Eurasian Journal of Educational Research 78 (2018) 67-86

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

Investigation of Group Invariance in Test Equating Under Different Simulation Conditions*

Hatice INAL[1], Duygu ANIL[2]

ARTICLE INFO

Article History:
Received: 1 Oct. 2018
Received in revised form: 23 Oct. 2018
Accepted: 8 Nov. 2018
DOI: 10.14689/ejer.2018.78.4

Keywords
test equating, group invariance, differential item functioning, simulation study

ABSTRACT

Purpose: This study aimed to examine the impact of differential item functioning in anchor items on the group invariance in test equating for different sample sizes. Within this scope, the factors chosen to investigate the group invariance in test equating were sample size, frequency of sample size of subgroups, differential form of differential item functioning (DIF), frequency of items in the anchor test with differential item functioning, directionality of differential item functioning and mean differences in subpopulation ability levels.

Research Methods: The current study was conducted by using item response theory true score equating under equivalent groups anchor test design. REMSD index was used for investigating group invariance in test equating. This study was designed as a comparison of equating results on 96 simulation conditions. The R language and SPSS software was utilized for analysis and 100 replications were performed for each condition. The effect of the conditions held in the study on group invariance in test equating was evaluated by taking average of REMSD. Also, ANOVA was performed to determine significant effect of each factor on group invariance in test equating. **Findings:** The findings of the study showed that differential form DIF was the factor that had the most prominent impact on group invariance in test equating. **Implications for Research and Practice:** Within the scope of the results of the study, group invariance affected by factors of DIF were only in instances in which DIF in anchor items was differential across test forms.

*This study is based on a brief summary of the doctoral dissertation entitled "Investigation of Group Invariance in Test Equating Under Different Simulation Conditions" prepared in the Educational Measurement and Evaluation Program, Hacettepe University, Turkey.

[1] Mehmet Akif Ersoy University, TURKEY, e-mail: haticeinall@hotmail.com, ORCID: https://orcid.org/0000-0002-2813-0873

[2] Hacettepe University, TURKEY, e-mail: aduygu@hacettepe.edu.tr, ORCID: https://orcid.org/0000-0002-1745-4071

Introduction

The scores obtained from tests are among significant sources of information for important decisions made in many subjects. In fact, utilization of tests for the purposes of selection and placement of students, as well as for the determination of occupational life of individuals raises the importance of the decisions to be made according to the results of tests. Different forms of large-scale and centralized tests, which are administered at certain intervals with the aim of placing individuals in a school or work, are developed and implemented for some reasons such as safety. Although the administered forms are aimed at the same purpose, it is necessary to equate the test forms to obtain comparable scores. In psychometry, equating is referred to as the process of converting points so that scores obtained from different forms of a test can be comparable and interchangeable (Dorans and Holland, 2000; Holland, 2007; Kolen and Brennan, 2004). According to Kolen and Brennan (2004), prior to 1980, the issue of equating was ignored by most researchers carrying out assessment and evaluation studies, except for psychometrics responsible for equating. However, the importance of equating began to be understood in the early 1980s. The increase attached to the importance of equating led to a rise in the number and diversity of programs employing different forms of a test, and test specialists in charge of these programs employed in equating scores in different forms (Kolen and Brennan, 2004). Thus, issues of accountability in education and fairness in examination have also begun to be taken into consideration. These developments have led to a further increase in the importance of equating among assessment experts and test takers.

The accuracy of test equating process is of critical significance for testing practices to be performed in a fairer manner, and for making the right decisions about the future of the individuals. Angoff (1984) reported that following a successful equating process, it is possible to monitor individual development of test takers on different applications of a test, identify changes in a group performance at a given time interval, and compare students taking a test at different times of the year. A successful equating means that individuals who take the easier form of a test do not gain an unfair advantage over individuals taking the more difficult form of the test, and that the difference in the scores of individuals who take different forms of a test results from the difference in the achievement level of individuals. For the equating process to be carried out successfully, put another way, for the scores obtained from different forms to be used interchangeably by means of the equating process, several conditions, including the condition of group invariance, must be met. (Dorans and Holland, 2000; Kolen and Brennan, 2004). The group invariance of equating is achieved by the fact that the function used to equate the scores on different scales is not dependent on subgroups (Angoff, 1971; Dorans and Holland, 2000; Flanagan, 1951). Thus, it can be said that the equating function, each form of which is placed on a common scale, must be the same in different groups or subgroups so that the scores obtained from different forms can be used interchangeably (Kolen, 2004). Violation of group invariance in the equating causes the individuals belonging to the different groups who must have the same score to receive different equated scores. In other words, the violation of group invariance

in test equating compromises the principle that a test should be comparable and fair for different groups (Dorans, 2004, 2008; Huggins and Penfield, 2012).

Based on the decisions to be made from the tests, the test scores should provide the most accurate information possible (Kolen and Brennan, 2004). Therefore, no matter what the test scores are used for, the accuracy of the decisions made based on a test score should be demonstrated (Kane, 2013; Messick, 1995; Zumbo, 1999). For this purpose, all items in a test must measure the construct aimed to be measured by a test in a similar way. To this end, Differential Item Functioning (DIF) analyses are performed (Kane, 2013). DIF is used to examine how the performance pertaining to an item in a test shows variation across different subgroups. Therefore, it is necessary to carry out DIF analysis prior to parameter estimation or the test equating to identify or delete items showing DIF, if necessary. On the other hand, it is generally assumed in equating studies that all items are cleared from DIF effects. However, in some circumstances, a well-structured test may contain many items showing DIF, but the deletion of the items showing DIF from the test will result in a decline in the construct validity of the test, and an increase in error in the ability parameter estimations. In addition, there may be circumstances in which the test is invalid in case items showing DIF are deleted. Therefore, it is important to examine conditions that may minimize the effect on equating in the presence of items showing DIF during equating.

Angoff and Cook (1988) stated that DIF and test equating could not be considered as two separate subjects. Hence, studies investigating the effect of items showing DIF on equating have been carried out in the literature (Atalay Kabasakal, 2014; Chu, 2002; Chu and Kamata, 2005; Demirus, 2015; Han, 2008; Huggins, 2012). Items showing DIF not only increase test equating error but may also cause bias against some test takers. Therefore, under ideal conditions, the presence of items showing DIF should be investigated before test equating is performed and if any, the equating process should be initiated upon deleting items showing DIF. However, in studies in which DIF and test equating are carried out concurrently, it is also aimed to examine to what extent and in which conditions DIF influences test equating. When the relevant literature is reviewed, it is seen that the effect of the variables such as the form containing items showing DIF, the ratio of the items showing DIF, DIF effect size, the sample size, the ratio of the sub-group sizes, and the test length on test equating was examined. In this study, it was also investigated how the presence of DIF in the test items posed a problem in terms of group invariance along with equating. Different indices were utilized for examining equating in terms of group invariance. In this study, the group invariance of the equating carried out under the simulation conditions specified was determined by REMSD (Root Expected Mean Square Difference) index. It was also investigated how the group invariance indices changed under conditions in a way that the subgroup sizes were equal, and the ratio of subgroup sizes was 1: 2 in the presence of items showing DIF. Besides, the effect of the sample size on equating results was investigated by various studies in the literature (Hanson and Beguin, 2002; Hu et al., 2008; Kolen and Brennan, 2004; Lee and Ban, 2010), and it was found that the small sample size in those studies could cause erroneous equating results. However, in case

that items showing DIF were present in different sample sizes, it was thought that equating results needed to be investigated in terms of group invariance.

DIF analysis is carried out to examine whether a subgroup in an item gains an advantage over the other, but it is also conducted because it is a prerequisite for some measurement techniques (Shepard et al., 1984). For example, the presence of items showing DIF in equating can adversely affect the equating results (Kim and Cohen, 1992; Shepard et al., 1984). Test equating also encompasses the process of placing parameter estimations on the same scale if there are individuals taking tests containing different items. Items showing DIF may increase the number of errors in test equating or parameter estimation and may also cause bias against some individuals. For this reason, it is necessary to carry out the DIF analysis to identify and, if necessary, delete the items showing DIF prior to the parameter estimation or test equating. However, in some circumstances, a well-structured test may contain a large number of items showing DIF, but the deletion of items showing DIF in the test will result in a decline in the validity of the test and increase the error in the ability parameter estimations. In addition, deletion of items showing DIF will increase the cost of test development studies. For this reason, it is important to identify the items showing DIF during equating and develop and utilize the methods that will minimize the effect of these items on equating (Hidalgo-Montesinos and Lopez-Pina, 2002).

The aim of this study was to determine group invariance indices in various conditions generated from DIF and sample size, and to compare the results obtained in equating carried out based on common items in case that common items showed DIF. In line with the purpose of the study, it was aimed to determine the optimum condition in terms of equating under DIF and sample size factors in equating by means of common items. It was thought that the results of this study would be useful for the researchers and the assessment and evaluation centers implementing large-scale testing applications for designing the appropriate equating process to equate different test forms. It could also be said that this study would contribute to the theoretical research aimed at determining the necessary conditions in order to obtain the most accurate equating results.

Method

Research Design

In this study, it was aimed to compare different equating designs by using simulation data generated according to the specified conditions. Thus, it is fair to say that this study, in which the optimum conditions for equating design were investigated under the specified conditions, was a simulation study.

Equating Design

In this study, "the common item equivalent groups" deign was used to equate the two test forms. In this study, two test forms called F0 and F1 were placed in the forms of common items and the data were simulated. Common items were used to establish the equating relationship between F0 and F1. Hambleton et al. (1991) state that the number of items required for common items should correspond to approximately 20-

25% of the number of items in the test. Therefore, each form to be equated consists of binary scored 40 items, 25% or 10 of which are common items.

Simulation Conditions

The aim of this simulation study was to evaluate the effectiveness of equating designs under the conditions specified in the study. Simulation factors used in the study included the sample size, sample size ratio, the form containing the items showing DIF, the ratio of the items showing DIF, the direction of DIF and the mean ability difference between the groups. Table 1 illustrates the simulation conditions specified.

Table 1

Simulated Conditions of the Study

Factors	**Number of Conditions**		**Level**
Sample Size	2	Level I: Level II:	1500 3000
Ratio of Sample Sizes	2	Level I: Level II:	1:1 1:2
Differential Form	2	Level I: Level II:	2 Form 1 Form
Ratio of Items in the Anchor Test with DIF	3	Level I: Level II: Level III:	%20 %40 %60
Directionality of DIF	2	Level I: Level II:	Unidirectional Bidirectional
Group Mean Ability Difference	2	Level I: Level II:	0 1 SD

Sample Size and Ratio of Sample Sizes: It has been revealed by several studies that sample size has a significant impact on the implementation and interpretation of equating designs (Cui and Kolen, 2008; Hanson and Beguin, 2002; Hu et al., 2008; Kolen and Brennan, 2004; Lee and Ban, 2010, Sahin and Anil, 2017; Zhao, 2008;). In this study, equating was conducted in the total group, the focus group, and the reference group. Therefore, the focus group sample size was taken as at least 500, the smallest sample was taken as 750 in the reference group, and the smallest sample size was taken as 1500 in the total group in order to ensure that equating was conducted in accordance with the literature. A large sample size was taken as 1500 in the focus group, as 2000 in the reference group, and as 3000 in the total group.

On the other hand, it is evident that the sample size of the focus group is smaller in real test applications. For this reason, two sample size ratios were determined as 1: 1 and 1: 2 among focus and reference groups. Thus, four conditions were created for the sample as 500: 1000 and 750: 750 for the small sample (N = 1500) and as 1000: 2000 and 1500: 1500 for the large sample (N = 3000).

DIF: There are past studies which investigated the effect of DIF on test equating (Chu and Kamata, 2005; Tong and Um, 2007). In this study, the ratio items showing DIF, the direction of DIF, and the effect of mean ability difference between focus and

reference group on group invariance were investigated. All of the items showing DIF that were discussed in this study were included in the common items. Furthermore, all of the DIFs were simulated in line with the study uniform, resulting from the differentiation of the b parameter. In the past studies carried out by means of simulated data, it is seen that different values of DIF amount are used (Atar, 2007; Kristjansson, 2001; Wang and Su, 2004;). In this study, 0.90 value, which is a high level of DIF, was selected (Dorans and Holland, 1993).

When two test forms to be equated are applied to different groups at different times, one common item shows DIF in one of the forms, whereas DIF may not be seen in the same common item in the other form. Accordingly, in this study, group invariance in test equating was examined in case when DIF was seen in the common items in both forms, or only in common items in one form.

Due to the presence of more than one item showing DIF in the real test conditions, two, four and six of the 10 common items in the study were simulated in a way to include DIF. Therefore, ratio of the items showing DIF were established as 20%, 40%, and 60%.

DIF is carried out in two ways in a test. Firstly, if all items showing DIF in a test operate in a way to create an advantage for one subgroup only, this points to a unidirectional DIF. More specifically, DIF is unidirectional if one of the groups gains an advantage over the other in all items. DIF is bidirectional if some of the items showing DIF in a test create an advantage for a subgroup, while the other items create an advantage for the other sub-group (Bolt and Stout, 1996).

In this study, it was aimed to investigate how the group invariance of equating changed under the unidirectional and bi-directional conditions of DIF. Unidirectional DIF will always be generated in a way that the reference group is advantageous.

Group Mean Ability Difference: In the literature, while .5 standard deviation between the mean ability distributions of the groups is not found to be significant, the difference of .75 standard deviation is considered moderate, and the difference of 1 standard deviation is considered as an indicator of a significant differentiation. Differences of .5, .75 and 1 standard deviations between the mean ability levels of the compared groups are reported as values commonly found in real test results (Tian, 1999).

In this study, lack of mean ability difference in groups was taken as a condition and 1 SD difference in groups was taken as a secondary condition. Under the first condition, the ability distributions of the reference (R) and focus (F) groups were generated in a way to show the unit normal distribution [R ~ N (0,1) and F ~ N (0,1)]. For the other condition, the ability distribution of the reference group had the unit normal distribution [R ~ N (0,1)], while the ability distribution of the focus group was generated as F ~ N (-1,1).

Data Generation

R program was used to generate data. In this study, each data set was derived 100 times so that the results could be consistent and generalizable. Within the scope of the study, it was necessary to define DIF in some common test items. Therefore, the groups in which F0 and F1 forms were applied were divided into focus groups and reference groups. In addition, in order to calculate the group invariance, it was necessary to carry out equating in the focus, reference and total groups. Therefore, in the data generation phase, data generation was carried out in the focus group, the reference group and the total group for F0 form and the F1 form. The data generation process consisted of three stages which included the generation of item parameters, the generation of ability parameters, and the generation of data sets that contained item responses for both forms to be equated.

Data Analysis

The R program and SPSS program were used to analyze the data. In the first stage of data analysis, the equating process was carried out. Then, in order to evaluate the accuracy of the equating based on the results obtained from the equating process, group invariance indices were obtained. Finally, variance analysis of the group invariance indices obtained was performed according to the factors discussed in the study.

Equating Process

In line with the aim of the study, the equating process of F0 and F1 forms was conducted separately on the total groups, focus groups and reference groups. The equating process was carried out in three stages for each group. Firstly, item calibration was performed with BILOG codes which were batched in R. Then, the parameters pertaining to the forms to be equated were scaled on the same metric with the average sigma method. Following the scaling, Item Response Theory (IRT) true score equating was performed and the equated true scores of F1 form were obtained. The equating process was carried out 100 times for 96 conditions. After each equating process was completed; the equating table containing the equated score of each score obtained as a result of total group equating, the equated score obtained by the equating of the reference groups, and the equated score obtained by the equating of the focus groups were created.

Item parameters can be simultaneously or separately calibrated depending on whether the computer program used for the item parameters IRT analysis is performed once or twice (Hanson and Beguin, 2002; Kim and Cohen, 1992; Petersen et al., 1983; Wingersky et al., 1987). In this study, item responses were calibrated separately for each group and for both test forms. Then, A and B equating coefficients should be calculated based on the item parameters of common items in order to place the parameters of both forms on the same scale. Different methods are available for placing the items and ability parameters in different calibrations on a common scale. These are mean-sigma method (Marco, 1977), the mean-mean method (Loyd and Hoover, 1980), and the characteristic curve method (Haebara, 1980; Stocking and Lord,

1983). In this study, the mean- sigma method was used. In the mean-sigma method, the A and B equating coefficients that must be calculated to place the parameters on the same scale are obtained via mean and standard deviation of the b parameter. With this method, the item and ability parameters of both forms were scaled in each group.

After item difficulty parameters were placed on the same scale, IRT true score equating was used to develop a relationship between correct number scores, in other words true scores in the old and new forms. In the IRT true score equating method, the true score associated with the ability level of a test form is equivalent to the true score associated with the ability level of the other form:

$$\tau_X(\theta_i) = \tau_Y(\theta_i)$$

IRT true score equating is completed in a three-stage process. Firstly, a true score τ_X is selected from the X form. Then, the θ_i value corresponding to the selected true score τ_X is determined. Finally, the τ_Y true score corresponding to the θ_i value in the Y form is found. This process is repeated for all the true score values included in the X form (Kolen and Brennan, 2004).

In this study, the accuracy of equating was evaluated in terms of group invariance. Group invariance in equating is the case when the function used to equate the scores on different scales is not dependent on subgroups (Angoff, 1971; Dorans and Holland, 2000; Flanagan, 1951). Failure to achieve equating group invariance results from the fact that respondents in different subgroups with the same raw score have different expected scores on the equated scale. In this case, even if the exams in which the forms have been applied for the same purpose, problems arise while comparing the scores of the students taking the exam. Various methods are employed to determine group invariance in equating. In this study, the REMSD indice introduced by Dorans and Holland to determine the group invariance in test equating was used (2000). REMSD group invariance indice is shown in the following equation in an unstandardized way:

$$REMSD = \sqrt{\sum\nolimits_{x=1} P_x \left\{ \sum\nolimits_k w_k [d_k(x)]^2 \right\}}$$

where x: score level of the test form; k: Subgroup level; d_k (x): The difference between the equated score calculated based on the equating function of the subgroup k at an x score level with the equated score calculated based on the total equating function; wk: The weight that is determined with the help of the ratio of the test-takers with the subgroups for each subgroup (von Davier, Holland and Thayer, 2004; von Davier and Wilson, 2008).

The value found with REMSD stands for the distance between the sub-group equating functions and the total equating function at each x-point level. In a group invariance study, one REMSD is obtained.

In this study, since the number of replications was 100, group invariance indices in each condition were reported by taking the means of the calculations obtained from replications.

In assessing group invariance in equating, DTM (Difference That Matters) criterion taken as half of the raw point unit proposed by Dorans et al. (2003) and Dorans (2004) was used. Although it is not a rule of thumb to evaluate group invariance with the DTM criterion, it may be ignored that the difference between a score equated in the total group and the equated score(s) in the subgroup(s) is less than 0.50, by considering the DTM = 0.50 criterion, and interpretations are made accepting that if it is more than 0.50, it is considered to be significant (Kolen and Brennan, 2014). Thus, in this study, group invariance indices that were below 0.50 were considered as an indicator that group invariance was achieved, and there was not a problem in terms of the group invariance in the equating conducted. Similarly, the fact that group invariance indices were above 0.50 indicated that group invariance was not achieved, and there was a problem in terms of group invariance in the equating conducted.

Results

The aim of this study was to determine group invariance indices in various conditions generated from DIF and sample size, and to compare the results obtained in equating carried out based on common items in case that common items showed DIF. In line with the aim of the study, Table 2 present the means of REMSD group invariance indices obtained as a result of 100 replications pertaining to the conditions in order to determine accuracy, and to evaluate performance of the equating carried out.

When the REMSD group invariance values in Table 2 were analyzed, it was seen that the difference between the equated scores varied between 0.2 and 2.3. In addition, when the common items showing DIF in the table were present in both forms, the REMSD values generally took values smaller than DTM = 0.5, whereas in case that common items showing DIF were present in a single form, it was observed that most of the REMSD indices took values higher than DTM = 0.5. Thus, in general, it can be interpreted that the presence of the common items showing DIF in a single form would pose a problem in terms of group invariance in equating. The condition when the difference between the equated scores was highest in 1500 sample size was the condition when the sample size ratio of the focus and reference groups was 1:1, the unidirectional DIF defined in 60% common items in a single form was seen, and the mean ability difference between the focus and the reference group was 1. When Table 2 was examined, it was observed that difference between the equated scores increased with a rise in the ratio of the items showing DIF. Furthermore, it can be said that in unidirectional DIF cases, group invariance values were increased compared to the bi-directional conditions, and that bidirectional DIF did not constitute a problem for group invariance. It was seen that the group invariance values in conditions when the ability difference between the groups was 1, it was greater than the values in conditions when it was 0. Thus, it can be interpreted that, as the average ability difference between the groups increased, the difference between the group invariance value, in other words difference between the equated scores would increase. According to Table 2, it can be said that the change in the sample size ratio did not make a big difference on group invariance.

Table 2

REMSD Group Invariance Indices According to Study Conditions

				Differential Form			
				2 Form		1 Form	
				Sample Size		Sample Size	
RSS	RID	DD	GMD	1500	3000	1500	3000
1:1	%20	1	0	0,464	0,255	0,948	1,096
			1SD	0,722	0,347	0,885	0,976
		2	0	0,401	0,274	0,474	0,591
			1SD	0,601	0,383	1,498	0,485
	%40	1	0	0,385	0,340	1,616	1,315
			1SD	0,530	0,510	1,761	1,725
		2	0	0,463	0,343	0,398	0,506
			1SD	0,526	0,320	0,848	0,848
	%60	1	0	0,539	0,357	2,132	2,153
			1SD	0,584	0,597	2,207	1,714
		2	0	0,429	0,331	0,626	0,940
			1SD	0,662	0,678	1,386	1,660
1:2	%20	1	0	0,386	0,283	0,683	0,653
			1SD	0,483	0,430	0,870	0,890
		2	0	0,373	0,300	0,382	0,302
			1SD	0,529	0,491	0,602	1,348
	%40	1	0	0,391	0,238	1,246	1,545
			1SD	0,628	0,504	1,530	1,392
		2	0	0,388	0,255	1,447	0,521
			1SD	0,535	0,473	0,665	1,096
	%60	1	0	0,446	0,338	2,044	1,851
			1SD	0,611	0,609	2,204	1,916
		2	0	0,412	0,328	0,622	0,795
			1SD	0,585	0,429	0,824	0,874

RSS: Ratio of Sample sizes, RID: Ratio of Items in the Anchor Test with DIF, DD: Directionality of DIF, GMD: Group Mean Ability Difference

In Table 2, if the common items in both forms to be equated showed DIF, the ratio of the variables formed under DIF, put another way, ratio of the items showing DIF , and the direction of DIF did not cause a large difference in the values of group invariance in the test equating, but if the common items in only one of the two forms

to be equated showed DIF, it was observed that the ratio of items showing DIF and the direction of DIF led to the group invariance indices in which there were greater differences between the conditions in the test equating. Moreover, when the common items in both forms to be equated showed DIF, the biggest difference developed in group invariance when the mean ability difference between the focus and the reference group changed. If only one of the two forms to be equated had common items showing DIF, a remarkable difference was obtained in group invariance with a change in DIF direction compared to other factors.

Table 3 demonstrates variance analysis results of REMSD group invariance values according to all factors discussed in the study.

Table 3

ANOVA Results for REMSD Group Invariance Indices

Factors		Mean	F	η^2
Sample Size	1500	0,833	186,907*	0,019
	3000	0,763		
Ratio of Items in the Anchor Test with DIF	20%	0,606	693,237*	0,127
	40%	0,790		
	60%	0,996		
Directionality of DIF	1	0,965	1411,710*	0,129
	2	0,630		
Differential Form DIF	1 Form	1,148	9955,451*	0,510
	2 Form	0,448		
Ratio of Sample Sizes	1:1	0,830	66,754*	0,007
	1:2	0,766		
Group Mean Ability Difference	0	0,700	1352,987*	0,124
	1 SD	0,895		

(*p<.05)

When Table 3 is examined, it is seen that REMSD differed significantly according to all variables discussed in the study. As the size of the study group increased, REMSD was significantly reduced. In addition, as the ratio of the common items showing DIF got higher, REMSD increased significantly as well. According to the Post-Hoc test conducted to determine in which subgroups REMSD showed significant increases according to the ratio of common items showing DIF, it was found that there was a significant difference in all dual subgroup comparisons in the variable of ratio of common items showing DIF. It was revealed that REMSD got significantly higher values in conditions when unidirectional DIF was present compared to conditions in which bidirectional DIF was present. In addition, smaller REMSD values were calculated when the difference between the groups was 0, compared to when the difference between the groups was 1.

When the effect sizes in Table 3 were examined, according to the classification developed by Cohen (1992) for the effect size, it is fair to say that the effect of the ratio of sub-group sizes on the REMSD was non-significant; the effect of the study group size was small; the effect of ratio of the common items showing DIF, DIF direction and

the mean ability difference between the groups was moderate and finally, the effect of the form containing common items showing DIF was large.

Discussion, Conclusion and Recommendations

In the context of this study, it was investigated how items showing DIF affected the group invariance in test equating in different sample sizes in case common items showed DIF. In line with this scope, the variables used in order to examine the group invariance of the equating were sample size, ratio of the sample size, the form containing items showing DIF, the ratio of the items showing DIF, the direction of DIF, and the mean ability difference between the groups. Based on the variables discussed, the data were generated, and it was examined which methods affected the equating in the negative sense. To this end, the accuracy of equating was examined in terms of group invariance, and the non-standardized REMSD index was used.

Huggins (2012, 2014) conclude that the most important variable affecting the group invariance was that items showing DIF are defined in a single form or in both forms. Similarly, it was also found in the present study that the most important factor affecting the group invariance in the test equating was the form variable which contained common items showing DIF in case that DIF was seen in common items. The case that common items showing DIF were in a single form caused DIF to negatively affect the group invariance in equating compared to the case of presence of items showing DIF in two forms. It was concluded that in the presence of DIF in the common items, if the common items in both forms to be equated showed DIF, the variables formed under DIF, in other words the ratio of the items showing DIF and DIF direction, did not affect the group invariance in the test equating; however, if the common items showed DIF in only of one of two forms to be equated, the variables formed under DIF, in other words, the ratio of the items showing DIF and the DIF direction affected the group invariance in the test equating. This result is consistent with the findings by Huggins (2012). In the case that DIF is seen in the common items, while the most important factor affecting the group invariance is the ability difference in focus and reference groups, if the common items in both forms to be equated show DIF, the most important factor affecting the group invariance is the DIF direction if the common items show DIF in only one of the two forms to be equated. The increase in the mean ability difference between the groups leads to a significant increase in the difference between the equated scores, that is the values that the group invariance obtains. In an equating study conducted by Huang (2010) on different ability groups with real data, it was concluded that the group invariance did not change in different ability groups. We can say that the difference obtained in this study may be caused by the conditions examined in the study.

In case that common items showing DIF were seen in only one form, a higher difference was obtained between the equated scores in the unidirectional DIF condition compared to two-directional DIF conditions. In other words, in case that common items showing DIF were seen in one form, significantly higher group invariance values of the unidirectional DIF were calculated compared to the conditions

in which directional DIF was present. In the study conducted by Huggins (2012), it is seen that the most influential variable affecting the group invariance is DIF direction in case when the common items showing DIF are seen in a single form. The reason why the direction of DIF is the variable that affects the group invariance most might be associated with that bidirectional DIF may cause less DIF at the form level when compared with the unidirectional DIF, as DIF is defined in favour of the focus group in one part of the items, and in favour of the reference group in another part (Drasgow, 1987; Nandakamur, 1993; Penfield and Camilli, 2007).

In case that common items showing DIF were seen in a single form, the higher the group invariance values were obtained as the ratio of the common items showing DIF increased. In other words, the difference between the equated scores increased with a rise in the ratio of the common items showing DIF. This result is consistent with the findings of the studies carried out by Han (2008) and Huggins (2012, 2014).

Study group size is one of the most commonly used variables in test equating studies. In the literature, it has been established with various studies that less erroneous equating results could be achieved with a larger working group size (Cui and Kolen, 2008; Hanson and Beguin, 2002; Zhao, 2008). Similar to these studies that examined the performance of test equating with an equating error, this study presented consistent findings when the equating performance was evaluated in terms of group invariance. As the working group size grew, significantly lower group invariance values were obtained.

In this study, the data were generated and equated according to 3-parameter logistic model. Different studies can be conducted by generating data according to different models. In addition, it is possible to investigate how the group invariance changes in the test equating in case that items showing DIF are present among the common items or in the test to be equated according to different equating patterns. On the other hand, for different sample sizes, different equating methods in different test lengths can be used to investigate the group invariance of test equating. In this study, the effect of common items showing DIF on the group invariance of test equating was investigated by using simulation data. Likewise, it may be advisable to conduct a study on how test equating affects group invariance by means of a real data set or a simulation study along with a real data set in case that DIF is seen in the common items.

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, D.C: American Council on Education

Angoff, W.H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (College Board Report No. 88-2). New York: College Entrance Examination Board.

Atar, B. (2007). *Differential item functioning analyses for mixed response data using irt likelihood-ratio test, logistic regression, and gllamm procedures*. Unpublished doctorate dissertation. The Florida State University.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika*, *23*(1), 67-95.

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., & 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.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Demirus, K.B. (2015). *Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi.* Yayınlanmamış doktora tezi. Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü. Ankara

Dorans, N.J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43-68.

Dorans, N. J., (2008). *Three facets of fairness*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281-306.

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: Lawrence Erlbaum.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three advanced placement program exams. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to advanced placement program examinations* (pp. 79-118), Research Report 03-27. Princeton, NJ: Educational Testing Service.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

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

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24.

Hidalgo Montesinos, M. D., & Lopez Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and Lord statistic. *Educational and Psychological Measurement, 62*(1), 32.

Holland, P.W. (2007). A framework and history for score linking. In N.J. Dorans, M. Pommerich, & P.W. Holland's (Eds.), *Linking and aligning scores and scales* (pp. 5- 30). NY: Springer

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187–220). Westport, CT: Praeger Publishers.

Huang, J. (2010). *Population invariance of linking functions of curriculum-based measures of math problem solving*. Unpublished doctorate thesis. University of Miami, Florida.

Huggins, A.C. (2012). *The effect of differential item functioning on population invariance of item response theory true score equating*. Unpublished doctoral dissertation. University of Miami, Florida.

Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement, 74(*4), 627-658.

Huggins, A.C., & Penfield, R.D. (2012). An instructional NCME module on population invariance in linking and equating. *Educational Measurement: Issues and Practices, 31*, 27-40.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kim, S.H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*(1), 51–66.

Kolen, M., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (Second ed.). New York: Springer.

Kolen, M.J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*, 3-14.

Lee, W., & Ban, J. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education, 23*, 23–48.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137–156.

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in Item Response Theory. *Educational Sciences: Theory and Practice, 17*(1n), 321-335.

Tian, F. (1999). *Detecting differential item functioning in polytomous items*. Unpublished doctoral dissertation. Faculty of Education, University of Ottawa.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating and their relationship to population invariance. *Journal of Educational Measurement, 41*, 15-32.

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of item-response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*(1), 11-26.

Yang, W.L. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*, 33-41.

Yang, W.L., Dorans, N.J., & Tateneni, K. (2003*)*. Sample selection effects on AP multiple-choice score to composite score scaling. In N.J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to advanced placement program examinations* (ETS Research Report No. RR- 03-27) (pp. 57-78). Princeton, NJ: Educational Testing Service.

Test Eşitlemede Grup Değişmezliğinin Farklı Simülasyon Koşulları Altında İncelenmesi

Atıf:

Inal, H., & Anil, D. (2018). Investigation of group invariance in test equating under different simulation conditions. *Eurasian Journal of Educational Research, 78*, 67-86, DOI: 10.14689/ejer.2018.78.4

Özet

Problem Durumu: Psikometride, bir testin farklı formlarından elde edilen puanların karşılaştırılabilmesini ve birbiri yerine kullanılabilmesini sağlayan puanları dönüştürme süreci eşitleme olarak adlandırılmaktadır. Eşitleme işlemin hatasız olması, gerçekleştirilen test uygulamalarının daha adil olması ve bireylerin geleceği ile ilgili doğru kararlar alınmasında kritik önem arz etmektedir. Başarılı bir eşitleme, bir testin daha kolay formuna alan bireylerin testin daha zor formuna giren bireylere göre

haksız bir avantajı olmadığı ve bir testin farklı formlarını alan bireylerin puanlarındaki farklılığın, bireylerin başarı düzeyindeki farklılıktan kaynaklandığı anlamına gelmektedir. Eşitleme sürecinin başarılı bir şekilde yürütülmesi için, başka bir değişle eşitleme süreci yardımı ile farklı formlardan alınan puanların birbiri yerine kullanılabilmesi için eşitlemenin, grup değişmezliği şartının da yer aldığı birtakım şartları karşılanması gerekmektedir. Eşitlemenin grup değişmezliği, farklı ölçekler üzerindeki puanları eşitlemek için kullanılan fonksiyonun alt gruplara bağlı olmamasıyla sağlanır.

Diğer yandan, bir testte yer alan bir maddeye ait performansın farklı alt gruplarda nasıl değiştiğini incelemek için Değişen Madde Fonksiyonu (DMF) kullanılmaktadır. Bundan dolayı parametre kestirimi ya da test eşitleme çalışmasına başlanmadan önce DMF gösteren maddeleri belirlemek ve gerekirse silmek için DMF analizi yürütmek gerekmektedir. Ancak bazı koşullarda iyi yapılandırılmış bir test çok sayıda DMF gösteren madde içerse de testten DMF gösteren maddelerin silinmesi testin yapı geçerliliğinin düşmesine ve yetenek parametre kestirimlerindeki hatanın artmasına neden olur. Ayrıca DMF gösteren maddelerin silinmesi durumunda testin geçersiz olacağı durumlar oluşabilir. Bu nedenle, eşitleme sırasında DMF gösteren maddelerin varlığında da eşitlemeye olan etkisini en aza indirebilecek koşulların incelenmesi önemlidir.

Araştırmanın Amacı: Bu çalışmanın amacı ortak maddelere dayalı olarak yapılan eşitlemelerde, ortak maddelerin DMF göstermesi durumunda; DMF ve örneklem büyüklüğünden oluşturulan çeşitli koşullara göre grup değişmezliği indislerini belirlemek ve elde edilen sonuçları karşılaştırmaktır. Araştırmanın amacı doğrultusunda; ortak maddeler yardımıyla eşitlenmede, DMF ve örneklem büyüklüğü faktörleri altında simülasyon yardımıyla eşitleme açısından en optimum durum belirlenmeye çalışılmıştır. Çalışmada kullanılan simülasyon faktörleri; örneklem büyüklüğü, örneklem büyüklüğü oranı, DMF gösteren maddelerin bulunduğu form, DMF gösteren madde oranı, DMF yönü ve gruplar arası ortalama yetenek farkıdır.

Araştırmanın Yöntemi: Bu araştırmada, belirlenen koşullara göre üretilen simülasyon veriyi kullanarak farklı eşitleme tasarımlarının karşılaştırılması amaçlanmaktadır. Böylece belirlenen koşullar altında eşitleme tasarımı için optimum koşullar incelendiği bu araştırmanın bir simülasyon çalışması niteliği taşımakta olduğu söylenebilir.

Bu araştırmada iki test formunu eşitleyebilmek için "denk gruplarda ortak madde/test deseni" kullanılmıştır. Ortak maddeler, F0 ve F1 arasındaki eşitleme ilişkisini kurmak için kullanılmıştır. Eşitlenecek her bir form %25'si yani 10 tanesi ortak madde olmak üzere ikili puanlanmış 40 maddeden oluşmaktadır.

Verilerin üretilmesinde R programından yararlanılmıştır. Bu çalışmada sonuçların tutarlı ve genellenebilir olabilmesi için her veri seti 100 defa türetilmiştir. Çalışma kapsamında bazı ortak test maddelerinde DMF tanımlamak gerekmektedir. Bu nedenle eşitlenecek olan F0 ve F1 formlarının uygulandığı gruplar, odak grup ve referans grup olarak ikiye ayrılmıştır.

Verilerin analizinde R programı ve SPSS programından yararlanılmıştır. Veri analizinin ilk aşamasında eşitleme süreci yürütülmüştür. Daha sonra eşitleme sürecinden elde edilen sonuçlara dayalı olarak eşitlemenin doğruluğunu değerlendirmek amacıyla eşitlemede grup değişmezliği indisleri elde edilmiştir. Son olarak da elde edilen grup değişmezliği indislerinin çalışmada ele alınan faktörlere göre varyans analizi yapılmıştır.

Çalışmanın amacı doğrultusunda F0 ve F1 formlarını eşitleme süreci toplam gruplar, odak gruplar ve referans gruplar üzerinde ayrı ayrı yürütülmüştür. Eşitleme süreci her bir grup için üç aşamada gerçekleştirilmiştir. İlk olarak R da batch edilen BILOG kodlarıyla madde kalibrasyonu işlemi yapılmıştır. Daha sonra eşitlenecek formlara ait parametreler ortalama sigma yöntemiyle aynı metrik üzerine ölçeklenmiştir. Ölçeklemenin akabinde, MTK gerçek puan eşitlemesi yapılarak F1 formuna ait eşitlenmiş gerçek puanlar elde edilmiştir. Eşitleme süreci çalışma kapsamındaki 96 koşul için 100'er kere gerçekleştirilmiştir. Her bir eşitleme süreci tamamlandıktan sonra; her bir puanın toplam grupların eşitlenmesi sonucunda elde edilen eşitlenmiş puanının, referans grupların eşitlenmesi sonucunda elde edilen eşitlenmiş puanının ve odak grupların eşitlenmesi sonucunda elde edilen eşitlenmiş puanının yer aldığı eşitleme tabloları oluşturulmuştur.

Bu çalışmada eşitlemenin doğruluğunu grup değişmezliği açısından değerlendirilmiştir. Eşitlemede grup değişmezliğin belirlenmesinde çeşitli yöntemler kullanılmaktadır. Bu çalışmada Dorans ve Holland (2000) tarafından test eşitlemede grup değişmezliğini belirlemek amacıyla geliştirilen REMSD indisinden yararlanılmıştır. Tekrar sayısı 100 olduğu için her bir koşulda grup değişmezliği indisleri tekrarlardan elden edilen hesaplamaların ortalamaları alınarak raporlanmıştır.

Eşitlemede grup değişmezliğinin değerlendirilmesinde, Dorans ve diğerlerinin (2003) ve Dorans' ın (2004) önerdiği ham puan biriminin yarısı olarak alınan DTM (Difference That Matters) kriterinden yararlanılmaktadır. DTM= 0.50 kriteri alınarak bir puanın toplam gruptaki bir eşitlenmiş puan ile alt grup(lar)daki eşitlenmiş puan(lar) arasındaki farklılığın 0.50'den daha az olmasının yok sayılabilir; 0.50'den daha fazla olmasının ise manidar olduğu kabul edilerek yorumlar yapılmaktadır.

Araştırmanın Bulguları: REMSD grup değişmezliğinin çalışmada ele alınan tüm değişkenlere göre manidar farklılık gösterdiği görülmektedir. Çalışma grubu büyüklüğü arttıkça REMSD grup değişmezliği indisinin manidar olarak azaldığı görülmektedir. Ayrıca, DMF gösteren ortak madde oranı arttıkça ise manidar şekilde REMSD grup değişmezliği indisi de artmaktadır. REMSD grup değişmezliği indisinin DMF gösteren ortak madde oranına göre gösterdiği manidar farklılığın hangi alt gruplar arasında olduğunu belirlemek için yapılan Post Hoc testine göre ise DMF gösteren ortak madde oranı değişkeninin tüm ikili alt grup karşılaştırmalarında manidar fark olduğu belirlenmiştir. REMSD grup değişmezliği, tek yönlü DMF'nin söz konusu olduğu koşullarda iki yönlü DMF'nin söz konusu olduğu koşullara göre manidar olarak daha büyük değer aldığı görülmektedir. Ayrıca gruplar arası yetenek

farkının 0 olduğu koşullar, gruplar arası yetenek farkının 1 olduğu koşullara göre daha küçük REMSD değerleri hesaplanmıştır.

Araştırmanın Sonuç ve Önerileri: Bu çalışmada veriler 3 parametreli lojistik modele göre üretilerek eşitleme yapılmıştır. Farklı modellere göre veri üretip farklı çalışmalar oluşturulabilir. Ayrıca, farklı eşitleme desenlerinde ortak maddelerde ya da eşitlenecek testte DMF gösteren maddelerin yer alması durumunda da test eşitlemede grup değişmezliğinin nasıl değiştiği incelenebilir. Diğer yandan, farklı örneklem büyüklükleri için farklı test uzunluklarında farklı eşitleme yöntemleri kullanılarak test eşitlemenin grup değişmezliği araştırılabilir. Bu çalışmada DMF gösteren ortak maddelerin test eşitlemenin grup değişmezliğine etkisi simülasyon verisi kullanılarak gerçekleştirilmiştir. Benzer şekilde gerçek bir veri setinde ya da simülasyon çalışması ile birlikte gerçek veri seti kullanılarak ortak maddelerde DMF görülmesi durumunda test eşitlemenin grup değişmezliğini nasıl etkilediği hususunda bir çalışma yapılması önerilebilir.

Anahtar Kelimeler: Test eşitleme, grup değişmezliği, simülasyon çalışması, değişen madde fonksiyonu.