Elementary Education Online, 2017; 16(1): 217-227,
İlköğretim Online, 16(1), 217-227, 2017. [Online]:http://ilkogretim-online.org.tr
doi: http://dx.doi.org/10.17051/io.2017.34481

Investigating Group Invariance of Equating Results

Eşitleme Sonuçlarının Gruptan Bağımsızlığının İncelenmesi

Neşe Öztürk Gübeş, *Mehmet Akif Ersoy University, Faculty of Education,* nozturk@mehmetakif.edu.tr
Hülya Kelecioğlu, *Hacettepe University, Faculty of Education*, hulyaebb@hacettepe.edu.tr

ABSTRACT. In this study, it was investigated that Item Response Theory (IRT) equating results whether or not group invariant. Group invariance of equating functions means that equating is same for everyone in the population. The raw scores which were taken from 9th grade 2009 ÖBBS D form of Social Sciences were equated to 2009 ÖBBS B form of Social Sciences with IRT true-and observed- score equating methods. Equating study was conducted with using equivalent groups design. The subgroups were generated with regard to examinees' self-perceived competence in geography and history lessons. The results indicated that under the both IRT true-score and observed-score equating methods, equating results were group sensitive.
Keywords: Test Equating, Group Invariance, IRT True-Score Equating, IRT Observed-Score Equating

ÖZ. Bu araştırmada, Madde Tepki Kuramı (MTK) gerçek-puan eşitleme ve gözlenen-puan eşitleme sonuçlarının eşitlemenin gruptan bağımsızlık varsayımını sağlayıp sağlamadığı incelenmiştir. Gruptan bağımsızlık, eşitleme fonksiyonun elde edildiği gruba bağlı olmaksızın evrendeki her birey için aynı olması demektir. Bu çalışmada, 2009 yılında yapılan 9. Sınıf Öğrenci Başarılarının Belirlenmesi Sınavı (ÖBBS)'nın Sosyal Bilgiler alt testi D kitapçığından elde edilen puanlar B kitapçığından elde edilen puanlara eşitlenmiştir. Eşitleme çalışması, eşdeğer gruplar deseni kullanılarak yürütülmüştür. Eşitleme sonuçlarının gruptan bağımsızlığı öğrencilerin tarih ve coğrafya derslerindeki başarı algısına göre incelenmiştir. Araştırmanın sonucunda, MTK gerçek-puan ve gözlenen-puan eşitleme sonuçlarının elde edildikleri gruba duyarlı olduğu görülmüştür.
Anahtar Kelimeler: Test Eşitleme, Gruptan Bağımsızlık, MTK Gerçek-Puan Eşitleme, MTK Gözlenen-Puan Eşitleme

INTRODUCTION

For preserving test security and removing negative effects of administering the same test form several times, generally parallel test forms are used in the large-scale assessments. Although all test forms are designed for the same purpose, administering different test forms in each test situation may cause problems of non-equivalent test forms. The scores which are taken from different test forms cannot be accurately compared unless the two test forms are equivalent. Suppose that the two students apply for the same college scholarship that is based on test scores. The two students take the tests on different test dates, and Student 1 earns a higher test score than Student 2. We can explain this difference in two ways: Student 1 is more successful than Student 2 or Student 1 took an easier form than Student 2. In the second case, the difference in scores might be due to differences in the difficulty of test forms rather than in the achievement levels of the students (Kolen & Brennan, 2004). To avoid this problem, we need to equate the different forms of the same test to maintain test score comparability. Test equating is necessary to be fair to students taking different test forms. Braun and Holland (1982) defined test equating as "making numerical adjustment to the scores obtained each form of the test to compensate for the form to form variation in difficulty" (p.10).

An equating study can be conducted with various data collection designs. The major three equating designs are single group design, random or equivalent groups design and common-item nonequivalent groups design (Kolen, 1988; Kolen & Brennan, 2004). In the single group, design the same group takes both test forms to be equated. In the equivalent groups design two randomly selected groups take different forms of the tests. In this design, a spiraling process can be used to randomly assign forms. With spiraling procedure; first examinee receives Form X, the second

Geliş tarihi: 20/04/2016, Kabul tarihi: 26/08/2016, Yayımlanma tarihi: 01/01/2017

examinee Form Y, the third examinee Form X, and so on. This spiraling process leads randomly equivalent groups taking Form X and Form Y. The last design is common-item nonequivalent groups design. This design often is used when more than one form per test date cannot be administered because of test security or other practical concerns. In this design two groups of examinees take different forms of a test; each form contains a common set of items (internal anchor) or a common anchor test (external anchor) is given with the forms (Cook & Eignor, 1991; Kolen, 1988; Kolen & Brennan, 2004). In this study, equivalent groups design was used to collect data for equating.

Besides different equating designs, there are different equating methods. We can classify these methods as traditional equating methods and item response theory (IRT) equating methods (Cook & Eignor, 1991; Kolen & Brennan, 2004). The three traditional equating methods are mean equating, linear equating, and equipercentile equating. In mean equating, the means on the two forms are set equal; the Form Y scores are converted so that their mean will equal the mean of the scores on Form X. In linear equating, the means and standard deviations on the two forms are set equal. In other words, linear equating based on the assumption that, apart from differences in means and standard deviations, the distribution of the scores on Form X and Form Y are the same. The equipercentile equating involves determining which scores on two forms have the same percentile rank. In equipercentile equating, Form Y scores converted using equipercentile equating have approximately the same mean, the same standard deviation, and distributional shape (skewness, kurtosis, etc.) as do scores on Form X (Crocker & Algina, 1986; Kolen, 1988).

As above mentioned, item response theory (IRT) can be used to equate tests. IRT equating can be viewed as a three-step procedure. In the first step, item parameters are estimated with a particular IRT model or models. In the current study, the three-parameter logistic (3PL) model was assumed for items. The second step involves placing item parameter estimates from separate calibration runs on the same scale. For some equating designs (e.g. random groups design) this second step is not necessary. If concurrent calibration run is used, the item parameter estimates for two test forms automatically on the same scale. In the third step, equating is conducted using either the IRT true-score equating or the IRT observed-score equating method (Cook & Eignor, 1991).

Test equating is a statistical process for producing interchangeable scores across test forms. Achieving test score interchangeability requires satisfying equating properties (Kolen & Brennan, 2004). Dorans and Holland (2000) reported five equating requirements that are basic to all of test equating.

i. The equal construct requirement: The tests should measure the same constructs. For example a test of reading can only be equated with another test measuring reading.
ii. The equal reliability requirement: The tests should have the same reliability.
iii. The symmetry requirement: Equating transformations must be symmetric. This requirement implies that if a raw score of 26 on Form X converts to a raw score of 27 on Form Y, then a raw score 27 on Form Y must convert to raw score of 26 on Form X.
iv. The equity requirement: It should be a matter of indifference for an examinee to be tested by either one of two tests that have been equated.
v. Group invariance or population invariance requirement: Group invariance requirement implies that "Equating relationship is the same regardless of the group of examinees used to conduct the equating" (Kolen & Brennan, 1995, p.12). For example, if group invariance requirement holds, the same equating relationship is found for gender groups or geographic region groups.

Group invariance of equating functions means that equating is valid for everyone in the population and it is directly related with the test fairness and equity. Equating functions should not be strongly influenced by the population of examinees on which they are derived. However, equating results cannot be completely group invariant, but it might hold approximately (Dorans & Holland, 2000; von Davier, 2007; Yi, Harris, & Gao, 2008). As Brennan (2008) noted "population invariance is a matter of degree" (p. 102). If group invariance does not hold to a sufficient degree the equating

might be appropriate for the target population as a whole but inappropriate for the some subgroups. Suppose that there are two subgroups of examinees and two test forms to be equated. If one subgroups of examinees has lower scores on one form than the other subgroup and vice versa occurs for the other test form, the lower scoring group always disadvantaged by the use of the total-group equating function. Therefore, using the same equating function for different subpopulations of examinees cannot provide fair equating results (Holland & Dorans, 2006; von Davier, 2007).

Despite the fact that concept of group invariance in equating has been discussed since 1950s (Kolen, 2004 gives more detailed information) but its popularity has grown in recent times because of increasing sensitivity for test fairness and equity (Huginns & Penfield, 2012). Several studies (e.g. Dorans, 2004; Dorans, Holland, Thayer, & Tateneni, 2002; Dorans, Liu &, Hammond, 2008; Liu & Holland, 2008; von Davier & Wilson, 2008; Yang & Gao, 2008) have examined group invariance property of equating results based on racial/ethnic background, gender, geographic region, and other demographic variables to obtain subgroups. These studies found little sensitivity of equating results for subgroups formed on the basis of naturally occurring variables. On the other hand, there are relatively few researches (Cook & Petersen, 1987; Harris & Kolen, 1986; Yi et al., 2008) which have looked at group invariance for subgroups that differed in ability. Except for Harris and Kolen's (1986) study, these studies have shown that if subgroups are constructed using variables that are related to the construct being measured, equating functions may be different for populations and subpopulations. For example, in their study Yi et al. (2008) divided examinees into different subgroups based on various measures of ability (average composite scores for test centers, whether they had taken a physics course, and self-reported science grade point average (GPA)) using a science achievement test. They found that if the subgroups' abilities are related to performance on the science test (e.g., examinees' self-reported GPAs or if examinees had taken a physics course), then equating results more group dependent.

In equating literature, it is assumed that when groups used to equate test forms are similar in ability, the equating functions appears to be population independent but if there are large difference between groups it can cause significant problems. The large difference in mean ability of the equating samples can be reason of failure of the equating properties (Kolen & Brennan, 2004). In this study, our purpose is to examine group invariance of equating functions by dividing groups based on self-perception in geography and history lessons which related to construct being measured and can lead subgroups different in ability. Another aim is to show importance of equating samples to satisfy population invariance which is one of the important property to provide test score interchangeability.

METHOD

Data and Equating Design

The data used in this study are item responses from Öğrenci Başarılarının Belirlenmesi Sınavı (ÖBBS) which is a national assessment that is used for assessing elementary and secondary grade students' achievements in Turkish, mathematics, science, social science, and English domains in Turkey. We used data from the 9th grade ÖBBS, which was administered in 2009. ÖBBS aims to assess secondary school students' achievements in Turkish Literature, mathematics and geometry, science (physics, chemistry, and biology), social science (history and geography) and English. To prevent copying and allow for the sampling wide range of content, four different booklets (A, B, C, and D) are used in ÖBBS. As mentioned, students' achievements in five domains are assessed so one booklet is comprised of five tests. There are 15 questions in each test for a total of 75 questions in one booklet. In these booklets, A-C and B-D are designed parallel in construct, content and difficulty (Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı [EARGED], 2010). While choosing data for this study we considered whether tests held IRT unidimensionality assumption or not. So we conducted dimensionality assessment (principal component analysis) and saw that social science tests were unidimensional. We concluded that they were measuring students' social science achievement.

While administrating ÖBBS, four booklets are alternated when they are packaged. While the booklets are handed out, the first examinee receives Booklet A, the second examinee receives Booklet B, the third examinee receives Booklet C, and the fourth examinee receives Booklet D and so on. Based on this spiraling process, we assume that the examinees that get different booklets belong to randomly equivalent groups. Therefore, this research was conducted with using equivalent groups design. The study group of this research is consisted of 15270 and 15323 9th grade students which had taken Booklet B and Booklet D. Hereafter; we call social science tests in Booklet B and D as "Form B" and "Form D".

Procedure

We completed data analyses in four stages. At the first stage; for each test form, we generated subgroups based on examinees' self-perceived competence in geography and history lessons. To provide an extensive resource for interpreting achievement results and to track changes in curriculum and instructional practices, ÖBBS asks students and their teachers to complete questionnaires. To examine population invariance of equating functions we used information based on variables related to the construct: examinees' self-perceived competence in geography and history lessons. ÖBBS asks students, "How successful do you find yourself in geography lesson?" and "How successful do you find yourself in history lesson?". There are four possible answers: "very successful", "successful", "a little successful", and "not successful". To provide two subgroups, we combined the "very successful" and "successful" categories under "successful category"; "a little successful" and "not successful" categories under "not successful category".

At the second stage; the three-parameter logistic IRT model was used to calibrate item parameters with the BILOG-MG 3.0 (Zmowski, Muraki, Mislevy, & Bock, 2003) computer program. The item parameters were calibrated for each form on the total group and each subgroup respectively.

At the third stage; the IRT true-score-and observed- score equating methods were used to equate test scores. The computer program PIE (Hanson & Zeng, 2004) was used for equating.

At the fourth and last stage; to evaluate group invariance of equating functions root mean square difference (RMSD) and root expected mean square difference (REMSD) indexes were calculated. RMSD and REMSD indexes were introduced by Dorans and Holland (2000). The RMSD which is conditional on score level and defined as

$$RMSD\ (x) = \frac{\sqrt{\sum_j w_j [e_{Pj}(x) - e_P(x)]^2}}{\sigma_{Y_P}}, \quad (1)$$

In equation (1), P denotes the total population, and $\{P_j\}$ denotes a partition of P into mutually exclusive and exhaustive subpopulations, $P_1, P_2, \ldots, P_j$. In the application of this study, P denotes entire test administration, whereas the P_js are defined by examinees' self-perceived competence in geography and history lessons (successful – not successful). Furthermore, in equation (1), w_j is the weight which could be the relative proportion of P_j in P or some other set of weights that sum to unity. In this study, w_j is the relative proportion of P_j in P. The equating function that equates Form X to Form Y, computed for the whole population, is $e_P(x)$, and for the subpopulation P_j this equating function is $e_{P_j}(x)$. Finally, in equation (1), σ_{Y_P} denotes the standard deviation of Y scores in P. The resulting value for the RMSD represents the typical distance between the subpopulation equating functions and the overall equating function at the each score level x.

The $REMSD$ is a weighted average of the $RMSD\ (x)$ values where the weighting at each value of x is the proportion of test takers scoring at x (Huggins & Penfield, 2012). The $REMSD$ is computed using

$$REMSD = \frac{\sqrt{\sum_j w_j E_p\{e_{Pj}(x) - e_P(x)\}^2}}{\sigma_{Y_P}}, \quad (2)$$

where x denotes a score point from the total group P, and E_p denotes averaging over the distribution of X in P.

To evaluate whether the magnitude of RMSD and REMSD is essentially significant we calculated Difference That Matters (DTM; Dorans & Feigenbaum, 1994). DTM is used as a benchmark for RMSD and REMSD statistics; depends on the reporting scale of a particular test program. A difference between equating results larger than a half score unit means a DTM (Brennan, 2008; Dorans, 2004; Yi et al., 2008). In this study, we used the criteria of .5 score points as DTM. Because the RMSD and REMSD are standardized dividing by σ_{Y_P}, we divided DTM by the standard deviation of D Form (3.50) and we obtained Standardized Difference that Matters (SDTM) value, .143. When RMSD or REMSD statistics exceed the SDTM, we consider the differences between equating results is practical significance (Dorans, 2003).

RESULTS

Table 1 shows the sample sizes of the subgroups and total group for the two forms. The sample sizes of subgroups are approximately equal for Form B and D. Successful self-perceived competence in geography lesson group comprise 52.8% of the total group for Form B and 53.3% for Form D. Similarly, successful self-perceived competence in history lesson group comprise 61.3% of the total group for Form B and 60.8% for Form D.

Table 1. *Raw Score Descriptive Statistics of Form B and Form D for the Total, Self-Perceived Competence in Geography and History Lessons*

	Form B				*Form D*			
Group	*N*	*%*	*M*	*SD*	*N*	*%*	*M*	*SD*
Successful in Geography	8066	52.8	8.20	3.51	8168	53.3	8.63	3.48
Unsuccessful in Geography	7204	47.2	7.11	3.39	7155	46.7	7.41	3.46
Successful in History	9363	61.3	8.22	3.48	9309	60.8	8.68	3.47
Unsuccessful in History	5907	38.7	6.85	3.37	6014	39.2	7.10	3.39
Total	15270	100	7.69	3.50	15323	100	8.06	3.52

Note. N= number of examinees; M= mean, SD= standard deviation

The average number-correct scores and standard deviations for subgroups and total groups are also summarized in Table 1. The successful self-perceived competence groups for both geography and history lessons have slightly higher mean scores than unsuccessful self-perceived groups on both forms. The differences of the mean raw scores between two subgroups are based on examinees' self-perceived competence in geography lesson are 1.09 for Form B and 1.22 for Form D. The differences between subgroups based on examinees' self-perceived competence in history lesson are 1.37 for Form B and 1.58 for Form D. Also, the difference between two forms' mean raw score is only 0.37 for the total group. As Yang and Gao (2008) indicated that similarity of average raw scores both for total and subgroups also provided evidence of the groups taking different forms are fairly equivalent.

The consistency of the equating results which are obtained from the subgroups and equating methods can be examined by looking at the percentages of examinees at different score points in the total group (Yi et al., 2008). Table 2 presents the percentages of examinees and equated raw scores at raw score points (5, 8, and 11) under the two equating methods and two grouping variables for Form D. The reason for choosing these three score points is that about 75% of examinee scores were 5 or higher, about 50% scored 8 or higher, and about 25% scored 11 or higher.

Table 2 shows that for Form D under the IRT true-score equating method at a raw score point of 5, the equating results would be the same total and subgroups. As seen in Table 2, although the equivalent score is the same for the total and successful examinees' self-perceived competence in the geography lesson subgroup, the equivalent score obtained from unsuccessful subgroup is 1 score point higher than the total group. Under the IRT observed-score equating method at raw score point of 5, the equated score estimated from the successful group is 1 score point higher than the total group and 2 points higher than the total group estimated from unsuccessful group. About 9% of (1409) examinees would be affected if the equating results obtained from the subgroups used. At a raw score point of 8, under the both IRT true- and observed-score equating methods the equivalent score is the same for the total and all subgroups. But, at the 11 raw score the equivalent score is 1 point lower than obtained from the total group. Similarly, if the equated results obtained from the subgroups, then about 8% (1244) of examinees would be affected.

Table 2. *Percentages of Examinees at Three Raw Score and Equated Raw Score Points Based on Different Equating Methods and Grouping Variables*

		IRT True-Score Equating			IRT Observed-Score Equating		
Raw Score	N (%)	TGES	SGES	UGES	TGES	SGES	UGES
Self-Perceived Competence in Geography Subgroups							
5	1409 (9.2)	4	4	**5**	3	**4**	**5**
8	1412 (9.2)	7	7	7	7	7	7
11	1244 (8.1)	11	**10**	**10**	11	**10**	**10**
Self-Perceived Competence in History Subgroups							
5	1409 (9.2)	4	4	**5**	3	**4**	**5**
8	1412 (9.2)	7	7	7	7	7	7
11	1244 (8.1)	11	**10**	11	11	**10**	11

Note. TGES=total group equated score; SGES= successful group equated score; UGES= unsuccessful group equated score. Bold type indicates that the subgroup equated raw score are different from the total group equated score.

Table 2 also illustrates equating results at raw score 5, 8, and 11 when subgroups were divided in based on examinees' self-perceived competence in history lesson. For both raw scores 5 and 8, a similar pattern is observed as examinees' self-perceived competence in geography. For the raw score of 11, both under the IRT true- and observed-score equating methods, equivalent score is the same for the total and examinees' unsuccessful self-perceived competence in history lesson subgroups, but different for the total and successful subgroups. So, only if the successful subgroup's equating result was used, about 8% (1244) of examinees would be affected.

Conversion differences (total group equated score minus subgroup equated score) between the total and subgroups based on examinees' self-perceived competence in geography and history lessons are plotted in Figure 1.

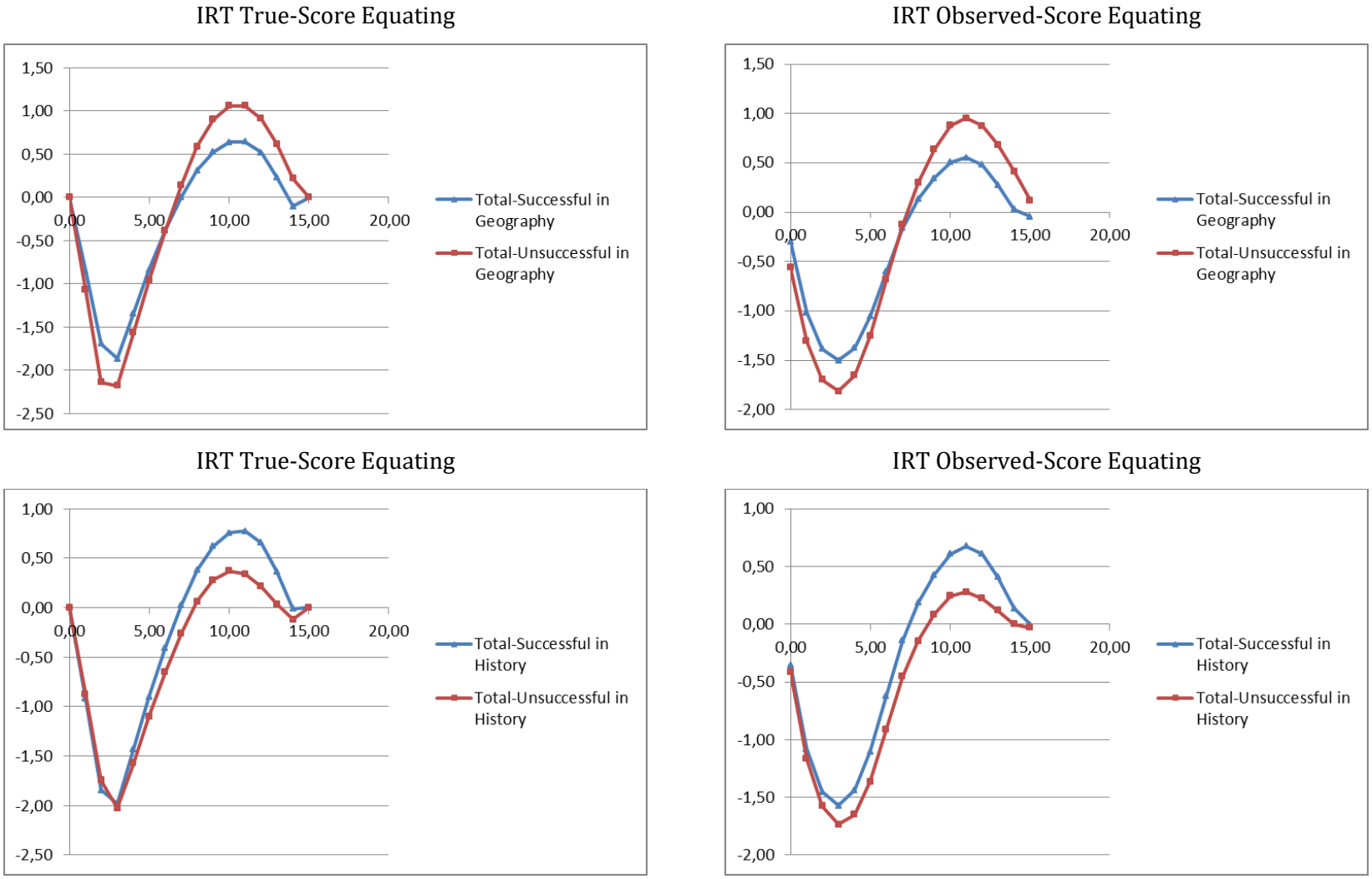

Figure 1. *Conversion Differences when Groups Defined based on Examinees' Self-Perceived Competence in Geography and History Lessons.*

Using different subgroups based on the self-perceived competence in geography and history lessons provided similar equated raw score conversions at the middle score points. Under the both IRT true-score and observed-score equating methods, for instance, at 7 and 8 raw score points the equated score differences between the total and subgroups are approximately zero (see Figure 1). Also, under the both equating methods the conversion differences between the total and subgroups are less than 1 point for the many raw score points between 6 and 15 (see Figure1). But, under the both equating methods at raw scores between 1 and 5 the absolute conversion difference between the total and all subgroups are larger than 1 score point for the many score points (see Figure 1).

Table 3 presents the RMSD at each score points and REMSD for the raw entire score scale. The RMSD and REMSD which are larger than SDTM appears with an asterisk in the Table 3. RMSD statistic describes the difference between the total and the subgroup equating functions across subgroups at each score level, and REMSD is a measure of overall differences between the total and subgroup equating functions across subgroups and score levels. Table 3 shows that for both self-perceived competence in geography and history lessons RMSD values are smaller than SDTM only at middle and extreme points of the raw score scale. Otherwise, RMSD values are larger than SDTM. Also, for both equating methods and subgroups, all of REMSD values are larger than SDTM. This result indicates that IRT equating functions for each subgroup differ in a significant way from the total group. Also, results indicate that for all subgroups IRT true-score equating results have larger REMSD values than IRT observed-score equating.

Table 3. *RMSD and REMSD of Forms B and D across Different Subgroups*

	Self-perceived Competence in Geography Lesson		Self-perceived Competence in History Lesson	
Raw Score	IRT OSE	IRT TSE	IRT OSE	IRT TSE
0	0.125	0.000	0.108	0.000
1	0.332*	0.274*	0.319*	0.258*
2	0.439*	0.547*	0.429*	0.516*
3	0.472*	0.577*	0.468*	0.571*
4	0.431*	0.415*	0.436*	0.424*
5	0.329*	0.256*	0.346*	0.280*
6	0.183*	0.109	0.213*	0.148*
7	0.042	0.028	0.087	0.046
8	0.064	0.132	0.048	0.085
9	0.143	0.207*	0.096	0.147*
10	0.201*	0.246*	0.143	0.181*
11	0.219*	0.247*	0.158*	0.183*
12	0.198*	0.208*	0.141	0.151*
13	0.146*	0.130	0.093	0.082
14	0.081	0.048	0.031	0.022
15	0.026	0.000	0.006	0.000
REMSD	0.256*	0.274*	0.246*	0.258*

Note. OSE= observed-score equating; TSE= true-score equating; RMSD= root mean square difference; REMSD= root expected mean square difference; *RMSD > SDTM

CONCLUSION and DISCUSSION

In this study, the group invariance of equating functions based on construct related subgroups were examined. The results indicated that under the both IRT true-score and observed-score equating methods, equating results were group sensitive. The score conversion derived from subgroups were similar in the middle (7 and 8) and extreme (0, 14, and 15) part of the raw score range, otherwise different. Especially, at the most of lower raw scores (except 0) under the both IRT true-score and observed-score conversion difference between the total and subgroups more than 1 score points (see Figure 1). As Yi et al. (2008) noted that raw score distribution for each form may affect equating results. In this study, the raw score distribution for each form indicated that there were more observations in the middle of the score range and fewer examinees through two ends. But

there were relatively more observations in the upper extreme scores than lower extreme scores. Therefore, the sampling errors in the lower extreme scores may result in conversion differences observed in Figure 1. Similar results also have been found in Yi et al.'s (2008) study. The findings of our study showed that as the conversion differences between total and subgroups decreased the equating function became group independent. For example, at the mid raw score points (7 and 8) conversion differences were near to zero (see Figure 1). Also these score points provided population invariance property based on their RMSD (x) values.

Lack of equating function group invariance indicates that the differential difficulty of the two tests is not consistent across the groups. Invariance can hold if the relative difficulty changes as a function of score level in the same way across subpopulations. If the relative difficulty of two test interacts with group membership then invariance does not hold (Dorans, 2004). In our study, especially at the lower raw score points there was a differential difficulty of two tests across the two groups. For example, examinees who felt themselves unsuccessful in history lesson found social science test harder than the examinees who felt themselves successful in history lesson. The same social science test scores were converted to higher equated score for feeling themselves unsuccessful in history lesson subgroups (see Table 3). We can also see the effects of relative difficulty changes as a function of score level in the opposite way across subpopulations in Figure 1. While equated scores derived from the total group were higher than derived from both subgroups at the low raw scores, the total group equated scores were lower than subgroups' equated scores at the higher raw scores.

Based on the self-efficacy theory developed by Bandura (1986), the people with high self-perception of capabilities display high motivation and attain high achievement. Because of that we divided examinees subgroups based on examinees' self-perceived competence in geography and history lessons which are related to the construct being measured. As Petersen (2008) noted "if the selection variable for constructing the subgroups is related to the construct being measured, I would expect to the equating results to exhibit population dependence" (p. 100). Dorans (2004) emphasized that important subpopulations are likely to affect equating functions in some degree, although group invariance never holds exactly, it should be expected to hold well enough under suitable assessment conditions. Our findings confirm both Petersen's (2008) and Dorans's (2004) views and supports Yi et al.'s (2008) findings.

The results also showed that the IRT true-score equating method was more group sensitive than the IRT observed-score equating method with respect to subgroups of self-perceived competence in geography and history lessons. Under all subgroups, the IRT true-score equating had larger REMSD values than the IRT observed-score equating method. The RMSD and REMSD indexes were originally suggested by Dorans and Holland (2000) for observed score linking functions. Later, von Davier and Wilson (2008) extended these statistics to the IRT true-score equating methods. In theory, if the assumptions of IRT hold, then the IRT true-score equating is invariant over all subpopulations. But in general, the group invariance of the IRT true-score equating does not hold while equating functions are used with observed scores (Brennan, 2008). Our study confirms the Brennan's proposal. The reason for higher sensitivity of the IRT true-score equating functions may be that the IRT true-score equating introduces more assumptions like "the relationship between true scores holds also for observed scores" (von Davier & Wilson, 2008, p. 13) than the IRT observed-score equating.

To provide test score interchangeability after test equating, group invariance of equating should be satisfy in some degree. As von Davier (2007) recommended group invariance of equating results should be examined routinely in operational test works. In future studies, for achieving better equating results, group invariance of equating results should be evaluated based on different grouping criteria and also what characteristic of data cause population dependence should be investigated.

REFERENCES

Bandura, A. (1986). *Social foundation of thought and action: A social cognitive theory.* Prentice Hall, NJ: Englewood Cliffs.

Braun, H. I., & Holland, P. W. (1982). Observed score equating: A mathematical analysis of some ETS equating

procedures. In P. W. Holland & D. B. Rubin (Eds.). *Test equating* (pp.9-49). New York: Academic Press.
Brennan, R. L. (2008). A discussion of population invariance. *Applied Psychological Measurement*, *32* (1), 102-114.
Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice*, *10*, 191-200.
Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225-244.
Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Philadelphia: Harcourt Brace Jovanovich College Publishers.
Dorans, N. J. (Ed.). (2003). *Population invariance of score linking: Theory and applications to Advanced Placement program examinations* (ETS Research Report RR-03-27). Princeton, NJ: Educational Testing Service.
Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227 246.
Dorans, N. J., Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrance, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT.* ETS Research memorandum. No. RM-94-10. Princeton, NJ: Educational Testing Service.
Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281-306.
Dorans, N.J., Holland, P. W., Thayer, D.T., & Tateneni, K. (2002, April). *Invariance of score linking across gender groups for three Advanced Placement Program exams.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, La.
Dorans, N., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32,* 81-97.
Eğitim, Araştırma ve Geliştirme Daire Başkanlığı. (2010). *Ortaöğretim ÖBBS raporu 2009*. Retrieved from http://egitek.meb.gov.tr/dosyalar/obbs/OBBS_2009.pdf
Hanson, B., & Zeng, L. (n.d.). *PIE: A computer program for IRT equating*. (Windows Console Version, Revised by Cui, May 20, 2004) [Manual]. Unpublished manuscript, College of Education, University of Iowa, Iowa City, Iowa
Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*(1), 35-43.
Holland, P. W., & Dorans, N. J. (2006). Linking and Equating. In R. L. Brennan (Ed.) *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger Publishers.
Huggins, A. C., & Penfield, R. D. (2012). An NCME instructional module on population invariance in linking and equating. *Educational Measurement: Issues and Practice*, *31*(1), 27-40.
Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice*, *7,* 29-36.
Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, *41*(1), 3-14.
Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.
Liu, M., & Holland, P. W. (2008). Exploring population sensitivity of linking functions across three law school admission test administrations. *Applied Psychological Measurement, 32,* 27-44.
Petersen, N. S. (2008). A discussion of population invariance of equating. *Applied Psychological Measurement, 32*, 98-101.
von Davier, A. A. (2007). Potential solutions to practical equating issues. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.). *Linking and Aligning Scores and Scales* (pp.89-105). New York: Springer.
von Davier, A. A., & Wilson, C. (2008). Assumption of item response theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*, 11-26.
Yang, W.-L., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based College Level Examination Program Examination. *Applied Psychological Measurement, 32,* 45-61.
Yi, Q., Harris, D., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement*, *32*, 62-80.
Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOGMG 3.0 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Eşitleme Sonuçlarının Gruptan Bağımsızlığının İncelenmesi

ÖZET

Çalışmanın Amacı ve Önemi: Bu araştırmada, Madde Tepki Kuramına (MTK) dayalı test eşitleme sonuçlarının test eşitlemenin gruptan bağımsızlık varsayımını sağlayıp sağlamadığı incelenmiştir. Gruptan bağımsızlık, eşitleme ilişkisinin eşitlemenin yürütüldüğü gruba bağlı olmaksızın aynı kalması demektir. Eşitleme fonksiyonlarının alt gruplardan bağımsız olması, eşitlemenin tüm bireyler için geçerli olmasını ve elde edilen puanların birbiri yerine kullanılabilmesini sağlar.

Yöntem: Bu araştırmanın çalışma grubunu, 2009 yılı ÖBBS uygulamasında B kitapçığını alan 15270 ve D kitapçığını alan 15323 9. sınıf öğrencisi oluşturmaktadır. Araştırmanın verileri, ÖBBS kapsamında ortaöğretim 9. Sınıf öğrencilerinin tarih ve coğrafya derslerindeki kazanımlarını ölçmek amacıyla hazırlanan sosyal bilimler testinden alınan puanlar ile Öğrenci Anketi'nde "Coğrafya dersinde kendinizi ne derece başarılı buluyorsunuz?" ve "Tarih dersinde kendinizi ne derece başarılı buluyorsunuz?" sorularına verilen yanıtlardır. Araştırma, eşdeğer gruplar test eşitleme deseni kullanılarak yürütülmüştür.

Verilerin analizi dört aşamada gerçekleştirilmiştir. Birinci aşamada, B ve D kitapçıklarını (formlarını) alan öğrenciler coğrafya dersinde kendini başarılı bulanlar-başarısız bulanlar ve tarih dersinde kendini başarılı bulanlar-başarısız bulanlar olmak üzere dört gruba ayrılmıştır. İkinci aşamada, her bir alt grup ve grupların tamamı için 3PL model ile madde parametreleri kestirilmiştir. Üçüncü aşamada, MTK gerçek-puan ve gözlenen-puan eşitleme yöntemleri ile puanlar eşitlenmiştir. Dördüncü aşamada, farklı alt gruplardan elde edilen eşitleme fonksiyonlarının gruptan bağımsızlığını değerlendirmek üzere RMSD ve REMSD indeksleri hesaplanmıştır.

Bulgular: Her bir ham puanının gruptan bağımsızlığı RMSD ve ölçek puanlarının tamamının gruptan bağımsızlığı REMSD indeksleri hesaplanarak incelenmiştir. RMSD ve REMSD değerlerinin istatistiksel olarak önemli olup olmadığına standartlaştırılmış DTM (SDTM) değeri ile karşılaştırılarak karar verilmiştir. Her iki eşitleme yöntemi ile elde edilen puanlar için tüm alt gruplarda sadece puan ölçeğinin ortalarında ve uç noktalarında RMSD değerinin SDTM değerinden küçük olduğu diğer noktalarda büyük olduğu görülmüştür. Ayrıca, her iki eşitleme yöntemi ile tüm alt gruplardan elde edilen REMSD değerlerinin SDTM değerinden büyük olduğu görülmüştür. Bu bulguya dayalı olarak, MTK eşitleme yöntemleri sonucu elde edilen eşitleme sonuçlarının coğrafya dersinde kendini başarılı bulan-bulmayan ve tarih dersinde kendini başarılı bulan-bulmayan alt gruplardan etkilendiğini söylenebilir.

Sonuç ve Tartışma: Bu araştırmanın sonucunda MTK gerçek-puan ve gözlenen-puan eşitleme sonuçlarının öğrencilerin coğrafya ve tarih derslerinde kendilerini başarılı bulup bulmamalarına göre oluşturulan gruplardan etkilendiği görülmüştür. Eşitleme fonksiyonlarının gruptan bağımsız olmaması, iki testin güçlüğünün gruplar arasında tutarlı olmadığını gösterir. Gruptan bağımsızlık varsayımı ancak testlerin göreceli güçlüğü alt gruplarda değişmediğinde sağlanabilir. Eğer iki test formunun göreceli güçlüğü bir gruba ait olup olmama ile etkileşiyorsa gruptan bağımsızlık varsayımı sağlanamaz (Dorans, 2004). Test eşitleme çalışmasından sonra elde edilen eşitlenmiş puanların birbiri yerine kullanılabilirliği için gruptan bağımsızlığın bir ölçüde sağlanması gerekir. İlerleyen araştırmalarda, daha başarılı eşitleme sonuçları elde etmek için eşitleme sonuçlarının gruptan bağımsızlığı farklı gruplamalara göre incelenmelidir.