International Journal of Assessment Tools in Education
2018, Vol. 5, No. 3, 417–427

DOI: 10.21449/ijate.409826

Published at http://www.ijate.net http://dergipark.gov.tr/ijate Research Article

A Comparison of Traditional and Kernel Equating Methods

Çiğdem Akın Arıkan [1], **Selahattin Gelbal** [2,*]

[1]Ordu University, Education Faculty, Measurement and Evaluation Department, 52200, Ordu, Turkey
[2]Hacettepe University, Education Faculty, Measurement and Evaluation Department, 06800, Ankara, Turkey

Abstract: In this study, the equated score results of the kernel equating (KE) method compared with the results of traditional equating methods—equipercentile and linear equating and 9th grade 2009 ÖBBS Form B of Social Sciences and 2009 ÖBBS Form D of Social Sciences was used under an equivalent groups (EG) design. Study sample consists of 16.249 students taking booklets B and another 16.327 students taking D in that test. The analysis of the test forms was carried out in four steps. First, descriptive statistics were calculated for the data and then it was checked whether the data obtained from the two booklets satisfy the equating conditions. In the second step, the booklets were equated according to methods. Lastly, the errors for each equating methods were calculated. Kernel equating results were nearly same to the results from the corresponding traditional equating methods. In Kernel equating, when parameter h was selected as optimal, equated scores provided almost identical results as traditional equipercentile equating. When it was selected large, this time the equated scores provided results almost identical to traditional linear equating. It is concluded that Kernel equating methods are relatively more the most appropriate equating method method than traditional equating methods.

ARTICLE HISTORY

Received: 26 March 2018
Revised: 13 May 2018
Accepted: 29 May 2018

KEYWORDS

Kernel equating,
Traditional equating Methods,
The equivalent groups design

1. INTRODUCTION

Comparison of test scores obtained from different forms has been a centre of attention for psychometrics for nearly a century. Further, discovery of methods for comparing scores is almost as old as the field of psychometrics itself (Holland, 2007). Discussion and development on the topic of comparison of scores were started in 1910s by Otis (1916, 1918), Kelley (1914), Starch (1913), Weiss (1914) and Pinter (1914) and have not come to an end up to day (Holland, 2007). In order to compare scores from different forms, various equating methods have been introduced which are based on different theories and statistics. The purpose of all equating methods is to compare scores collected from different test forms. Although different test forms are developed with similar content and statistics, they may vary in difficulty, so test equating is needed. The statistical process which allows comparing and interchangeable use of scores

CONTACT: Çiğdem AKIN ARIKAN ✉ akincgdm@gmail.com Ordu University, Education Faculty, Measurement and Evaluation Department, 52200, Ordu, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

obtained from different test forms is called equating (Holland, 2007; Kolen & Brennan, 2004). There are certain conditions that must be met for equating test forms. These conditions include equality, symmetry, group invariance, equal construct and equal reliability (Dorans and Holland, 2000).

- Equal construct: Both tests must be measure of the same characteristics, trait or skills.
- Equal reliability: The tests must have the same level of reliability.
- Symmetry: In equating function, form X can be transformed into form Y and vice versa at the same time.
- Equality: Lack of difference resulting of taking form X or Y of individuals.
- Group invariance: Equating relationship is independent of groups.

After meeting the equating conditions, test equating design is selected, one of the most important steps of test form equating. Equating designs are divided into three as single-group design, equivalent groups design, and common-item test design. In single-group design, the oldest and the simplest equating design, the same individuals are given both test forms. Since forms are answered by the same individuals, no error emerges due to ability levels of individuals (Kolen & Brennan, 2004). In equivalent groups design, the test forms are divided into two randomly and administered separately to the groups regarded equivalent. In this design, individuals in each group responds to only one test (Livingston, 2014). However, since each of the groups consists of different individuals, the difference of distribution of individuals' ability is reflected as bias in equating (Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). In common-item test design, two tests consist of different questions and groups are given either test which contain common items. On the common items in the forms, equating is performed. For this reason, common items should be selected in a way to represent the test properly (Kolen, 1988; Kolen & Brennan, 2004). Once the equating design is determined, decision is made regarding which equating method to use. Equating methods are divided as true-score equating and observed-score equating. While true-score equating relates to equating methods based on the Item Response Theory; whereas observed-score equating includes linear, average, equipercentile equating, IRT observed-score equating, and Kernel equating, a new approach. Present study was conducted with linear equating, equipercentile equating, Kernel linear equating and Kernel equipercentile equating methods among observed-score equating methods.

In linear equating, it is assumed that all properties except the mean and standard deviation of test forms are equal (Kolen & Brennan, 2004). In addition, the difference in difficulty between test forms varies in a constant amount through the entire score scale (Albano, 2016). Concerning X new form and Y reference (old) form, , μ_X ve μ_Y gives means of the forms, and S_X ve S_Y gives standard deviations of the forms. On this basis, linear equating is obtained from the following equation.

$$Y = \frac{S_Y}{S_X}x + \left[\mu_Y - \frac{S_Y}{S_X}\mu_X\right] \quad (1)$$

In the case of equipercentile equating, cumulative frequency of each test form is first calculated and then the scores corresponding to the same percentile based on the frequencies are equated. When the forms are equated with equipercentile equating, the scores obtained from the forms are similar in mean, standard deviation, and distribution (kurtosis, skewness, etc.) (Kolen, 1988). However, Livingston (2014) stated that, in equipercentile equating, equated points obtained from the new form have almost identical distribution with the points on the reference form, and that the distributions are not identical because the scores are discrete.

In equivalent groups design, both linear and equipercentile equating methods can be used. While equipercentile equating methods use score distribution curves to explain difficulty

difference between forms, linear equating methods use linear estimates. Hence equipercentile equating is more general than linear equating methods (Kolen & Brennan, 2004).

As for the Kernel equating, an observed-score equating method, was defined by Holland and Thayer (1989) and improved by von Davier, Holland and Thayer (2004). In traditional equipercentile equating methods, cut-off score distribution is made continuous by using linear estimates. On the other hand, Kernel equating employs the Gauss Kernel approach after which it is also named. In the latter, discrete distributions are made continuous so that scores are equated on the basis of the continuous distributions (Lee & von Davier, 2011; Ricker & von Davier, 2007). Kernel equating also includes linear equating functions even though it is an equipercentile equating function (von Davier et al. 2004). One of the important parameters in the Kernel is the parameter h, which is the continuation parameter. If the parameter h is chosen as ideal, equipercentile equating is obtained; but if it is selected as large, linear equating function is obtained. In addition, the results obtained through the ideal h parameter approach traditional linear equating, while those obtained through the large h parameter approach traditional equipercentile equating.

In the Kernel model, test forms are equated in five steps:

Presmoothing: It refers to using the log-linear statistical model for smoothing of score distributions. In this step, estimation of score probabilities varies depending on the score equating design. Equivalent groups design is a univariate distribution; however, common-item test design is a bivariate distribution in nonequivalent groups. von Davier et al. (2004) indicated four statistical properties in selection of estimating point probabilities as;

- **Consistency;** as the sample size increases, estimated values approach the population parameter.
- **Efficiency;** deviation of the score probabilities estimated from the population values is at the minimum level possible.
- **Positivity;** score probabilities estimated for each score are positive.
- **Integrity;** smoothed score distributions match with observed score distribution. To get good fit in univariate distributions, five or six moments of test forms must be used (von Davier et al., 2004).

Estimation of score probabilities according to the equating design Gauss Kernel approach is used to make the cut-off score distributions continuous at the relevant stage. Still, Lee and von Davier (2011) suggested logistics and uniform kernel approaches as alternatives. Test forms are equated by using the continuous distributions obtained in the previous step. It is calculation of the standard error of equating and difference of the standard error of equating. Kernel equating can be used in common-test designs with single-group, equivalent groups, balanced group, and nonequivalent groups (von Davier et al., 2004).

It is used for not only international tests such as PISA and TIMMS but also nation-wide examinations held by the Ministry of National Education and Student Selection and Placement Centre in Turkey. Student Achievement Determination Exam at National Scale (SADE-ÖBBS) is among them. ÖBBS has taken every three years since 1992 at elementary education. It applies to different areas and grade levels. Then, the applications were reported in 2002, 2005 and 2008, respectively. ÖBSS helps determine the adequacy levels of the education and instruction environment offered to students, objectives and skills and make an assessment so as report to relevant authorities accordingly (EARGED, 2010). It was last held in year 2008 to determine achievement levels of elementary students in subjects such as Turkish language, science and technology, mathematics, social studies and English language. Then, in 2009, the Secondary Education Development Project was launched to identify achievements of students in lessons, monitor their progress, and propose recommendations according to results. The

project was implemented on the 9th and 10th grade students (EARGED, 2010). In this scope, in order to ensure test security in practice and to avoid violations of the rules; booklets A, B, C and D were prepared which are very similar in both scope and difficulty. However, these booklets were not equated although they consist of different questions. Even though test developers introduce test forms similar in content and statistical aspects, these forms may differ in difficulty.

This study was conducted on the scores obtained from the ÖBBS given to 9th graders in 2009. In particular, the scores obtained from social studies test in booklet B were equated to booklet D by using linear equating, equipercentile equating, Kernel linear equating and Kernel equipercentile equating methods. It was aimed to find out the best method as a result. The other aim of the study is to compare results obtained from the Kernel equating and traditional equating methods.

2. METHOD

2.1. Population and Sample

The population of the study is composed of the elementary 10 th grade students who took the 2009 ÖBBS covering Turkish Literature - Language and Expression, Mathematics, Sciences, Social Studies and English Language tests. There are four booklets as A, B, C and D, for each of the tests in ÖBSS. The booklets were put into two pairs by arranging A and C, and B and D in parallel, respectively (EARGED, 2010). Study sample consists of 16.249 students taking booklets B and another 16.327 students taking D in that test.

Data is constituted by scores obtained from social studies test as a part of ÖBBS held in 2009. The test contains 15 (fifteen) questions which target history and geography lessons. Each booklet contains different questions in ÖBBS. So, there are two pairs of booklets considering the traits they measure resulting in booklets A and C as a pair and B and D as another pair. This study was carried out on the latter pair of booklets, B and D.

2.2. Equating Design and Data Analysis

Despite containing different questions, the booklets were prepared in parallel and applied spirally to the students. For this reason, equivalent groups were formed for B and D booklets at random. In this study, the social studies test was equated through the use of the equating pattern for equivalent groups.

The analysis of the test forms was carried out in four steps. First, descriptive statistics were calculated for the data and then it was checked whether the data obtained from the two booklets satisfy the equating conditions. In the second phase, the two booklets were equated according to the equivalent group design and equated according to the methods. In the second step, the booklets were equated with equivalent groups design and equated socres were gained accordingly. Lastly, the amount of error resulting from each equating method was calculated. Analysis of the data was done with SPSS and FACTOR (Lorenzo-Seva & Ferrando, 2006). The equate R package (Albano, 2016) was used for traditional equating methods analyses and the kequate R package (Andersson, Branberg & Wiberg, 2013) was used for kernel equating methods analyses (R Core Team, 2017).

Step I: Descriptive Statistics and Testing of Equating Conditions

At first, descriptive statistics of the social studies tests were calculated. The findings are given in Table 1. According to Table 1, for the booklets B and D, score distributions exhibit positive coefficients of skewness below 1. On the other hand, kurtosis and skewness coefficients are negative and smaller than 1 in score distribution of both forms. Büyüköztürk (2007) stated that kurtosis and skewness coefficients between -1 and +1 refer to normal distribution.

Table 1. Descriptive statistics

Descriptive Statistics	Booklet B	Booklet D
N	16249	16327
Mean	7.60	7.97
Standard Deviation	3.50	3.52
Variance	12.28	12.45
Skewness	.172	.050
Skewness Standard Error	.019	.019
Kurtosis	-.81	-.89
Kurtosis Standard Error	.038	.038

Test forms can be equated provided that certain requirements are satisfied. Dorans and Holland (2000) list these requirements as unidimensionality, equal reliability, and similar difficulty. To confirm unidimensionality of the data, principal components factor analysis based on the tetrachloric correlation used for two-category data was performed. The factor analysis was conducted with FACTOR 10.7 (2017) program.

Table 2. Factor Analysis Results for Booklets B and D

	Booklet B		Booklet D	
Component	Eigenvalue	V.A.O (%)	Eigenvalue	V.A.O (%)
1	4.91	32.7	5.02	33.5
2	0.96	6.4	0.99	6.6

In Table 2, when factors with eigenvalue greater than 1 are taken to calculate the number of factors, there is only one factor with eigenvalue greater than 1 for both booklets. Explanatory variance of the first factor is 32.7% for booklet B, while it is 33.5% for the other booklet. In single-factor scales, explained variance ratio at and above 30% is regarded adequate (Büyüköztürk, 2007). So, both booklets can be said to have one single general factor.

In order to test whether reliability of booklets B and D is equal, the reliability coefficient of KR-20 was calculated. Again, reliability coefficients were accepted as the correlation coefficients and Fischer's Z transformation was performed to check if there is a difference between the two reliability coefficients (Akhun, 1984). The results are presented in Table 3.

Table 3. Comparison Results of Reliability of Booklets

Booklet	KR-20	Zr	Z	p
B	0.75	0.99	1.29	0.51
D	0.76	1.00		

Table 3 suggests that the booklets B and D meet the prerequisite of equal reliability ($p>.05$). The difference between average difficulties of the booklets B and D was examined with two ratio difference test (Baykul, 1996), indicating no significant difference between difficulty levels of the two booklets ($p>.05$). It can be suggested that both booklets are equal in average difficulty.

Step II: Booklet equating: The booklets B and D were equated by using linear and equipercentile equating, Kernel equating method (linear and equipercentile) among traditional equating methods in equivalent grups design.

Step III: At this stage, the weighted error squares mean (WMSE) and RMSD (Root Mean Squared Difference) were calculated for evaluating the errors randomly involved in the

equating process and MSD (Mean Signed Difference) indices were calculated for evaluating the systematic errors.

Below are shown the equations for calculating WMSE, RMSD and MSD coefficients:

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{k-1} fi(X_E - X_{Crit})^2}{\sum_{i=1}^{k} f_i}} \quad (2)$$

$$\text{WMSE} = \frac{\sum_{i=1}^{k-1} fi(X_E - X_{Crit})^2}{\sum_{i=1}^{k} f_i \, S^2 y} \quad (3)$$

$$\text{MSD} = \frac{1}{N} \sum_{j=1}^{N} (X_E - X_{crit}) \quad (4)$$

3. FINDINGS

In 2009 ÖBBS social studies test, booklet B was equated with booklet D by using linear equating, equipercentile equating, Kernel linear equating and Kernel equipercentile equating methods. The parameter h, which is the continuation parameter in the kernel equating methods, was selected by the kequate package. It was found to be hx=0.539 and hy=0.538 for Kernel equipercentile equating method; and hx=3503.590 and hy=3530.941 for Kernel linear equating method. Equated scores obtained from the equating methods and raw scores are given in Table 4.

Table 4. Equivalent scores of Booklet D corresponding to raw scores in Booklet B

Booklet B	LE	Dif.	EE	Diff.	KE-LE	Diff.	KE-EE	Dif.
0	0.31	0.31	0.01	0.01	0.31	0.31	0.02	0.02
1	1.32	0.32	1.12	0.12	1.32	0.32	1.11	0.11
2	2.33	0.33	2.19	0.19	2.33	0.33	2.18	0.18
3	3.34	0.34	3.24	0.24	3.34	0.34	3.23	0.23
4	4.35	0.35	4.29	0.29	4.34	0.34	4.28	0.28
5	5.35	0.35	5.31	0.31	5.35	0.35	5.31	0.31
6	6.36	0.36	6.41	0.41	6.36	0.36	6.40	0.40
7	7.37	0.37	7.50	0.50	7.37	0.37	7.49	0.49
8	8.38	0.38	8.53	0.53	8.37	0.37	8.52	0.52
9	9.38	0.38	9.54	0.54	9.38	0.38	9.54	0.54
10	10.39	0.39	10.53	0.53	10.39	0.39	10.53	0.53
11	11.40	0.40	11.45	0.45	11.4	0.40	11.47	0.47
12	12.41	0.41	12.35	0.35	12.41	0.41	12.35	0.35
13	13.41	0.41	13.21	0.21	13.41	0.41	13.21	0.21
14	14.42	0.42	14.10	0.10	14.42	0.42	14.11	0.11
15	15.43	0.43	15.02	0.02	15.43	0.43	15.04	0.04

**Dıff: difference; LE: Linear Equating; EQ: Equipercentile Equating; KE-LE: Kernel Linear Equating; KE-EQ: Kernel Equipercentile Equating*

According to Table 4, the raw scores belonging to Booklet B take values in the range of 0-15 points, but the equated scores obtained through traditional linear equating and Kernel linear equating methods vary from 0.31 to 15.43. The equated scores obtained from both methods demonstrate that the equated scores are the same except for a few conditions. As a result of the linear equating methods, the scores equated throughout the entire score scale were

found larger than the raw scores and difficulty level did not vary throughout the scale. So, it can be said that booklet B is more difficult than D. The equated scores obtained from traditional equipercentile equating method were found to be between 0.01 and 15.02, while Kernel equipercentile equating method yielded results in the range of 0.02 to 15.04. The two methods were seen to generate the same equated scores except in a few conditions. It was found out that the scores equated with equipercentile equating methods were greater than the raw scores throughout the entire score scale and difficulty level did not vary throughout the scale. This finding suggests that booklet B is more difficult than booklet D according to equipercentile methods. Moreover, a linear relationship was detected between raw scores from booklet B and equated scores obtained through both linear equating methods and equipercentile equating methods. Such relationships are shown in Figure 1 and Figure 2.

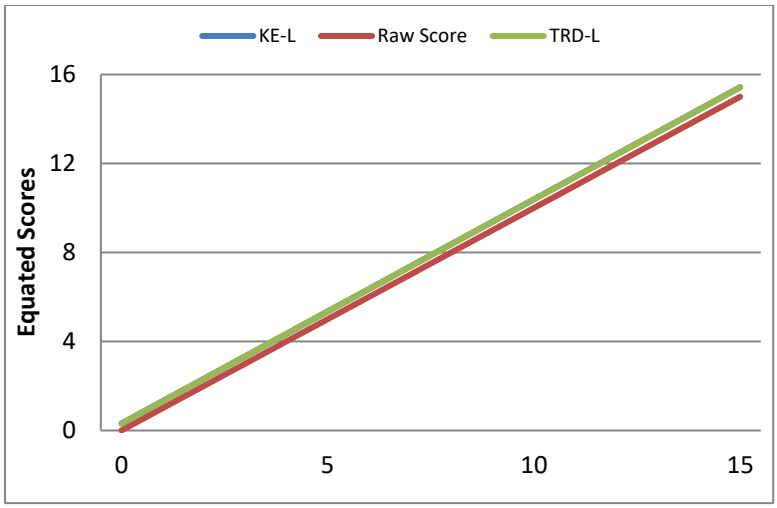

Figure 1. Raw scores in booklet B and equated scores based on linear equating methods

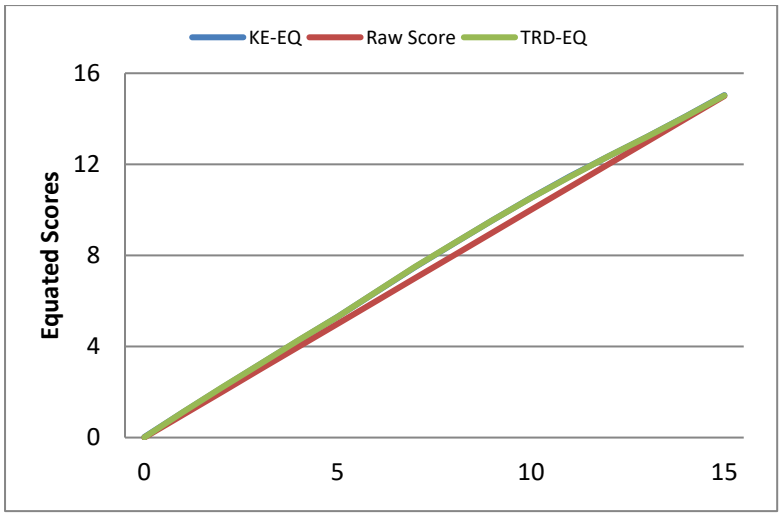

Figure 2. Raw scores in booklet B and equated scores based on equipercentile equating methods

The graphs above indicate almost the same difference between the equated scores based on linear equating methods and raw scores throughout the whole scale. In the case of equipercentile equating methods; while the difference between equated scores and raw scores is smaller in extreme scores, the difference increases in in the medium score scale (in the range of 5 to 12). One superiority of Kernel equating to traditional equating methods is that it

calculates the standard error of equating for each score. Figure 3 displays the graph for the standard error of the equating obtained over the entire scale score according to Kernel equipercentile and Kernel linear equating methods.

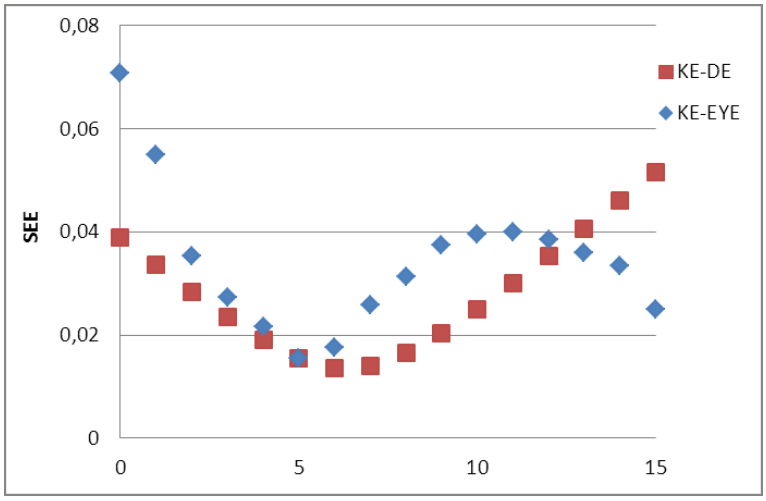

Figure 3. Standard errors from Kernel linear and equipercentile equating methods

Figure 3 shows that linear equating methods yield higher standard errors in marginal ends but lower standard errors in intermediate band. In the case of equipercentile equating method, standard error proved high in lower end but low in higher end. Also in linear equating, it has a decreasing tendency from 0 to 7 points but increasing from 7 to 15 points. In equipercentile equating, it again shows a decreasing tendency from 0 towards 5 points while increasing from 5 towards 11 and then falling from 11 to 15 back. The average standard error was calculated as 0.035 and 0.028 for Kernel equipercentile equating and linear equating, respectively, which shows a smaller error in linear equating method.

The booklets B and D for social studies test from 2009 ÖBBS were were equated with linear and equipercentile equating methods. WMSE, MSD and RMSD were calculated to find out which of the equating methods includes a lower level of error. The obtained values are given in Table 5.

Table 5. WMSE, RMSD and MSD values from the equating methods

Equating Method	WMSE	RMSD	MSD
GE-EYE	0.013	0.400	0.373
GE-DOĞ	0.011	0.377	0.376
KE-EYE	0.012	0.399	0.372
KE-DOĞ	0.011	0.375	0.374

Table 5 shows the smallest WMSE and RMSD values as a result of Kernel linear equating method, while the largest WMSE and RMSD values are given by traditional equipercentile equating method. In addition, in terms of WMSE and RMSD values while traditional linear equating and Kernel linear equating provided similar results; traditional equipercentile equating and Kernel equipercentile equating methods provided nearly same results. Apart from that, MSD index referring to systematic error indicates that Kernel equipercentile equating has the smallest error while the opposite is reported by traditional linear equating method. In consideration of all findings here, the smallest random error was reached with Kernel linear equating method and the smallest systematic error was achieved by Kernel equipercentile equating method.

4. DISCUSSIONS AND CONCLUSION

In this study, we compared the results obtained through traditional equating methods with Kernel equating with the intention of finding out the superior one.

It was found out that Kernel equating methods as a recently introduced approach produced results comparable to traditional equating methods. In Kernel equating, when parameter h was selected as optimal, equated scores provided almost identical results as traditional equipercentile equating. When it was selected large, this time the equated scores provided results almost identical to traditional linear equating. These results seem to be in accord with Mao, von Davier and Rupp (2006), von Davier et al (2006) and Grant, Zhang and Damiano (2009). In their study, Mao et al. (2006) compared traditional equating methods and Kernel equating methods in equivalent groups and nonequivalent groups in common-item test pattern, and noted quite similar results particularly as a result of linear equating in equivalent groups design. In another sample, von Davier et al. (2006) compared Tucker, chained equating, frequency estimation, Levine observed-score equating and post-stratification equating from traditional equating methods against Kernel chained equating (h optimal), Kernel chained equating (h large), and Kernel post-stratification (h large) from Kernel equating methods. They reported comparable results in both Kernel and traditional equating methods**.** Furthermore, Grant et al. (2009) comparing performances of Kernelchained equating, Kernel post-stratification, Tucker, chained equipercentile and Levine equating methods found out that Kernel chained equating yield similar results to chained linear equating method with large bandwidth. In addition, they came up with minor differences only between Kernel post-stratification equating and Tucker and Levine equating results.

In relation with error; comparison of WMSE, RMSD, and MSD values from Kernel equipercentile, Kernel linear equating, traditional linear and equipercentile equating methods recorded the smallest random error with linear equating methods applying to both traditional and Kernel equating methods. The smallest systematic error was found in Kernel equating methods. In particular, Kernel linear equating method generated the lowest random error, while the highest level of the same type of error was generated by traditional equipercentile method. It can be argued that our results show similarity with the literature. Kelecioğlu and Öztürk Gübeş (2013) carried out equating on 2009 ÖBBS social studies test A and C books according to linear and equipercentile equating with random group design and found out that linear equating involves the smallest random error. In a comparative study by Zhu (1998) on RMSD and MSD coefficients from linear equating and unsmoothed and postsmoothed equipercentile equating methods, it was found out that linear equating method involves the least random and systematic error but unsmoothed equipercentile equating show the most random and systematic errors. Yet, our findings seem to be in dispute with Zhu (1998) in relation with systematic error because Kernel equipercentile equating and traditional linear equating methods yielded the smallest and the largest random error, respectively. The result does not coincide with the findings by Zhu (1998).

When the equated scores obtained from the linear equating methods are examined, it is seen that the scores take values out of the raw score scale range. Kolen and Brennan (2014) suggested that it is an expected effect and Livingston (2014) argued that it is peculiar to linear equating methods. Kolen and Brennan (2014) proposed two alternative ways in this case. The first is to allow the points that are not in the raw score range. The second is to accept the scores outside the raw score range as the lowest and highest raw scores. In other words, it refers to taking all equated points below as 0 and taking all equated points above 15 as 15. Concerning equipercentile equating method, Kolen and Brennan (2014) claim that equated scores might deviate from the raw score range by -.5 up to +5, which is a desirable feature of equipercentile equating. That is to say, the points obtained with equipercentile equating methods need to be

valued between -0.5 and 15.5. Considering this range for both Kernel and traditional equipercentile equating, both methods seem to be in the range determined.

We also found out that the standard error obtained from Kernel equipercentile equating method is greater than Kernel linear equating and it tends to increase in marginal points. This seems to be consistent with findings from other studies in the literature. In a study by Choi (2009) comparing Kernel equating and traditional equating methods, Kernel linear equating methods were found to give lower standard error compared to Kernel equipercentile equating methods. According to Mao (2006), higher errors in end points with Kernel equating methods are due to the fact that the point scale in Gauss Kernel continuation method falls in the range of $+\infty$ and $-\infty$.

In the light of the all findings, it is understood that Kernel linear equating method has the least random error and Kernel equipercentile equating has the least systematic error in equating of booklets B and D of social studies test from 2009 ÖBBS. Yet, the error values obtained seem to be very close, which may be due to similar distribution of the test forms. Livingston (2014) contended that if the new and old forms exhibit the same distribution of points, linear and equipercentile equating methods would yield almost identical results and even equated scores could overlap.

Departing from the discussion above, it was concluded that Kernel equating methods are more suitable than others for equating booklets B and D of social studies test in the context of 2009 ÖBBS. Present study was planned to compare equatings of the foregoing documents through the use of traditional and Kernel equating methods. In the future, a similar study could be conducted in other equating designs and methods (equating methods based on Item Response Theory, local equating, etc.). Besides RMSD, MSD and WMSE, other evaluation criteria such as invariance indices and DTM could be employed in future studies as well.

ORCID

Çiğdem AKIN ARIKAN https://orcid.org/0000-0001-5255-8792
Selahattin GELBAL https://orcid.org/0000-0001-5181-7262

5. REFERENCES

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

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, *74*(8), 1-36.

Andersson, B., Branberg, K., & Wiberg, M. (2013). Performing the Kernel Method of Test Equating with the Package kequate. *Journal of Statistical Software*, *55*(6), 1–25.

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

Büyüköztürk, Ş. (2007). *Sosyal bilimler için veri analizi el kitabı* (8.Baskı). Ankara: Pegem A Yayıncılık.

Choi, S. I. (2009). *A comparison of kernel equating and traditional equipercentile equating methods and the parametric bootstrap methods for estimating Standard errors in equipercentile equating*. Unpublished doctoral dissertation.University of Illinois at Urbana-Champaign.

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

Eğitim, Araştırma ve Geliştirmesi Daire Başkanlığı (EARGED). (2010). Ortaöğretim ÖBBS raporu 2009. Ankara, Milli Eğitim Bakanlığı.

Grant, M. C., Zhang, L., & Damiano, I. (2009). An Evaluation of Kernel Equating: Parallel Equating With Classical Methods in the SAT Subject Tests™ Program. *ETS Research Report Series, 2009* (1), i-25.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston: Academic Puslishers Group.

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

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.

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

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

Lee, Y. H., & von Davier, A. A. (2011). Equating through alternative kernels. In von Davier, A. (Ed.) *Statistical models for test equating, scaling, and linking* (pp. 159-173).Springer New York.

Livingston, S. A. (2014). Equating test scores (without IRT), (2nd. ed.). *Educational testing service.*

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*(1), 88-91.

Mao, X. (2006). *An investigation of the accuracy of the estimates of Standard errors for the kernel equating functions*. Unpublished doctoral dissertation, The University of Iowa.

Mao, X., von Davier, A. A., & Rupp, S. (2006). Comparisons of the Kernel Equating Method with the Traditional Equating Methods on Praxis™ Data. *ETS Research Report Series, 2006(2).*

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Ricker, K. L.,& Davier, A. A. (2007). The impact of anchor test length on equating results in a nonequivalent groups design. *ETS Research Report Series, 2007(2).*

von Davier, A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). An Evaluation of the Kernel Equating Method: A Special Study with Pseudo tests Constructed from Real Test Data. *ETS Research Report Series, 2006(1).*

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

Zhu, W. (1998). Test equating: What, why, how?. *Research Quarterly for Exercise and Sport, 69*(1), 11-23.