ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);128-145

An Investigation of Group Invariance in Test Equating According to Gender

Test Eşitlemede Grup Değişmezliğinin Cinsiyete Göre İncelenmesi

Hatice İNAL * Çiğdem AKIN ARIKAN **

Abstract
The aim of this study is to investigate the group invariance condition according to Tucker and Levine observed score equating among linear equating methods. In the study, the 4th and 6th booklets of the PISA 2012 Mathematics subtest were used. Booklets were equated according to group and gender sub-variables, and then group invariance of each condition and WMSE values were calculated. Within this scope, REMSD and RMSD (x) group invariance indexes were employed. The results of the study indicated that, when WMSE values, obtained according to equating methods, were compared, Tucker observed score equating method with regard to whole-group and gender sub-groups produced the lowest error. When RMSD and REMSD values obtained according to gender sub-groups were examined by linear equating methods, it was found that group invariance value is smaller than criterion value for Tucker equating method, while it was greater than criterion value for Levine equating method. Eventually, group invariance condition was met for Tucker observed score equating, but not for Levine observed score equating.

Keywords: Equating, group invariance, Tucker linear equating, Levine linear equating

Öz
Bu çalışmanın amacı doğrusal eşitleme yöntemlerinden Tucker ve Levine gözlenen puan eşitleme yöntemlerine göre eşitlemenin grup değişmezliği koşulunun incelenmesidir. Bu çalışmada PISA 2012 matematik alt testine ait 4. ve 6. kitapçıklardan elde edilen test puanları kullanılmıştır. Kitapçıklardan elde edilen puanlar tüm grup ve cinsiyet alt değişkenine göre eşitlenmiştir. Her bir koşula ait grup değişmezliği ve WMSE değerleri hesaplanmıştır. Bu bağlamda grup değişmezliği indekslerinden REMSD ve RMSD (x) kullanılmıştır. Araştırma sonucunda eşitleme yöntemlerine göre elde edilen WMSE değerleri karşılaştırıldığında hem tüm grup hem cinsiyet alt grubuna göre en az hata veren yöntemin Tucker gözlenen puan eşitleme olduğu görülmüştür. Doğrusal eşitleme yöntemleriyle cinsiyet alt grubuna göre elde edilen RMSD ve REMSD değerleri incelendiğinde, Tucker eşitleme yöntemi için kriter değerden küçük iken, Levine eşitleme yöntemi için kriter değerden daha yüksek çıkmıştır. Böylece grup değişmezliği koşulunun Tucker eşitleme yönteminde sağlanırken, Levine eşitleme yönteminde sağlanmadığı görülmüştür.

Anahtar Kelimeler: Eşitleme, grup değişmezliği, Tucker eşitleme, Levine eşitleme

INTRODUCTION

PISA (Programme for International Student Assessment), that enables countries to compare their educational indicators, was administered by OECD in every three years since 2000. PISA application assesses to the extent which students at the age group of 15 are equipped with the basic mathematics, science and reading knowledge and skills in order to help them be a part of the modern society. PISA application aims to determine the extent students' ability to utilize knowledge and skills to use them in real life, understand the new situations, resolve problems, make guesses about what they are unfamiliar with and make judgments. In PISA application, students are required to take the all test item sets that consist of science, mathematics and reading skills. The items sets are incorporated in 13

* Araş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, e-posta: hinal@hacettepe.edu.tr
** Araş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, e-posta: akincgdm@gmail.com

Geliş Tarihi: 23.11.2016
DOI: *10.21031/epod.301326* *Kabul Tarihi: 06.03.2017*

booklets and there are some common items to link all the booklets (OECD, 2014). Therefore, it is necessary to equate the scores in order to compare these scores obtained from different booklets.

Equating can be described as the statistical process, which regulates the differences between the tests, forms with the same content and difficulty level and enables the scores obtained from these form to be used interchangeably (Kolen, 1988). The aim of test equating is to make sure that the difficulty of the test form does not create any advantage or disadvantage to the test taker. There are some conditions that must be met in order to equate the test forms. These conditions include equality, symmetry, group invariance and unidimensionality (Hambleton & Swaminathan, 1985). Among these conditions, group invariance means equating function is independent from the sub-groups so that sub-groups do not affect the equating (Kolen, 2004). For example, when two forms of a test are equated, it is possible to obtain the same equated scores for the female and males only when the group invariance condition is met. When group invariance is not ensured, students with different gender and same skills can obtain different equated scores and thus students have advantage or disadvantage because of their genders. In other words, it is fair to say that group invariance is related to equality and objectivity in assessment and evaluation (Dorans, 2004, 2008). In literature, different group invariance criteria have been developed to assess the accuracy as well as fairness of equated scores. These criteria are based on controlling the correspondence of equation principles (Petersen, Kolen & Hoover, 1989).

Test equating is divided into two groups as traditional and Item Response Theory approaches. Traditional equating methods include mean equating, linear equating and equipercentile equating methods (Kolen & Brennan, 2004). Mean equating is based on the assumption that test forms differ with respect to difficulty levels and this difference is fixed across whole scale. For example, in mean equating how much did responders in the upper group found X form easier than Y form will be the same for the individuals in the lower group (Kolen & Brennan, 2014). The equation of mean equating is as follows:

$$m_y(x) = y = x - \mu(X) + \mu(Y) \qquad (1)$$

If reference and score distribution of the new form are not equal, equipercentile equating method is used. It is accepted that in the score distribution of X and Y forms, the scores that correspond to the same percentile rank are equal. Equipercentile equating consists of two steps. First, cumulative frequencies of two forms are transferred to a table and cumulative frequency table is drawn. Second the scores that correspond to the same percentile rank are equated. With the scores that are obtained via equipercentile equating method, score distribution of the new form and reference form becomes similar (Livingston, 2004; Kolen, 1988).

When features of two test forms are the same except from means and standard deviations, linear equating is used (Crocker & Algina, 1986; Kolen & Brennan, 2014). In other words, the scores that correspond to the same standard scores (Z scores) are accepted as equal. If the standard deviations of test forms are equal, linear and mean equating will yield the same results. If raw scores and equated scores are given in the same graph, their linear relationship can be illustrated. Linear equating equation is presented in equation 2.

$$\frac{y-\bar{Y}}{s_y} = \frac{x-\bar{X}}{s_X} \qquad (2)$$

In linear equating, if the groups, which take the forms differ in terms of their skills, anchor items are used. Different linear equating methods have been developed to equate the forms, which have common items (Livingston, 2004).

Linear Equating Methods for the Non -Equivalent Groups

Non-Equivalent groups Anchor Test-NEAT, the common items pattern, is administered when it is not possible to administer the test form more than once due to test reliability in non-equivalent groups. In NEAT pattern, both forms incorporate some common items and these forms are administered on the non-equivalent groups. Equating relationship between the test forms is established via common items. Common items are classified as internal and external. If the score obtained from the common items is added to the test score of the test taker, it is called internal anchor, if not, it is called as external anchor. In linear equating for the NEAT pattern, equating relationship prediction is made over a single group by combining non-equivalent Group 1 and Group 2. Braun & Holland (1982) & Angoff (1971) named this group as synthetic. Group 1 and Group 2 classified as synthetic are weighted with w1 and w2. Weighting has two rules. The first of these rules is that the sums of two weights are completely equal (w1+w2=1) and the second one is that each weight equals to zero or is bigger than zero (w1, w2 ≥ 0). Even tough w1=w2=0,5 where two weights are equal are used in general, synthetic is used in (w1=1, w2=0) when group is only defined as new (Topczewski, Cui, Woodruff, Chen & Fang, 2013; Kolen & Breannan, 2014). In this study, the case in which both weights are equal was used.

Equation for linear equating in non-equivalent groups on common items pattern $ly_s(x)$ is the equation used for equating the X observed scores with Y observed scores and s stands for the synthetic group):

$$ly_s(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - (\mu_s(X)] + \mu_s(Y) \qquad (3)$$

μ_s (X) stands for the mean score of the new form obtained from the synthetic group, μ_s (Y) stands for the mean score of the reference form obtained from the synthetic group; σ_s(Y) stands for the standard deviation of the reference form obtained from the synthetic group, σ_s(X) stands for the standard deviation of the new form obtained from the synthetic group.

Four parameters of synthetic population in Equation 3, are indicated by the following Equations No. 4, 5, 6 and 7 for Group 1 and Group 2.

$$\mu_s(X) = w_1\mu_1(X) + w_2\mu_2(X) \qquad (4)$$
$$\mu_s(Y) = w_1\mu_1(Y) + w_2\mu_2(Y) \qquad (5)$$
$$\sigma_s^2(X) = w_1\sigma_1^2(X) + w_2\sigma_2^2(X) + w_1w_2[\mu_1(X) - \mu_2(X)]^2 \qquad (6)$$
$$\sigma_s^2(Y) = w_1\sigma_1^2(Y) + w_2\sigma_2^2(Y) + w_1w_2[\mu_1(Y) - \mu_2(Y)]^2 \qquad (7)$$

In non-equivalent groups, common items pattern $\mu_s(X), \mu_s(Y), \sigma_s^2(X)\ and\ \sigma_s^2(Y)$ cannot be calculated directly since Group 1 does not take X form and Group 2 does not take Y form. Therefore, some assumptions are required according to the equating methods used (Kolen & Brennan, 2004).

Linear equating methods that are used in non-equivalent groups common items pattern can be listed as Levine observed score equating, Levine true scores equating, chained linear equating, Braun-Holland Linear equating (Kolen & Brennan, 2014). Since a group can take only one form in non-equivalent groups common items pattern, linear equating also requires powerful statistical assumptions (Chen, Cui, Zhu & Gao, 2010). In this study, since Tucker and Levine observed score equating methods were used, only information about them was mentioned.

Tucker observed score equating

Tucker method was defined by Gulliksen in 1950 (Kolen & Brennan, 2014). The assumptions required for Tucker observed score equating method are related to regression and conditional variance. The first assumption requires the regression on the common item scores of total scores within both samples are equal. Conditional variance assumption requires variances of the total scores conditions are equal for both samples (Chen et al, 2010; Kolen & Brennan, 2014).

Levine observed scored equating

Levine originally developed the method in 1955 without considering the concept of a synthetic population. After improvements, this method became more general than Levine's (1955).

There are three assumptions of Levine observed score equating.

I. X, Y and common items measure the same characteristics and real scores of X, Y and common items are interlinked within both groups.

II. The regression of X and Y forms on common items are linear and equal within both groups.

III. Error variance of X and Y forms is equal within both groups (von Davier & Kong, 2003; Kolen & Brennan, 2014).

Group invariance is one of important condition to provide test score interchangeability. If group invariance wasn't met, it can be said that the equating couldn't be performed satisfactorily. Multiple studies that investigate the group invariance condition according to different subgroups s available (Dorans, 2004; Yang, 2004; von Davier & Han, 2004; Yin, Brennan & Kolen, 2004; von Davier & Wilson, 2008; Yang & Gao, 2008, Yi, Harris & Gao, 2008; Dorans, Liu & Hammond, 2008). However, although there are plenty of studies related to equating in Turkey (Kelecioğlu,1994; Şahhüseyinoğlu, 2005; Bozdağ & Kan, 2010; Kan, 2011; Kilmen, 2010; Gök, 2012; Öztürk, 2010; Kahraman, 2012; Kelecioğlu & Öztürk Gübeş, 2013; Mutluer, 2013; Demir & Güler, 2014; Atalay Kabasakal, 2014; İnci, 2014; Uysal, 2014), there is no more research which investigated group invariance of equating results. (Öztürk-Gübeş, N. & Kelecioğlu;2017).

The aim of this study is to equate test scores using Tucker and Levine observed score equating methods among linear equating methods according to non-equivalent groups common items pattern and to investigate whether or not group invariance condition of equating methods is met with respect to gender sub-groups. Additionally, in order to assess score equating, this study addressed how group invariance was applied by using real data.

Sub-problems

The purpose of the study is to investigate group invariance of the equated scores obtained from Tucker and Levine observed score equating method with respect to gender. For this purpose, these research questions were examined

1. How the results of Tucker and Levine are observed score equations for total score?
2. How the results of Tucker and Levine are observed score equating with regard to gender?
3. How the results of group invariance according to Tucker and Levine are observed score equating methods?
4. Which is the better option from Tucker and Levine equating methods to equate the test forms?

METHOD

This study aims to equate two booklets administered in Turkey in PISA 2012 (4th and 6th booklets) and assess the equating results. Therefore, this study can be considered as descriptive since the existing method and techniques were assessed via real data.

Population and Sampling

A total of 510 thousand students at the age of 15 participated in PISA application as the representatives of 28 million students from 65 countries in 2012. 4848 students from Turkey participated in PISA in 2012. The sample of the study consists of 741 students, who took 4th and 6th booklet of PISA in Turkish Descriptive statistics of the sample are presented in Table 1.

Table 1. Descriptive Statistics Regarding Gender

		Descriptive statistics				
Booklet	Gender	N	Mean	Standard deviation	Skewness	Kurtosis
Booklet 4	Female	182	13,302	6,873	,715	-,036
	Male	197	13,944	8,047	,555	-,719
Booklet 6	Female	178	12,601	7,088	,648	-,284
	Male	184	13,647	7,728	,768	-,082

Data Collection Tools

For data analysis, the data set of the mathematical literacy items by the Turkish students who participated into PISA 2012 application was used. There were 13 booklets in PISA 2012 application. The 4th and 6th booklets were used in this study. 4th booklet included 37, 6th booklet included 36 items. Since traditional equating methods are used in the present study, the most difficult item was excluded from the 4th booklet and the number of the items was equated. The data used in this study were downloaded from official website of OECD (http://pisa2012.acer.edu.au/). Later, correct answers, wrong answers and missing data were coded as 1, 0 and 0, respectively and all partially correct and correct answers to a couple of partially scored items were coded as 1 and the data to be analyzed was made ready.

Data Analysis

Data analysis was conducted at four steps. At the first step, it was examined whether or not the booklets met the equating conditions, at the second one the equated scores were obtained by using different equating methods, at the third one group invariance indexes were calculated in order to see how equating function obtained by each equating method differed across groups and at the final step error in each equating method was calculated.

I. Step: At the first step of data analysis, it was examined whether or not equating conditions are met.

To this end, primarily it was tested if the data was unidimensional. Tetrachoric correlation based principal components factor analysis, is used in order to determine the unidimensionality. This analysis was conducted with Factor 9.3 (2014) program developed by Lorenzo-Seva.

Table 2. Results of the factor analysis

	Booklet 4		Booklet 6	
Component	Eigenvalue	P.E.V (%)	Eigenvalue	P.E.V (%)
1	8.884	0.246	8.597	0.238
2	1.764	0.049	1.565	0.434

P.E.V (%): Proportion of explanation variance

The results of the factor analysis presented in Table 2 demonstrate that there is more than 4 times decline between the 1st factor and the 4th factor and the explanation variance of the second factor was quite low. Therefore booklets have a single general factor, which implies the tests meet the unidimensionality assumption.

Ratio test was administered in order to determine whether or not there was a significant relationship between the average difficulties of the forms (Baykul, 1996). The results of the test to compare the average difficulty of the booklets are presented in Table 3.

Table 3. Comparison of the average difficulty of the booklets

Booklets	$\bar{p}$	T	p
4	0.368	0.638	0.738
6	0.364		

*p<0.05

When Table 3 is examined, there is no statistically signifcant difference between difficulty levels of booklets (p>.05). In this case, the equality of average difficulty of the booklets to be equated, which is another condition for equation, is ensured.

KR-20 reliability coefficient was calculated in order detect if the booklets to be equated are equally reliable. Fischer’s Z statistics was carried out in order to detect if there was a difference between two reliability coefficients (Akhun, 1984). The findings regarding the differences in reliability coefficients are presented in Table 4.

Table 4. Comparison of the reliability of booklets

Booklets	KR-20	Z_r	Z	p
4	0.905	1.499	0.367	0.643
6	0.900	1.472		

When Table 4 is examined, it is seen that there is no significant difference between the reliability of the booklets at .05 alpha level/%95 confidence interval (p>.05). This demonstrates that the booklets meet the equal reliability condition.

T test and Levine test were used to test difference between the mean scores and variances of the booklets, respectively. The findings regarding the analyses are presented in Table 5.

Table 5. Comparison of the means and variances of the booklets

		t test			Levene’s test		
Booklets	N	$\bar{X}$	t	p	S^2	F	p
4	379	13.635	0.917	0.359	56.300	0.665	0.415
6	362	13.132			55.184		

When Table 5 is examined, it is seen that there is not a significant difference between the means and variances of the booklets at .05 level.

At the end of the analyses regarding the necessary conditions for equating, it was seen that the tests are one-dimensional, are equal in reliability, variances and average difficulty.

II. Step: At the second step of the data analysis, equated scores were obtained by using Levine and Tucker equating methods. Tucker and Levine observed score equating was performed in Microsoft Excel program.

III. Step: At the third step of the data analysis, group invariance indexes were calculated in order to assess whether equated scores obtained via each equating method differed in female and male subgroups. In this study, RMSD(x) and REMSD indexes developed by Dorans and Holland (2000) were employed in order to determine group invariance.

RMSD(x) (Root Mean Square Difference): The value found by RMSD(x) denotes the distance between the subgroup equating functions and total equating function at a x score level. In literature, studies indicating that RMSD(x) can be adapted to other equating method and patterns are available (von Davier, Holland & Thayer, 2004; von Davier & Wilson, 2008). These studies indicate that RMSD(x) can be reported in the form of other equating methods and patterns by eliminating the denominator of the equation in an unstandardized way.

x: Determined score level of the test form

j: Subgroup level

$e_{p_j}(x) - e_p(x)$**:** The difference between the equated score calculated based on the equating function of the subgroup j at an x score level with the equated score calculated based on the total equating function

w$_j$: The weight that is determined with the help of the ratio of the test-takers with the subgroups for each subgroup

σ_{YP}**:** Standard deviations of the scores in Q group (Q stands for the one and only group that is examined in single-group or random groups pattern) are defined with the following equation

$$RMSD\,(X) = \frac{\sqrt{w_j\left[e_{p_j}(x) - e_p(x)\right]^2}}{\sigma_{YP}} \qquad (8)$$

and with the help with this equation, it is possible to determine group invariance in case of single-group or equivalent groups equating pattern and linear equating function (Dorans & Holland, 2000).

$$RMSD\,(X) = \sqrt{w_j\left[e_{p_j}(x) - e_p(x)\right]^2} \qquad (9)$$

Dorans and Holland (2000) described the score level independent state of RMSD(x) as REMSD (Root Expected Mean Square Difference).

$$REMSD\,(X) = \frac{\sqrt{w_j E_p\left[e_{p_j}(x) - e_p(x)\right]^2}}{\sigma_{YP}} \qquad (10)$$

In this equation, E$_p$ stands for the mean score of the distribution found with the help of the differences between the equated scores. A group invariance study yields one REMSD. In literature, some studies stating that REMSD can be adapted to other equating method and patterns are available (von Davier, Holland & Thayer, 2004; von Davier & Wilson, 2008). These studies indicate that RMSD (x) can be reported in the form of other equating methods and patterns by eliminating the denominator of the equation in an unstandardized way.

In assessing the group invariance in equating, DTM criterion, which is taken as the half of the raw score unit and recommended by Dorans, Holland, Thayer & Tateneni (2003) and Dorans (2004) is utilized. It is not a certainly set rule to assess the group invariance based on DTM scope. In this study interpretations were made by considering that the difference smaller than 0.50 between equated score of the whole group and the equated score of a sub-group(s) is negligible and difference bigger than 0.50 is significant (Kolen & Brennan, 2014).

IV. Step: At the final step of the data analysis, error of each equating method was calculated. In this study, weighted mean squares error (WMSE) was used in order to assess equating error.

WMSE (Weighted Mean Squares Error): It is used in order determine which method is the most suitable in line with the error of the scores equated according to different equating methods. Weighted mean squares error (WMSE) is calculated by comparing the equated scores corresponding to each raw score at the same skill level (Skaggs & Lissitz, 1986). Skaggs and Lissitz (1988) reported that WMSE index is quite similar to the total error indexes available in other equating studies. The equation for the calculation of WMSE coefficient is given below:

$$WMSE = \frac{\sum_{i=1}^{k-1} fi(X_E - X_{Crit})^2}{\sum_{i=1}^{k} f_i S^2 y} \tag{11}$$

k : The number of the items in Y test.
S^2y**:** Variance of the raw scores in Y test.
X $_{crit}$: i. raw score in Y test.
X$_E$: the score obtained via equating methods and that correspond to i. raw score in X test.
fi: i. raw score frequency in Y test

FINDINGS

The equated scores of PISA 2012 Mathematics sub-test obtained for Tucker and Levine observed score equating methods with respect to gender and the raw scores are presented in Table 6. The graphs regarding the raw scores obtained for both methods and equated scores are given in the appendix.

Table 6. Raw scores and the scores that correspond to these scores that are obtained via Tucker observed score equating methods

	Total		Female		Male	
Raw Score	Equated Score	Difference	Equated Score	Difference	Equated Score	Difference
0	0.945	-0.945	0.722	-0.722	1.007	-1.007
1	1.936	-0.936	1.935	-0.935	2.021	-1.021
2	2.927	-0.927	2.822	-0.822	3.034	-1.034
3	3.917	-0.917	3.787	-0.787	4.048	-1.048
4	4.908	-0.908	4.752	-0.752	5.062	-1.062
5	5.898	-0.898	5.717	-0.717	6.075	-1.075
6	6.889	-0.889	6.682	-0.682	7.089	-1.089
7	7.879	-0.879	7.647	-0.647	8.102	-1.102
8	8.870	-0.870	8.612	-0.612	9.116	-1.116
9	9.860	-0.860	9.577	-0.577	10.129	-1.129
10	10.851	-0.851	10.542	-0.542	11.143	-1.143
11	11.841	-0.841	11.507	-0.507	12.157	-1.157
12	12.832	-0.832	12.472	-0.472	13.170	-1.170
13	13.822	-0.822	13.437	-0.437	14.184	-1.184
14	14.813	-0.813	14.401	-0.401	15.197	-1.197
15	15.803	-0.803	15.366	-0.366	16.211	-1.211
16	16.794	-0.794	16.331	-0.331	17.224	-1.224
17	17.784	-0.784	17.296	-0.296	18.238	-1.238
18	18.775	-0.775	18.261	-0.261	19.252	-1.252
19	19.765	-0.765	19.226	-0.226	20.265	-1.265
20	20.756	-0.756	20.191	-0.191	21.279	-1.279
21	21.747	-0.747	21.156	-0.156	22.292	-1.292
22	22.737	-0.737	22.121	-0.121	23.306	-1.306
23	23.728	-0.728	23.086	-0.086	24.319	-1.319
24	24.718	-0.718	24.051	-0.051	25.333	-1.333
25	25.709	-0.709	25.016	-0.016	26.347	-1.347
26	26.699	-0.699	25.981	0.019	27.360	-1.360
27	27.690	-0.690	26.946	0.054	28.374	-1.374
28	28.680	-0.680	27.911	0.089	29.387	-1.387
29	29.671	-0.671	28.876	0.124	30.401	-1.401
30	30.661	-0.661	30.661	-0.661	31.415	-1.415
31	31.652	-0.652	30.806	0.194	32.428	-1.428
32	32.642	-0.642	32.642	-0.642	33.442	-1.442
33	33.633	-0.633	33.632	-0.632	34.455	-1.455
34	34.622	-0.622	34.066	-0.066	35.469	-1.469
35	35.614	-0.614	35.613	-0.613	36.482	-1.482
36	36.603	-0.603	36.028	-0.028	37.496	-1.496

When Table 6 is examined, shown raw scores range between 0-36. It is seen that equated scores for all groups range between 0.945 and 36.603. For women range between 0.722-36.028 and for men range between 1.007-37.496. As can be seen from the table, according to Tucker equating method, the equated scores between 26-29 range and at 31st raw scores are smaller than the raw scores and the other equated scores are bigger than the raw scores. It was also found that in males, equated scores are higher than the raw scores. Based on these findings, it can be said that 6th booklet was more difficult than 4th one for whole-group and males. Although this was the case for females in a general sense, this situation changes between 26-29 interval and 31st raw scores.

Table 7. Raw Scores and the scores corresponding to the raw scores that are obtained via Levine observed score equating method

	Total		Female		Male	
Raw Score	Equated Score	Difference	Equated Score	Difference	Equated Score	Difference
0	1.167	-1.167	1.159	-1.159	1.000	-1.000
1	2.164	-1.164	2.164	-1.164	2.032	-1.032
2	3.155	-1.155	3.356	-1.356	3.063	-1.063
3	4.146	-1.146	4.293	-1.293	4.094	-1.094
4	5.137	-1.137	5.229	-1.229	5.125	-1.125
5	6.128	-1.128	6.166	-1.166	6.157	-1.157
6	7.118	-1.118	7.103	-1.103	7.188	-1.188
7	8.109	-1.109	8.039	-1.039	8.219	-1.219
8	9.100	-1.100	8.976	-0.976	9.251	-1.251
9	10.091	-1.091	9.912	-0.912	10.282	-1.282
10	11.082	-1.082	10.849	-0.849	11.313	-1.313
11	12.073	-1.073	11.785	-0.785	12.344	-1.344
12	13.064	-1.064	12.722	-0.722	13.376	-1.376
13	14.054	-1.054	13.659	-0.659	14.407	-1.407
14	15.045	-1.045	14.595	-0.595	15.438	-1.438
15	16.036	-1.036	15.532	-0.532	16.469	-1.469
16	17.027	-1.027	16.468	-0.468	17.501	-1.501
17	18.018	-1.018	17.405	-0.405	18.532	-1.532
18	18.775	-0.775	18.341	-0.341	19.563	-1.563
19	19.999	-0.999	19.278	-0.278	20.594	-1.594
20	20.990	-0.990	20.214	-0.214	21.626	-1.626
21	21.981	-0.981	21.151	-0.151	22.657	-1.657
22	22.972	-0.972	22.088	-0.088	23.688	-1.688
23	23.963	-0.963	23.024	-0.024	24.719	-1.719
24	24.954	-0.954	23.961	0.039	25.751	-1.751
25	25.944	-0.944	24.897	0.103	26.782	-1.782
26	26.935	-0.935	25.834	0.166	27.813	-1.813
27	27.926	-0.926	26.770	0.230	28.844	-1.844
28	28.917	-0.917	27.707	0.293	29.876	-1.876
29	29.908	-0.908	28.643	0.357	30.907	-1.907
30	30.899	-0.899	30.898	-0.898	31.938	-1.938
31	31.889	-0.889	30.517	0.483	32.969	-1.969
32	32.880	-0.880	32.880	-0.880	34.001	-2.001
33	33.871	-0.871	33.871	-0.871	35.032	-2.032
34	34.854	-0.854	33.948	0.052	36.064	-2.064
35	35.853	-0.853	35.852	-0.852	37.095	-2.095
36	36.836	-0.836	35.877	0.133	38.127	-2.127

As can be seen from Table 7, while the raw scores between 0-36 score interval, the equated scores change between 1.167 and 36.836 for the whole-group, 1.159-35.877 for females and 1-38.127 for males. The results of the Levine observed score equating indicate that raw scores for whole-group and males are lower than the equated scores. However for females while raw scores are lower than equated scores between 0-23 raw score interval, they are lower and higher for some scores between 24-36 score interval.

In linear equating that regulates the difficulty difference of the forms across all scale scores, it was revealed that in both methods used in the study, there was a linear relationship between raw scores and equated scores for whole-group and males. There is no difference across whole number scale and only show difference between 24-36 score interval. It is fair to say that in Levine observed score equating 6th booklet was found to be more difficult than 4th one for whole-group and males and although it was the case for females in a general sense, this situation changes in raw scores between 24-31 interval.

The graphs of RMSD (x) index that correspond to each score in which group invariance of the Tucker and Levine observed score equating is examined according to the gender subgroup are presented in Figure 1 and Figure 2 respectively. The RMSD (x) values are given in the appendix in Table 1.

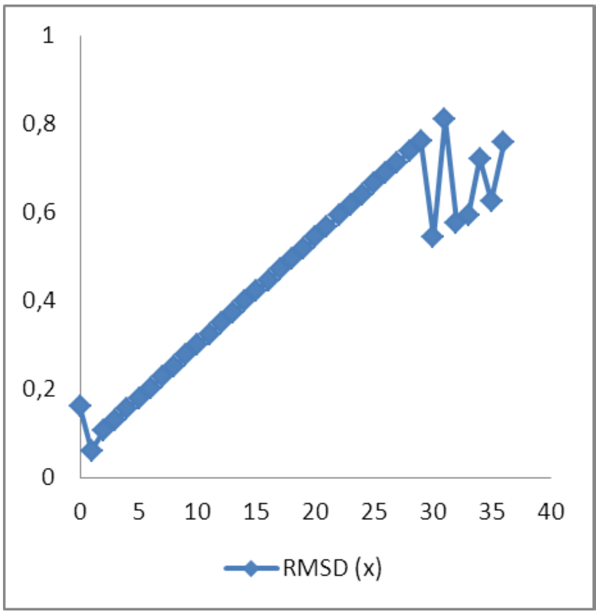


Figure 1. RMSD (x) for Tucker Equating

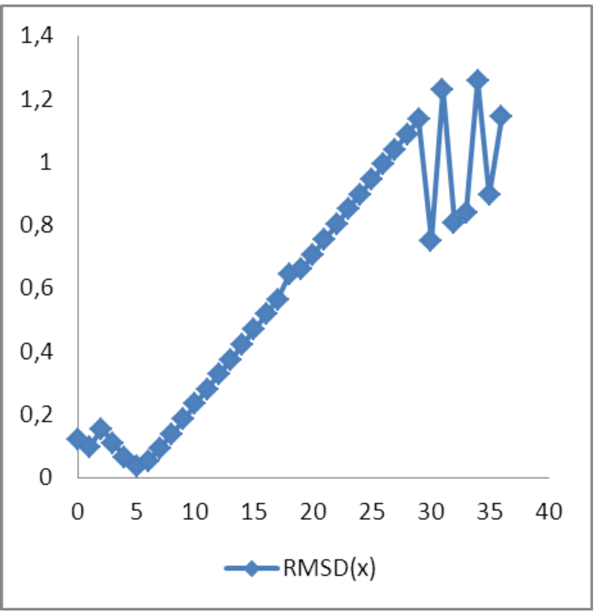


Figure 2. RMSD(x) for Levine Equating

When Figure 1 and Table 1 in appendix are examined, it is seen that RMSD (x) values range between 0.061 and 0.811 for Tucker equating and these values increased in simultaneously with the score in a general sense. However, this case differs when it comes to high scores. For Tucker equating method, the highest RMSD (x) value was obtained at 31 score level and the lowest one was obtained at 1 score level. In Figure 2, it is seen that RMSD (x) values for Levine Equating range between 0.034 and 1.257. Although it is seen that RMSD (x) values increased in simultaneously with the score in Levine equating method, it was found out that the increase was not linear at extreme values. In this method, the highest RMSD(x) score was obtained at 34 score level and the lowest one at 5 score level. According to RMSD values, there are some fluctuations in the extreme points of the scale in the graph for both equating methods. When the frequency of scores was examined, some extreme scores had fewer frequency than the others. Accordingly, fluctuations in the extreme points can be originated from the difference of frequencies.

In this study, it was found out that RMSD(x) values calculated with both methods were similar, however, RMSD(x) values for Tucker were smaller than the RMSD(x) values for Levine.

On the other hand, it is seen that RMSD(x) values that correspond to the scores between 1 and 18 for Tucker equating are lower than DTM. This means that the difference between the equated score in whole-group and equated scores in sub-groups is not significant. However, RMSD(x) values that correspond to the scores between 19 and 35 for Tucker equating are higher than DTM which means that the difference between equated score in whole group and equated scores in sub-groups is significant. For Levine equating, it is seen that RMSD(x) values that correspond to the scores between 1 and 15 for are lower than DTM. This means that the difference between the equated score in whole-group and equated scores in sub-groups is not significant. However, RMSD(x) values that correspond to the scores between 16 and 35 are higher than DTM. Therefore the difference between equated scores in whole-group and equated scores in sub-groups is significant.

RMSD (x) index that correspond to each score in which group invariance of the scores equated according to Tucker and Levine observed score equating in gender sub-group is examined is given above. REMSD values that are calculated at group invariance total score level are presented in Table 8.

Table 8. Values for Levine and Tucker Equating Methods

Equating Methods	REMSD
Tucker-Linear Equating	0.496
Levine-Linear Equating	0.668

As shown in the Table 8 Tucker equating, REMSD value was calculated as 0.496 and as 0.668 for Levine equating method. It is seen that REMSD value obtained for Tucker is lower than the REMSD value obtained for Levine. Besides Tucker equating RMSD(x) values are lower than DTM. This implies the difference between the equated score in whole-group and equated scores in sub-groups is not significant. However, for Levine equating, it is seen that RMSD(x) values are higher than DTM. This means that the difference between equated scores in whole-group and equated scores in sub-groups is significant.

WMSE (AHKO) coefficients were calculated according to Tucker and Levine equating methods and gender sub-group determined for invariance in order to find if Tucker or Levine is more suitable for the PISA 2012 4th and 6th booklets which included mathematics test. The information regarding coefficients is presented in Table 9.

Table 9. WMSE (AKHO) Values for Levine and Tucker Equating Methods

Equating Methods	Total	Female	Male
Tucker-Linear Equating	0.012	0.004	0.024
Levine-Linear Equating	0.199	0.112	0.035

Table 9 indicates that according to whole-group and sub-groups, the most suitable method regarding the mathematics sub-test in PISA 2012 included in 4th and 6th booklets is Tucker equating method. It is striking that in Tucker equating method, WMSE value obtained for males is quite higher than the WMSE value obtained for females. It is fair to say that WMSE coefficients obtained for males via both methods from sub-groups are similar.

DISCUSSION and CONCLUSION

In this study, equating errors of the scores obtained according to Tucker and Levine observed score equating methods were compared by equating with the 6^{th} and 4^{th} booklets of PISA 2012 Mathematics subtest and in order to assess the equitability of the scores, whether or not group invariance is was investigated according to RMSD (x) for each score and REMSD coefficients for total score.

When the scores obtained via linear equating are examined, it was seen that the scores obtained according to Tucker and Levine observed score equating take values out of raw score range. Livingston (2004) maintained that the scores equated in linear equating can go outside the raw score range and that does not create a problem for linear equating and is a characteristics specific to linear equating. Moreover, Livingston (2004) reported that the equated scores at very high and low scores can exceed the score range. This was observed at high scores in both equating methods according to female sub-group.

When WMSE values obtained based on the Tucker and Levine observed score equating methods are compared, it was found out that Tucker observed score equating produced lowest error for both whole-group and gender sub-group. While errors that are obtained according to Tucker and Levine observed score equating with regard to whole-group and female sub-group show difference, it can be said that errors that are obtained with regard to males sub-group are close. Similar results are obtained when the past studies are examined. A study by Demir & Güler (2014) compared frequency prediction equipercentile equating, Tucker, Levine and Braun-Holland Linear Equating methods and determined that the most appropriate method was Tucker equating method and also reported that Levine observed score equated method had the highest error. Topczewski et al., (2013) stated in their study in which they used a different version of Tucker, Angoff-Levine, congeneric -Levine and a different version of congeneric Levine by addressing the differences between the skills of the groups that Tucker equating method was the most suitable one in case that group variance is similar. Chen et al. (2003) performed Tucker and Levine observed score equating methods by using different skills distribution and tests with different difficulty levels and concluded that the results were similar when the difference between the group and tests forms was small.

When RMSD and REMSD values obtained according to gender sub-group via linear equating are examined, it was seen that the RMSD and REMSD values based on Tucker were lower than the ones based on Levine. Besides, the difference between the equated scores in whole-group and the scores equated for sub-groups is not significant for Tucker equating method, although it is significant for Levine equating method. That is to say that while group invariance is at an acceptable level for Tucker equating method, it is not the case for Levine equating method. In the study by von Davier and Han (2004) which compared RMSD values with respect to gender with Levine observed score and chained linear equating methods, it was observed that the equating function with the lowest changing equating rate belonged to Levine while the highest changing function belonged to Tucker method. It was found out that the present study and the relevant study results were not parallel. The study by Dorans, Liu & Hammond (2008) reported in their study in which they compared group invariance by gender with Tucker, Levine and Chained equating methods revealed that if the groups to be equated are similar in terms of average skills, Trucker equating method is more fruitful than Levine and Chained equation results. Also Yin, Brennan & Kolen (2004) investigated the group invariance of linear, parallel-linear and equipercentile equating of mathematics and science tests in their study. They reported that lower REMSD values were obtained via linear and parallel linear equating methods for mathematics tests, while lower REMSD via equipercentile equating was reported for the science test. It is seen that results of both studies support the current study.

Equitability of scores requires the same meaning regardless of when or when the equalized points are applied. Failure to achieve group invariance in equating function indicates that the difficulty difference of the old and new test forms in NEAT pattern is inconsistent across subgroups (Kim and Walker, 2009). Violation of group invariance condition in equating causes the individuals from different groups who are supposed to have the same score get different equated scores (Dorans, 2004, 2008). Group invariance is a prerequisite for equating. Failure to achieve group invariance is an indicator that equating has not succeeded completely. However, achieving group invariance does not necessarily mean that equated scores can be used interchangeably. This is because group invariance should not be taken as the only criterion in assessing the quality of the equation (Dorans, Liu & Hammond, 2008).

The usage of group invariance indexes made it possible to decide which equating method can be achieved better than the other. Based on the findings of current study, Tucker equating method was the best option in terms of equating 4th and 6th Mathematics Booklets of PISA 2012 and group invariance.

The difference observed in group invariance might be attributed to the difference between the whole and sub-group samples. The sample size of this study is 741, 381 and 360 for the whole group, males and females, respectively and a sample size between 50 and 100 is sufficient for Tucker and Levine observed score equating methods (Parshall, Du Bose Houghton & Kromrey, 1995; Skaggs, 2005; Babcock, Albano & Raymond, 2012). Since the sample sizes are sufficient in this study, it can be said that the difference in group invariance is not affected by the sample size.

In this study, 4th and 6th booklets of PISA 2012 mathematics sub test were equated by using Tucker and Levine observed score equating method in non-equivalent groups' common items pattern and it was investigated whether or not group invariance was achieved with regard to gender sub-group. A similar study can be carried out by using different equating methods, equating patterns and different samples. Also, whether or not group invariance condition was met with regard to gender sub-group was examined via RMSD (x) and REMSD indexes. In different studies, difference group invariance indexes can be used according to different sub-groups (socioeconomics, ethnic groups, countries etc.).

REFERENCES

Akhun, İ. (1984). İki korelasyon katsayısı arasındaki manidarlığın test edilmesi. *Ankara Üniversitesi Eğitim Fakültesi Dergisi. 17*, 1-7.

Angoff, W. (1996). Scales, norms, and equivalent scores. *Educational Measurement: Theories and Applications*, *2*, 121.

Atalay Kabasakal, K. (2014). *Değişen madde fonksiyonunun test eşitlemeye etkisi* (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Baykul, Y. (1996*). İstatistik: Metodlar ve uygulamalar* (3. Baskı). Ankara: Anı Yayıncılık

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational And Psychological Measurement, 72*(4), 608-628.

Bozdağ, S., & Kan, A. (2010). Şans başarısının test eşitlemeye etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 39*, 91-108.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. *Test equating*, 1982, 9-49.

Chen, H., Cui, Z., Zhu, R., & Gao, X. (2010). *Evaluating the effects of differences in group abilities on the Tucker and the Levine observed-score methods for common-item nonequivalent groups equating*. ACT Research Report Series (2010-(1)). Iowa City, IA: ACT.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. NewYork: Holt, Rinehart& Winston.

Demir, S., & Güler, N. (2014). Ortak maddeli denk olmayan gruplar desenine ilişkin test eşitleme çalışması. *International Journal of Human Sciences, 11*(2), 190-208.

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 equitability of tests: Basic theory and the linear case. *Journal of Measurement, 37*, 281-306.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender group for three Advanced Placement Program examinations. In N. J. Dorans (Ed.). *Population invariance o f score linking: Theory and applications to Advanced Placement Program examinations* (RR-03-27). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*, 81-98.

Gök, B. (2012). *Denk olmayan gruplarda ortak madde deseni kullanılarak Madde Tepki Kuramına dayalı eşitleme yöntemlerinin karşılaştırılması* (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston: Kluwer.

İnci, Y. (2014). *Örneklem büyüklüğünün test eşitlemeye etkisi* (Yayınlanmamış Yüksek Lisans Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Kahraman, H. (2012). *Düzgünleştirilmiş puanların eşitleme hatasına etkisi* (Yayınlanmamış Yüksek Lisans Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Kan, A. (2011). Test eşitleme: OKS testlerinin istatistiksel eşitliğinin sınanması. *Eğitim ve Bilim, 36*(160), 38-51.

Kelecioğlu, H. (1994). *Öğrenci Seçme Sınavı puanlarının eşitlenmesi üzerine bir çalışma* (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü Ankara.

Kelecioğlu, H., & Öztürk Gübeş, N. (2013). Comparing linear equating and equipercentile equating methods using random groups design. *International. Online Journal of Educational Sciences, 5*(1), 227-241.

Kilmen, S. (2010). *Madde Tepki Kuramına dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması* (Yayınlanmamış doktora tezi). Ankara Üniversitesi, Eğitim Bilimler Enstitüsü, Ankara.

Kim, S., & Walker, M. E. (2009). *Evaluating subpopulation invariance of linking functions to determine the anchor composition for a mixed-format test*. ETS Research Rep. No. RR-09-36. Princeton, NJ: Educational Testing Service.

Kolen, M. J. (1988). An NCME intructional module on traditional equating methodology. *Educational Measurement: Isuues and Practice, 7*, 29-36.

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

Kolen, M., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd Ed.). New York: Springer.

Kolen, M., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd Ed.). New York: Springer.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability*. ETS Research Report Series, 1955(2).

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.

MEB. (2010). *PISA 2009 projesi ulusal ön raporu*. MEB Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

MEB. (2013). *PISA 2012 ulusal ön raporu*. MEB Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.

Mutluer, C. (2013). *Yıl içinde farklı dönemlerde yapılan Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES) puanlarına ilişkin bir test eşitleme çalışması* (Yayınlanmamış yüksek lisans tezi). Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Enstitüsü, Bolu.

OECD. (2014). PISA Technical Report, OECD Publishing. http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

Öztürk, N. (2010). *Akademik personel ve lisansüstü eğitimi giriş sınavı puanlarının eşitlenmesi üzerine bir çalışma* (Yayınlanmamış yüksek lisans tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü. Ankara.

Öztürk Gübeş, N., & Kelecioğlu, H. (2017). Investigating group invariance of equating results. *Elementary Education Online, 16*(1), 217-227.

Parshall, C. G., Du Bose Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32*, 37-54.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.) *Educational Measurement* (pp.221-262). New York: Macmillan.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*, 495–529.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12*, 69–82.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*, 309-330.

Şahhüseyinoğlu, D. (2005). *İngilizce yeterlik sınavı puanlarının üç farklı eşitleme yöntemine göre karşılaştırılması* (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü. Ankara.

Topczewski, A., Cui, Z., Woodruff, D., Chen, H., & Fang, Y. (2013). *Comparison of four linear equating methods for the common-item nonequivalent groups design using simulation methods*. ACT Research Report Series (2013-(2). Iowa City, IA: ACT.

Uysal, İ. (2012). *Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması* (Yayınlanmamış yüksek lisans tezi). Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Enstitüsü, Bolu.

von Davier, A. A., & Han, N. (2004). *Population invariance and linear equating for the non-equivalent groups design*. (ETS RR-04-47). Princeton, NJ: Educational Testing Service.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York: Springer.

von Davier, A. A., & Kong, N. (2003). *A unified approach to linear equating for the non-equivalent groups design*. (ETS SR-03-31). Princeton, NJ: Educational Testing Service.

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*, 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.

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.

Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement, 28,* 274-289.

Yi, Q., Harris, D. J., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement, 32,* 62-80.

UZUN ÖZET

Giriş

Eşitleme benzer içerik ve güçlük düzeyinde geliştirilen test formları arasındaki farklılıkları düzenleyerek, bu formlardan elde edilen puanların birbiri yerine kullanılmasını sağlayan istatistiksel bir süreç olarak tanımlanabilir. Test eşitlemede amaç, kolay ya da zor test formunu alan bireye formun herhangi bir avantaj veya dezavantaj sağlamamasıdır. Test formlarının eşitlenebilmesi için eşitlik, simetri, grup değişmezliği ve tek boyutluluk gibi bazı koşulların karşılanması gerekmektedir. Bu koşullardan biri olan grup değişmezliği, eşitleme fonksiyonunun alt gruplardan bağımsız olması ve alt grupların eşitlemeyi etkilememesi anlamına gelmektedir. Bu araştırmanın amacı denk olmayan gruplarda ortak madde desenine göre doğrusal eşitleme yöntemlerinden Tucker ve Levine gözlenen puan eşitleme yöntemleriyle eşitlenmesi sonucunda cinsiyet alt grubuna göre eşitleme yöntemlerinin grup değişmezliği koşulunun sağlanıp sağlanmadığının incelenmesidir.

Yöntem

Bu araştırmanın örneklemini Türkiye'deki PISA 2012 uygulamasına katılan öğrenciler arasından, bu uygulama esnasında 4. ve 6. kitapçıkları alan 741 öğrenci oluşturmaktadır. Bu çalışmada veri toplama aracı için PISA 2012 uygulanmasındaki 4 ve 6 nolu kitapçıklarda yer alan maddeler kullanılmıştır.

Bu araştırmada verilerin analizi dört aşamada gerçekleştirilmiştir. Verilerin analizinin birinci aşamasında, eşitleme koşullarının sağlanıp sağlanmadığı test edilmiştir.

Bunun için ilk olarak, verinin tek boyutlu olup olmadığı test edilmiştir. Tek boyutluluğun belirlenmesi için iki kategorili veriler için kullanılan tetrakorik korelasyona dayalı temel bileşenler faktör analizi yöntemi seçilmiştir. Formların ortalama güçlükleri arasında anlamlı bir farkın olup olmadığını belirlemek için iki oran fark testi yapılmıştır. Eşitlenecek kitapçıkların eşit güvenirliğe sahip olup olmadığını görebilmek için KR-20 güvenirlik katsayısı hesaplanmıştır. İki güvenirlik katsayısı arasında fark olup olmadığı belirlemek için Fischer'ın Z istatistiği yapılmıştır. Eşitleme yapılacak kitapçıkların ortalama ve varyansları arasında fark olup olmadığı bağımsız gruplar t testi ve Levene testi ile incelenmiştir. Eşitleme için gerekli koşullar ile ilgili yapılan analizler sonucunda, testlerin tek boyutlu olduğu; güvenirliklerinin, varyanslarının ve ortalama güçlüklerinin eşit olduğu görülmüştür.

Veri analizinin ikinci aşamasında, eşitleme yöntemleri kullanılarak eşitlenmiş puanlar elde edilmiştir.

Veri analizinin üçüncü aşamasında, her bir eşitleme yöntemi ile elde edilen eşitlenmiş puanların cinsiyete göre nasıl değiştiğini değerlendirmek için grup değişmezliği indeksleri hesaplanmıştır. Bu çalışmada grup değişmezliğini belirlemek için Dorans ve Holland (2000) tarafından geliştirilen RMSD ve REMSD indeksleri kullanılmıştır. Eşitlemede grup değişmezliğinin değerlendirilmesinde DTM ölçütünden yararlanılmıştır. Bu çalışmada 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 anlamlı olduğu kabul edilerek yorumlar yapılmıştır.

Veri analizinin son olarak dördüncü aşamasında, her bir eşitleme yönteminde yapılan hata miktarı hesaplanmıştır. Eşitleme hatasını değerlendirmek için ağırlıklandırılmış hata kareleri ortalaması (WMSE) ölçütü kullanılmıştır.

Sonuçlar ve Tartışma

Doğrusal eşitleme sonucunda elde edilen puanlar incelendiğinde, Tucker ve Levine gözlenen puan eşitlenmesine göre elde edilen puanların, ham puan ranjının dışında değerler aldığı görülmüştür.

Tucker ve Levine gözlenen puan eşitleme yöntemlerine dayalı olarak elde edilen WMSE değerleri karşılaştırıldığında ise, hem tüm grup hem de cinsiyet alt grubuna göre en az hata veren yöntemin Tucker gözlenen puan eşitleme olduğu görülmüştür. Toplam grup ve kadın alt grubuna göre Tucker ve Levine gözlenen puan eşitlemesine göre elde edilen hata değerleri büyük farklılık gösterirken, erkekler alt grubuna göre elde edilen hata değerlerinin birbirine yakın olduğu bulunmuştur.

Doğrusal eşitleme ile cinsiyet alt grubuna göre elde edilen RMSD ve REMSD değerleri incelendiğinde, Tucker için elde edilen RMSD ve REMSD değerlerinin Levine için elde edilen değerlerden daha küçük olduğu görülmüştür. Ayrıca toplam gruptaki eşitlemiş puan ile alt gruplardaki eşitlenmiş puanlar arasındaki farklılığın Tucker eşitleme yöntemi için anlamsız olduğu; Levine eşitleme yöntemi için ise anlamlı olduğu sonucuna ulaşılmıştır.

Puanların eşitlenebilirliği, eşitlenmiş puanların ne zaman ya da hangi gruba uygulandığına bakılmaksızın aynı anlama gelmesini gerektirmektedir. Eşitleme fonksiyonundaki grup değişmezliğinin sağlanamaması, NEAT deseninde eski ve yeni test formlarının güçlüklerindeki farklılığın alt gruplar boyunca tutarlı olmadığını göstermektedir. Eşitlemede grup değişmezliğinin ihlali, aynı puana sahip olması gereken farklı gruplara ait bireylerin, farklı eşitlenmiş puanlar almasına neden olmaktadır. Grup değişmezliği eşitleme için bir önkoşuldur. Grup değişmezliğinin sağlanamaması eşitlemenin tam olarak gerçekleşmediğinin kanıtı olarak ele alınabilir. Ancak grup değişmezliğinin sağlanması da eşitlenmiş puanların birbiri yerine kullanılabileceği anlamına gelmez. Çünkü, eşitlemenin niteliğinin değerlendirilmesinde tek kriter grup değişmezliği değildir.

Bu çalışmada grup değişmezliği indekslerinin kullanımı, puanların eşitlenebilirliğinin hangi yöntemle daha iyi sağlandığına karar verilmesine olanak vermiştir. Elde edilen bulgulara dayalı olarak, PISA 2012 matematik 4. ve 6. Kitapçıkların eşitlenmesinde ve grup değişmezliği açısından en uygun yöntemin Tucker eşitleme yöntemi olduğuna ulaşılmıştır.

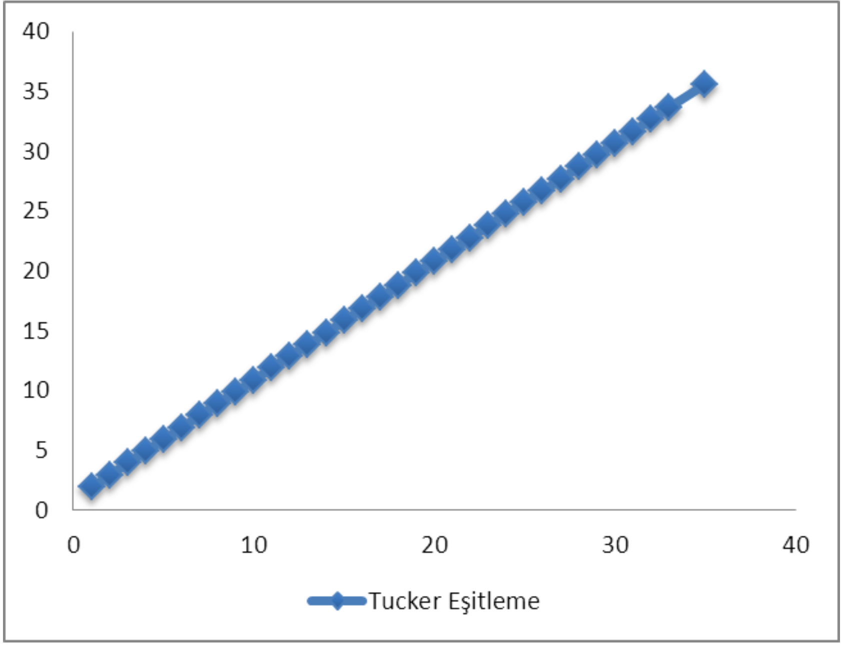
Bu araştırmada PISA 2012 matematik 4. ve 6. Kitapçıklar denk olmayan gruplarda ortak madde test deseninde Tucker ve Levine gözlenen puan eşitleme yöntemi kullanılarak eşitlenmiş ve cinsiyet alt grubuna göre grup değişmezliğinin sağlanıp sağlanamadığı incelenmiştir. Benzer bir araştırma farklı eşitleme yöntemleri, eşitleme desenleri ve farklı örneklemler kullanılarak yapılabilir. Ayrıca grup değişmezliğinin sağlanıp sağlanamadığı cinsiyet alt grubuna göre RMSD ve REMSD indeksleriyle yapılmıştır. Diğer çalışmalarda farklı alt gruplara (sosyo ekonomik, etnik grup, ülkeler vb.) göre farklı grup değişmezliği indeksleri ile yapılabilir.

APPENDICES

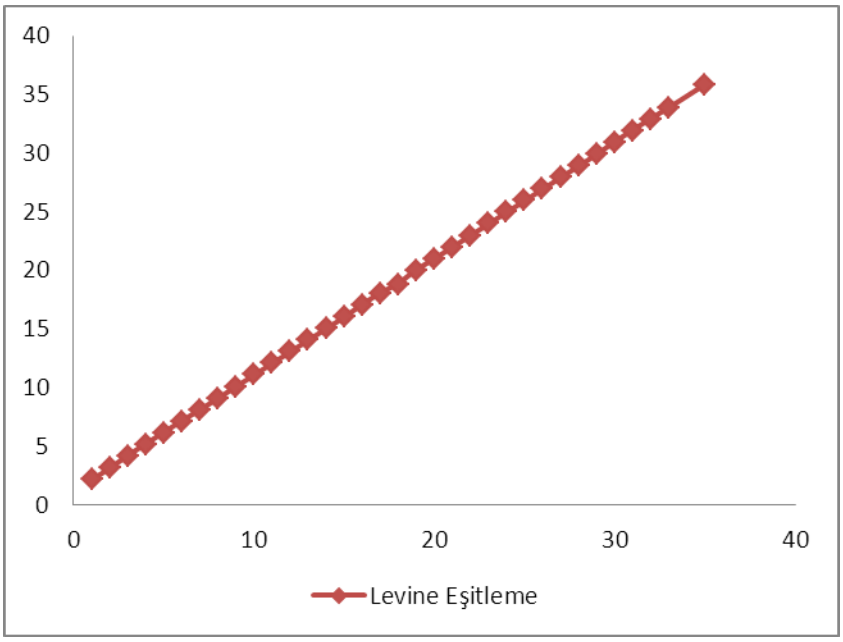
Appendix 1. RMSD (x) Values Regard to Equating Methods

Raw Score	Tucker Linear Equating	Levine Linear Equating
0	0,161	0,121
1	0,061	0,096
2	0,106	0,154
3	0,131	0,108
4	0,155	0,065
5	0,179	0,034
6	0,203	0,051
7	0,228	0,093
8	0,252	0,139
9	0,276	0,185
10	0,300	0,232
11	0,325	0,279
12	0,349	0,327
13	0,373	0,374
14	0,398	0,421
15	0,422	0,469
16	0,446	0,516
17	0,470	0,564
18	0,495	0,643
19	0,519	0,659
20	0,543	0,706
21	0,568	0,754
22	0,592	0,801
23	0,616	0,849
24	0,641	0,896
25	0,665	0,944
26	0,689	0,992
27	0,713	1,039
28	0,738	1,087
29	0,762	1,134
30	0,543	0,750
31	0,811	1,229
32	0,576	0,808
33	0,593	0,837
34	0,722	1,257
35	0,626	0,895
36	0,757	1,144

Appendix 2. The Graphics of Raw Score and Equated score Regard to Equating Methods



The Graphic of Raw Score and Equated score Regard to Tucker Linear Equating Method



The Graphic of Raw Score and Equated score Regard to Levine Linear Equating Method