasyon analizi uygulanmıştır.

Araştırmanın Bulguları: Yöntem bölümünde özetlenen bir testin DMF'li madde içeren ve DMF'li maddeden arındırılmış formlarından kestirilen yetenek kestirimlerini (θ_1 ve θ_2) arasındaki ilişki düzeyine bakmak için gerçekleştirilen spearman sıra farkları korelasyon analizi sonucunda elde edilen katsayıların 0'a yakın olmasından dolayı yetenek kestirimleri arasında pozitif ya da negatif yönlü bir ilişki görülmemiştir. Yetenek kestirimleri arasında ilişki görülmemesi ise bireylerin test sonuçlarındaki sıralamalarının değiştiğini işaret etmektedir. Başka bir ifadeyle test DMF'li maddeden arındırıldıktan sonra bireylerin testteki sıralamaları, bir önceki DMF'li madde içeren test formu sıralamalarına göre farklılaşmıştır. Bu tespit, çeşitli koşulların araştırıldığı tüm alt problemlerde benzer şekilde olmuştur. Başka bir ifadeyle madde sayısının 40 ve 60, örneklem büyüklüğünün 1000 ve 2000, DMF oranın %5 ve %10 olarak çaprazlandığı 8 farklı koşulda da testin DMF'li maddeden arındırılmasının bireylerin sıralamalarını değiştirdiğini belirlenmiştir.

Araştırmanın Sonuçları ve Öneriler: Bu çalışma ile bir testin DMF'li madde içeren ve DMF'li maddeden arındırılmış formlarından kestirilen yetenek kestirimleri arasında ilişki bulunmadığı, başka bir ifadeyle DMF'li maddelerin testten çıkarılmasıyla bireylerin başarı sıralamalarının değiştiği sonucuna ulaşılmıştır. Bir testin DMF'li maddelerden arındırılmasıyla testi alan bireylerin sıralamalarının farklılaşması o testin geçerliğini yani özelliğe sahip olanla olmayanı ayırt etme derecesini problemli hâle getirebilecektir. Öyle ki testte DMF'li madde bulunması testin geçerliğine önemli bir tehdit oluştururken bu maddelerin testten çıkarılmasıyla bireylerin sıralamaları değişiyorsa, yapılan arındırma işleminin önemli bir etkisinin olduğu görülmektedir. Bu durum, gerek ulusal gerekse de uluslar arası düzeyde bireyler hakkında hayati kararların alındığı, sonuçlarına bakılarak seçme ve yerleştirme işlemlerinin gerçekleştirildiği sınavların bireyler arasındaki farklılıkları ölçme derecelerinin sorgulanabilir olduğunu gösterebilmektedir.

Anahtar Kelimeler: Yanlılık, değişen madde fonksiyonu, yetenek kestirimi, arındırma.