Eurasian Journal of Educational Research 82 (2019) 27-44

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

A Comparison of Kernel Equating Methods Based on Neat Design*

Cigdem AKIN ARIKAN[1]

ARTICLE INFO

Article History:
Received: 12 Nov. 2018
Received in revised form: 22 Apr. 2019
Accepted: 29 May. 2019
DOI: 10.14689/ejer.2019.82.2

Keywords
Equating, equipercentile, linear, RMSD SEED, SEE

ABSTRACT

Problem Statement: Equating can be defined as a statistical process that allows modifying the differences between test forms with similar content and difficulty so that the scores obtained from these forms can be used interchangeably. In the literature, there are many equating methods, one of which is Kernel equating. Trends in International Mathematics and Science Study (TIMSS) aims to find out the knowledge and skills gained by the fourth and eighth-grade students in the fields of mathematics and science. TIMSS have different test forms, and these forms are equated through common items.

Purpose of the Study: This research aimed to compare the equated score results of the Kernel equating (KE) methods, which are chained, and post-stratification equipercentile and linear equating methods under NEAT design.

Methodology: TIMMS Science data were used in this study. The study sample consisted of 865 eighth-grade examinees who were given the Booklets 1 and 14 during the TIMSS application in Turkey. There were 39 items in Booklet 1, and 38 items in Booklet 14. Firstly, descriptive statistics were calculated and then the two Booklets were equated according to NEAT design based on Kernel chained, Kernel post-stratification equipercentile, and linear equating methods. Secondly, the equating methods were evaluated according to some criteria such as DTM, PRE, SEE, SEED, and RMSD.

Findings and Results: It was seen that results based on equipercentile and linear equating methods were consistent with each other, except for a high range of the score scale. PRE values demonstrated that KE equipercentile equating methods better matched with the discrete target distribution Y, and distribution of SEED revealed that KE equipercentile and linear methods were not significantly different from each other according to DTM.

* This study was presented at the 6th International Congress on Measurement and Evaluation in Education and Psychology in Prizren-Kosova, 5-8 September 2018.

[1] Ordu University, Faculty of Education, Ordu, TURKEY, e-mail: akincgdm@gmail.com, ORCID: https://orcid.org/ 0000-0001-5255-8792

Introduction

Equating can be defined as a statistical process that allows modifying the differences between test forms with similar content and difficulty so that the scores obtained from these forms can be used interchangeably (Kolen, 1988). For about 100 years, equating methods have attracted the attention of psychometrics and the development of new methods has not stopped. Equating methods include methods based on equipercentile equating, linear equating methods, IRT observed-score and true score equating, van der Linden local equating, Levine nonlinear method, and Kernel equating (von Davier, 2013). As for the Kernel equating, an observed-score equating method was defined by Holland and Thayer (1989) and then 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 Gaussian 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). KE is a flexible family of equipercentile-like equating functions that include the linear equating function as a special case (von Davier, Holland & Thayer 2004).

In the Kernel equating model, test forms are equated in five steps: presmoothing, estimation of score probabilities, continuization, equating, and standard error of equating. The first step is presmoothing that refers to using the log-linear statistical model for smoothing of score distributions. The goal of presmoothing is to achieve decreased sampling errors. In this step, the 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 the 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).

The second step is the estimation of score probabilities of X and Y scores according to the equating design that is obtained from step one. The third step is continuation where Gaussian Kernel approach is used to make the cut-off score distributions continuous at the relevant stage. In this step, the choice of bandwidths is essential. Von Davier et al. (2004) suggest the penalty function to automatically select the bandwidths. In addition to Gaussian Kernel approach, Lee and von Davier (2011) recommend logistics and uniform kernel approaches as alternatives. The fourth step

is equating. When the first three steps are done, test forms are equated by using continuous distributions. The last step is the standard error of equating (SEE). SEE is dependent on presmoothing, computing r and s from the smoothed data and equating function (von Davier et al., 2006).

Kernel equating can be used in single-group, equivalent groups, and non-equivalent groups (von Davier et al., 2004). Non-Equivalent groups Anchor Test-NEAT is used when the test form is applied more than once due to test safety. In NEAT design, both forms have common items and equating the relationship between the test forms is established through common items (Kolen & Brennan, 2004). In Kernel equating in NEAT pattern; Post- stratification (PSE), Levine observed-score linear, and Chained Equating (CE) methods are used (von Davier et al., 2004). In NEAT pattern, two different groups take two different test forms (X and Y) and the common test form (A). PSE uses the common test form to estimate the distribution of test forms across a group I and group II. In CE, the common test is used as a chain and the test form X is first connected to the common test form for group I. Then the common test form is connected to the version Y for group II (von Davier et al., 2004). Kernel equating includes both linear and equipercentile equating functions by manipulating bandwidths. If optimal bandwidths are selected, KE approximates the equipercentile equating function, and if large bandwidths are selected, KE approximates the linear equating (von Davier et al., 2006). The equating methods used in this paper are given in Table 1.

Table 1

Equating Methods

Linear	PSE-with large bandwidths
	CE –with large bandwidths
Equipercentile	PSE-with optimal bandwidths
	CE- with optimal bandwidths

One of the criteria for determining which method performs better in equating is the error. The equating method with a smaller rate of error can be said to be more appropriate. Furthermore, KE provides some measures, percent-relative error (PRE) and standard error of equating difference (SEED) when evaluating the equating results. PRE is a tool that assesses how well an equating function matches the discrete target distribution Y. SEED can be defined as a difference between the two equating functions and the range of ±2 SEED shows that the differences are because of sampling variability (Liu & Low, 2007). The equating methods are evaluated according to certain criteria: Difference That Matter (DTM), PRE, standard error of equating (SEE), SEED, and Root Mean Squared Difference (RMSD).

DTM: DTM is used to evaluate the difference between equated scores obtained from two distinct equating functions. Despite not being an established rule, it is generally determined to be .5, which is half of the raw point unit (1). If the difference between the two equated scores is less than .5, the scores are regarded similar; if the difference is bigger than .5, the equated scores are considered distinct (Holland & Dorans, 2006).

PRE: The percent-relative error (PRE) is a tool that compares the distribution of Y with the equated values, eY(X) and assesses how well an equating function matches the discrete target distribution Y (Von Davier et al., 2004). The PRE is calculated by the following formula.

$$PRE(p) = 100\frac{\mu_p\,(eY\,(X)) - \mu p(Y\,)}{\mu p(Y\,)} \qquad (1)$$

KE compares the first 10^{th} moment of Y and eY(X). If continuization step has been done cautiously, then the PRE values are frequently small (von Davier et al., 2004).

SEE: In Kernel equating, standard error of equating depends on three factors. The first is the combination of pre-smoothing, the second is the computation of smoothed data, and the third is the mathematical form of the smoothing and equating function (von Davier et al., 2006).

SEED: It is used to determine the accuracy of the difference between the two equating functions and suggest which synchronization function is more appropriate. SEED is also used to choose either linear or non-linear equating functions (Von Davier et al., 2004). Furthermore, ±2 SEED band is available in order to determine how the two equating functions vary depending on sample variability (Von Davier et al., 2004). If the variance between equating functions does not exceed the ± 2 SEED range, this means that the variance is due to sampling error (Liu & Low, 2007).

$$SEED_Y(x) = \sqrt{Var\left(\hat{e}_1(x) - \hat{e}_2(x)\right)} \qquad (2)$$

RMSD: Equating error is used to define the accuracy of equating. RMSD coefficient is used for the equating error.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{k-1} fi(X_E - X_{Crit})^2}{\sum_{i=1}^{k} f_i}} \qquad (3)$$

Xcrit: The raw score number i in test D

XE : The score obtained with equating methods and equal to the raw score number i in test X

fi: The frequency of the raw score number i in test D

Purpose of the Study

International tests applied in Turkey include TIMMS (Trends in International Mathematics and Science Study), PISA (Programme for International Student Assessment), and PIRLS (Progress in International Reading Literacy Study). TIMSS is a test held every four years since 1995, but Turkey did not participate in 1995 and 2003. TIMSS aims to find out the knowledge and skills gained by the fourth and eighth-grade students in the fields of mathematics and science (MEB, 2016). TIMSS 2015 Turkey test consists of 24 blocks with 14 test booklets for science and mathematics. The 24 blocks were placed in 14 test booklets, two in science and two in mathematics, and one of two blocks in science and mathematics is common to two of the booklets (MEB, 2016). To compare trends between the years, TIMSS assessments were converted into the same metrics. For that, TIMMS uses item response theory (IRT) scaling with concurrent calibration (Mullis, Martin & Foy, 2016). However, it is of great importance which equating method is chosen. For the purpose of the test, the equating method should be determined by taking into account the strengths and weaknesses of the methods. It is needed because the choice of an inappropriate equating method increases the equating errors, leading to unfair decisions. KE methods can be used especially when IRT (true score) equating methods are not favorable (Godfrey, 2007; Meng, 2012; Norman Dvroak, 2009). TIMMS didn't use Kernel equating methods for converting the scores into the same metrics. In Turkey, several tests such as KPSS and ALES hold different validity of periods and project subjects take different test forms in Measurement and Evaluation of Academic Skills (MEAS-ABIDE). Since test forms must be equated in order to compare or use the scores interchangeably, several studies have been used Kernel equating (Choi, 2009; Grant, Zhang & Damiano, 2009; Godfrey, 2007; Holland, von Davier, Sinharay & Han, 2006; Mao, 2006; Mao, von Davier & Rupp, 2005; Meng, 2012; Moses & Holland, 2007; Norman Dvorak, 2009; Ricker & von Davier, 2007; von Davier et al., 2006). When the literature is examined, it is seen that articles about Kernel equating are very limited in Turkey (e.g. Akın Arıkan, 2017; Akın Arıkan & Gelbal, 2018). Therefore, it is thought that this study will contribute to the other studies which can use KE when the assumptions of IRT equating methods are not meet.

The main purpose of this study was to compare Kernel equating methods with real data under NEAT design based on equipercentile and linear methods so as to detect the most appropriate equating method. For this main purpose, research questions were as follows:

1) What is the relationship between raw scores and equivalent scores obtained from different equating methods?

2) How do PRE, DTM, SEE, SEED and RMSD values differ according to equating methods?

3) Which is the best Kernel equating method to equate TIMSS science subtests under NEAT design?

Method

Research Design

In this study, TIMSS 2015 science tests (Booklet number 1 and 14) were equated with Kernel equating methods and the obtained equating results were compared with each other. In terms of this, this research was a descriptive study.

Research Sample

During the period when the TIMSS 2015 research was conducted, there were a total of 1,108,572 students at the 4th grade and another 1,187,893 students at the 8th grade in Turkey. Out of the population; 6456 of 4th graders and 6079 of 8th graders participated in the TIMMS application (MEB, 2016). The study sample consisted of 865 eighth-grade examinees who were given the Booklets 1 and 14 during the TIMSS application in Turkey.

Research Instrument and Procedures

For data analysis, the data set was used consisting of the pattern of responses given by the 8th-grade examinees to science literacy items in the TIMSS 2015 Turkey. In this study, the items in Booklet number 1 and 14 were used among fourteen booklets included in the TIMMS application. There were 39 items in Booklet number 1, and 38 items in Booklet number 14. The wrong and missing values were coded as 0 and the partial credit scores and all the correct answers were coded as 1 yielding the final data for analysis.

Data Analysis

The booklets were equated according to the methods of Kernel CE and Kernel PSE. The kequate package (Andersson, Branberg & Wiberg, 2013) was used for kernel equating methods analyses (R Core Team, 2017).

Results

In the first phase of data analysis, descriptive statistics were calculated and the findings are presented in Table 2.

Table 2

Raw Score Descriptive Statistics of Booklet 1 and Booklet 14

Descriptive Statistics						
TEST	N	Mean	Std. Dvt.	Variance	Skewness	Kurtosis
K1	435	18.58	7.58	57.38	0.20	-0.63
Anchor-K1	435	5.84	3.48	12.10	0.44	-0.30
K14	430	12.74	4.65	21.60	-0.15	-0.67
Anchor-K14	430	6.17	3.71	13.78	0.42	-0.73

Table 2 shows the mean and standard deviation values for both booklets according to the total tests and anchor tests. Anchor test of booklet 14 mean scores were higher than the anchor test of Booklet 1 mean scores. Moreover, since the skewness coefficient of score distribution in Booklet 1, the common test of Booklet 1, and the common test of Booklet 14 was positive, the distribution seemed to be skewed to the right of what was normal. In addition, since the skewness coefficient of score distribution of Booklet 14 was negative, it can be said that the distribution was skewed to the left than normal. It can be suggested that the distributions had kurtosis compared to normal because the kurtosis coefficients of score distribution of both forms were negative.

The bandwidths values were automatically calculated by kequate package. The obtained values for KE PSE equipercentile (PSE EQ) method were .6327 for hX and .6318 for hY; for KE PSE linear (PSE L) method, it was 7611.23 for hX and 7306.05 for hY. As for KE CE equipercentile (CE EQ) equating, the values are .633 for hX and .6322 for hY. Finally, 7575.07 for hX and 7342.40 for hY in KE CE linear (CE L) method. Table 3 displays PRE values for KE PSE and KE CE (equipercentile and linear) equating methods.

Table 3

The PRE Values for the KE Optimal and KE Linear for Equating X to Y

Pth Moment	**Post- stratification Equating (PSE)**		**Chained Equating (CE)**			
	PRE EQ	PRE L	CE EQ		CE L	
			X to A1	A1 to Y1	X to A1	A1 to Y1
1	0.000	0.000	-0.010	0.257	0.000	0.000
2	-0.001	0.000	-0.002	-0.443	0.000	0.000
3	-0.007	-0.217	-0.108	-0.144	-8.243	1.633
4	-0.018	-0.646	-0.087	-0.165	-10.372	3.564
5	-0.035	-1.281	0.014	-0.608	-10.577	4.516
6	-0.059	-2.112	0.191	-0.751	-7.683	3.822
7	-0.092	-3.126	0.439	-0.801	-2.471	1.328
8	-0.134	-4.309	0.758	-0.820	5.244	-2.795
9	-0.188	-5.645	0.845	-0.879	15.486	-8.221
10	-0.254	-7.117	0.902	-0.970	28.531	-14.573

PRE = Percent relative error, EQ= Equipercentile

Table 3 indicates that the PRE values stated a good match for PSE and CE equipercentile equating methods but a poorer match for both KE linear equating methods between the equating function computed at the discrete values of X and the

target distribution of Y. Both equipercentile and linear equating PRE(p) values for PSE were smaller than for both CE methods, indicating good matching of the moments of the distributions. Booklet 1 and Booklet 14 were equated according to Kernel chained (EQ –L) and Kernel post-stratification (EQ-L) equating methods. Table 4 displays the results of equating method.

Table 4

Equivalent scores of Booklet 14 corresponding to raw scores of Booklet 1

Booklet 1 Raw Score	PSE EQ	PSE L	CE EQ	CE L
0	-0.16	-0.69	-0.16	-0.58
1	0.69	0.27	0.69	0.36
2	1.55	1.23	1.55	1.31
3	2.43	2.19	2.43	2.25
4	3.33	3.15	3.33	3.20
5	4.25	4.11	4.24	4.14
6	5.18	5.07	5.16	5.09
7	6.11	6.03	6.09	6.03
8	7.04	6.99	7.03	6.98
9	7.98	7.95	7.96	7.92
10	8.93	8.91	8.90	8.87
11	9.87	9.87	9.84	9.81
12	10.82	10.83	10.78	10.76
13	11.77	11.79	11.72	11.70
14	12.72	12.75	12.66	12.65
15	13.67	13.71	13.60	13.60
16	14.62	14.67	14.54	14.54
17	15.58	15.63	15.47	15.49
18	16.54	16.59	16.41	16.43
19	17.50	17.55	17.35	17.38
20	18.46	18.51	18.29	18.32
21	19.42	19.47	19.22	19.27
22	20.38	20.43	20.16	20.21
23	21.35	21.39	21.11	21.16
24	22.32	22.35	22.05	22.10
25	23.28	23.31	23.01	23.05
26	24.25	24.27	23.97	23.99
27	25.22	25.23	24.93	24.94
28	26.19	26.19	25.91	25.88
29	27.16	27.15	26.89	26.83
30	28.13	28.11	27.87	27.77
31	29.11	29.07	28.86	28.72
32	30.09	30.03	29.86	29.66
33	31.08	30.99	30.87	30.61

Table 4 Continue…

Booklet 1 Raw Score	PSE EQ	PSE L	CE EQ	CE L
34	32.08	31.95	31.89	31.55
35	33.09	32.91	32.93	32.50
36	34.14	33.87	33.99	33.44
37	35.23	34.83	35.11	34.39
38	36.40	35.79	36.31	35.33
39	37.67	36.75	37.62	36.28

Table 4 showed that the raw scores from Booklet 1 got values from 0 to 39, but the results of PSE EQ equating showed that equivalent scores of Booklet 14 got points between -0.16 and 37.67, PSE L equating showed the values of -0.16 to 36.75, CE EQ equating yielded values from -0.16 to 37.62 and CE L equating showed values between -0.58 and 36.28. All raw scores of Booklet 1 were greater than Booklet 14 equivalent scores. This implies that Booklet 1 was easier than Booklet 14 throughout the score scale and there was a linear relationship between the raw scores and equivalent scores. Figure 1 and Figure 2 show the differences between the equivalent scores obtained according to the equating methods. Differences KE PSE EQ and KE PSE L and differences between KE CE EQ and KE CE linear are shown in Figure 1.

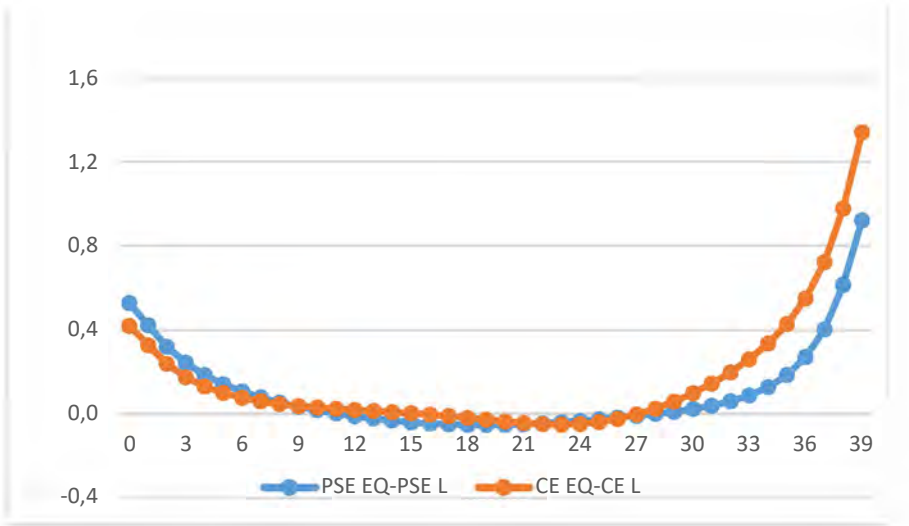

Figure 1. *Differences between KE PSE EQ and KE PSE L and differences between KE CE-EQ and KE CE L*

Figure 1 shows the raw-to-raw equating differences between KE PSE equipercentile and KE PSE linear and differences between KE CE- equipercentile and

KE CE linear, respectively. The results indicated that KE PSE equipercentile produced very similar results to KE PSE linear, except a high range of the score scale. KE CE equipercentile produced very similar results to KE CE linear, except between the scores of 36 and 39. The differences between KE PSE equating methods were smaller than DTM below 38 raw score points and the differences between KE CE equating methods were smaller than DTM below the raw score point of 36. Differences between KE PSE EQ and KE CE EQ and between KE PSE linear and KE CE linear are shown in Figure 2.

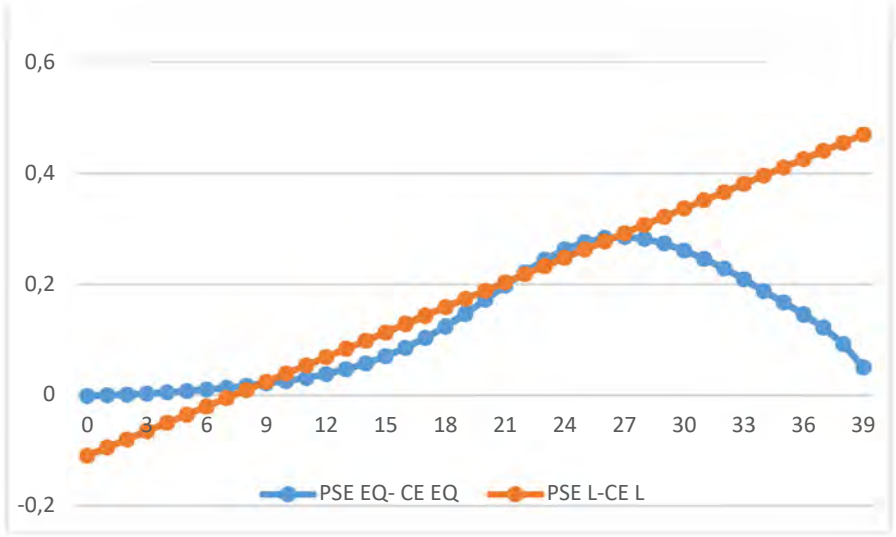

Figure 2. *Differences KE PSE EQ and KE CE EQ and differences between KE PSE L and KE CE L*

Figure 2 shows the raw-to-raw equating differences between KE PSE and KE CE equipercentile methods and differences between KE PSE and KE CE linear methods, respectively. The results indicated that KE PSE equipercentile method produced very similar results to KE CE equipercentile and KE PSE linear produced very similar results to KE CE linear. The differences between all equating methods were smaller than DTM. Figure 3 shows the values of the SEE obtained for each raw point from Kernel equipercentile and Kernel linear equating methods. The mean SEE values were found as .511 for KE PSE equipercentile; .573 for KE PSE linear; .526 for KE CE equipercentile, and .598 for KE CE linear methods.

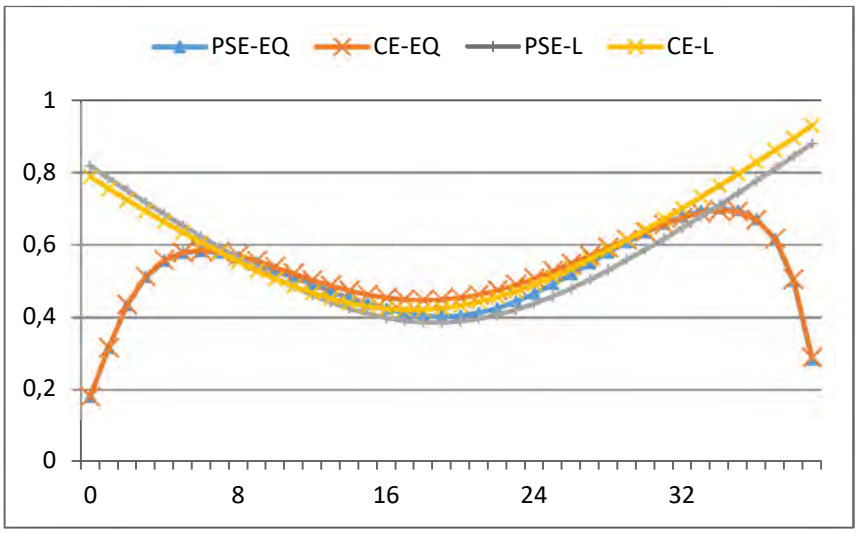

Figure 3. *SEE for each equating methods*

Figure 3 reveals that the standard error values were close to each other in the middle of the raw score scale (range of 8-32 points). On the other hand, at extreme points, Kernel equipercentile equating methods showed lower levels of standard errors while linear equating methods had higher standard errors. The SEE values for both equipercentile equating methods were nearly the same and the SEE values for both linear equating methods were close to each other. When we compared all equating methods, PSE method has a slightly smaller SEE for the middle of the raw score scale. SEED values between KE PSE EQ and KE PSE L, and between KE CE EQ and KE CE L were shown in Figure 4 and Figure 5.

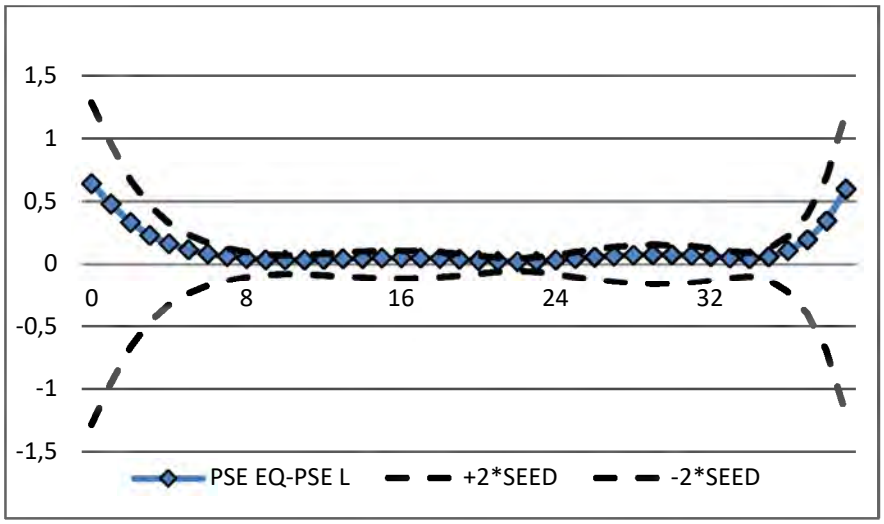

Figure 4. *SEED for equating methods: KE PSE EQ versus KE PSE L*

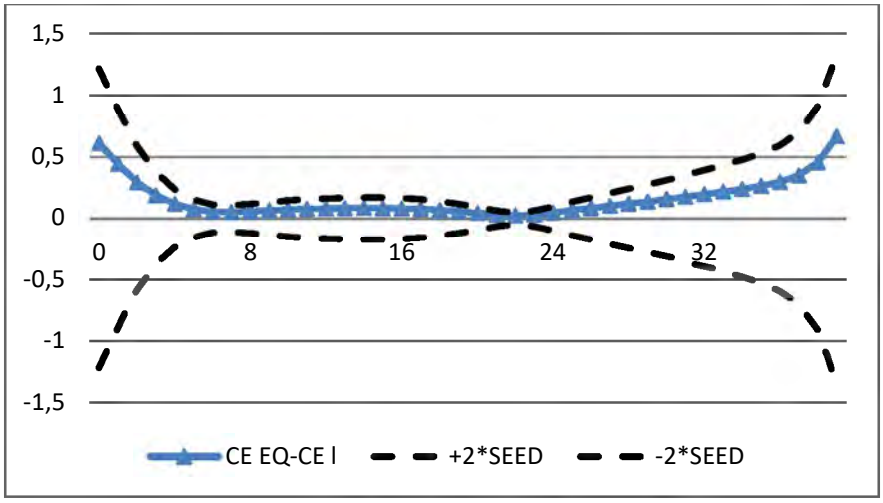

Figure 5. *SEED for equating methods: KE CE EQ versus KE CE L*

Figure 4 and 5 plot the differences and the SEED between PSE EQ and PSE linear, CE EQ and CE linear functions, respectively. Both plots indicated that the line was above the zero line for all score scale. EQ methods equated higher converted scores than linear equating methods, that is to say, EQ methods measured new test X as being harder than linear methods did (i.e., the X form was harder than the Y form). The

differences were smaller than one DTM, expected at the lower end and the higher end of the score scale. Moreover, the difference between the equating functions lied within ±2 SEED across the entire score range, in other words, EQ and linear functions were not significantly different from each other. RMSD coefficient was calculated to evaluate the random involved in the equating methods. The resulting coefficients are given in Table 5.

Table 5

The RMSD Values for Equating Methods

Equating methods	RMSD
PSE-EQ	2.044
PSE-L	2.043
CE-EQ	2.483
CE-L	2.528

It was seen in Table 4 that the equal RMSD coefficients existed in scores equated with KE PSE equipercentile and linear equating methods. The smallest RMSD (2.044 and 2.043) coefficients were obtained from scores equated with PSE method, while the largest RMSD coefficients were obtained through KE CE linear equating method. It can be inferred that whereas the least random error was yielded by KE PSE method, the maximum random error was given by chained linear equating method.

Discussion, Conclusion, and Recommendations

In this study, two Booklets (Booklet 1 and 14) used in TIMMS 2015 science test were equated by using the methods of KE PSE linear, KE PSE equipercentile, KE CE equipercentile and KE CE linear equating methods, and the resulting PRE, SEE, SEED and RMSD values were compared. When reviewing PRE values of Kernel PSE equipercentile and PSE linear, Kernel CE equipercentile and linear; PRE values demonstrate that equipercentile equating methods exhibit lower values than linear equating methods. In other words, it better matches the discrete target distribution Y. Distribution of SEED reveals that the difference between the equating functions lies within ±2 SEED across the entire score range. To put in another way, EQ and linear functions are not significantly different from each other. When the raw-to-raw equating differences between equating methods were examined, the results indicated that KE equipercentile seemed to produce very similar results to KE linear, except the high range of the score scale, and differences between KE PSE and KE CE equating methods were smaller than DTM, except the high range of the score scale. Comparison of the RMSD coefficients based on KE PSE and CE equating methods implies that post-stratification equating method offers the least random error, whereas chained linear equating method yields the maximum random error rates.

When Kernel equating methods were compared against mean SEE, linear equating methods had slightly higher than equipercentile methods. This finding seems incompliant with the findings of Choi (2009) and Liou, Cheng and Johnson (1997). In

his study, Choi (2009) compared the variables of sample size, test length, bandwidth, and presmoothing parameter with Kernel equating and traditional equating methods. He found out that linear Kernel equating methods yield lower standard errors than equipercentile methods. Apart from that, Liou, et al. (1997) found out that the Gaussian Kernel method reduces the standard error with wide bandwidth. While the same study revealed that selection of the parameter h decreases the standard error values, our study found out that the parameter h increased slightly the mean standard error. This difference may be due to the use of simulation data or large sample size in other studies. It was also found out that the KE linear equating methods yielded higher standard error rates at extreme points than the average scores. The results seem to be in conformity with findings of Mao (2006) and Mao, von Davier and Rupp (2006). The latter explained the higher standard errors at extreme values in Kernel equating methods with the use of the Gaussian Kernel method for the continuization of the cumulative score distribution. In the Gaussian Kernel continuization method, the score scale ranges from + ∞ to – ∞ and this leads to arising of increased mean error rates from extreme scores. When the RMSD coefficients obtained based on the KE, PSE, and CE equating methods were compared, the method with the least random error was found to be the post- stratification equating method, while the method with the most random errors was the chained linear equating method.

In this study, Booklets 1 and 14 in the TIMMS 2015 science test were equated in the NEAT design by using Kernel equating methods. A similar study can be carried out by means of equating methods based on the Item Response Theory and the Classical Test Theory, and the results can be compared to the results of this study. A similar study can also be performed for different subtests.

References

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software, 55*(6), 1-25.

Akin-Arikan, Ç. (2017). *Kernel Eşitleme ve Madde Tepki Kuramına Dayalı Eşitleme Yöntemlerinin Karşılaştırılması* [Comparison of kernel equating and item response theory equating methods] (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Akin-Arikan, Ç. & Gelbal, S. (2018). A Comparison of Traditional and Kernel Equating Methods. *International Journal of Assessment Tools in Education, 5*(3), 417-427. doi: 10.21449/ijate.409826

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, United States.

Godfrey, K. E. (2007). *A comparison of Kernel equating and IRT true score equating methods* (Unpublished doctoral dissertation). The University of North Carolina, United States.

Grant, M. C., Zhang, L., & Damiano, M. (2009). An Evaluation of Kernel Equating: Parallel Equating with Classical Methods in the SAT Subject Tests [TM] Program. (ETS RR-09-06). Princeton, NJ: Educational Testing Service.

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

Holland, P., von Davier, A., Sinharay, S., & Han, N. (2006). *Testing the untestable assumptions of the chain and post-stratification equating methods for the NEAT design* (ETS RR-06-17). Princeton, NJ: Educational Testing Service.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-37. doi: 10.1111/j.1745-3992.1988.tb00843.x

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

Lee, Y. H., & von Davier, A. A. (2010). Equating through alternative kernels. In A.A. von Davier (Ed.) *Statistical models for test equating, scaling, and linking* (pp. 159-173). Springer New York.

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the Kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*(4), 349-369. doi: 10.1177/01466216970214005

Liu, J., & Low, A. C. (2007). An Exploration of Kernel Equating Using SAT® Data: Equating to a Similar Population and to a Distant Population. (ETS RR-07-17). Princeton, NJ: Educational Testing Service.

MEB (2016). TIMSS 2015 Uluslararası Matematik ve Fen Eğilimleri Araştırması: TIMMS 2015 Ulusal Matematik ve Fen Bilimleri On Raporu 4. ve 8. Sınıflar. MEB Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü. Ankara. [Çevrim-içi http://timss.meb.gov.tr/wp-content/uploads/TIMSS_2015_Ulusal_Rapor.pdf] Erişim Tarihi: 15 Mart 2018.

Meng, Y. (2012). *Comparison of Kernel Equating and Item Response Theory Equating Methods* (Unpublished doctoral dissertation). University of Massachusetts Amherst, United States.

Mao, X. (2006). *An investigation of the accuracy of the estimates of standard errors for the Kernel equating functions* (Unpublished doctoral dissertation). University of Iowa, Iowa City, United States.

Mao, X., von Davier, A. A., & Rupp, S. (2006). Comparisons of the Kernel equating method with the traditional equating methods on PRAXISTM data (ETS RR-06-30). Princeton, NJ: Educational Testing Service.

Moses, T., & Holland, P. (2007). Kernel and traditional equipercentile equating with degrees of presmoothing (ETS RR-07-15). Princeton, NJ: Educational Testing Service.

Mullis, I. V. S., Cotter, K. E., Centurino, V. A. S., Fishbein, B. G., & Liu, J. (2016). Using Scale Anchoring to Interpret the TIMSS 2015 Achievement Scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 14.1-14.47). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-14.html

Norman Dvorak, R. K. (2009). *A comparison of Kernel equating to the test characteristic curve methods* (Unpublished doctoral dissertation). University of Nebraska, Lincoln, United States.

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

Ricker, K., & von Davier, A. A. (2007). *The impact of anchor test length on equating results in a nonequivalent groups design* (ETS RR-07-44). Princeton, NJ: Educational Testing Service.

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

von Davier, A. 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 pseudotests constructed from real test data* (ETS RR-06-02). Princeton, NJ: Educational Testing Service.

Kernel Eşitleme Yöntemlerinin Denk Olmayan Gruplarda Ortak Madde Test Deseninde Karşılaştırılması

Atıf:

Akin-Arikan, C. (2019). A comparison of kernel equating methods based on neat design. *Eurasian Journal of Educational Research,* 82, 27-44, DOI: 10.14689/ejer.2019.82.2

Özet

Problem Durumu: 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 (Kolen, 1988). Test eşitleme yöntemleri yaklaşık 100 yıldır psikometristlerin dikkatini çekmekte ve yeni yöntemler geliştirilmektedir. Eşitleme yöntemleri eşit yüzdelikli eşitlemeye dayalı yöntemler, doğrusal eşitleme yöntemleri, MTK gözlenen ve gerçek puan eşitleme, van der Linden yerel eşitleme, Levine doğrusal olmayan metot ve yeni bir yaklaşım olan Kernel eşitlemeyi kapsar (von Davier, 2013). Tek grup, eşdeğer grup ve denk olamayan gruplarda ortak madde test deseninde kullanılır (von Davier et al., 2004). Denk olmayan gruplarda ortak madde deseni (Non-Equivalent groups Anchor Test-NEAT), test güvenliği nedeniyle test formunun birden daha fazla uygulandığı durumlarda kullanılır. NEAT deseninde, her iki formda ortak maddeler yer alır ve test formları arasındaki eşitleme ilişkisi de ortak maddeler üzerinden kurulur (Kolen ve Brennan, 2014). Kernel eşitleme doğrusal ve eşit yüzdelikli eşitleme yöntemlerini içerir. NEAT deseninde zincirleme eştileme (doğrusal ve eşit yüzdelikli), son tabakalama (eşit yüzdelikli ve doğrusal), Levine gözlenen puan doğrusal eşitleme yöntemleri bulunmaktadır. Yeni bir yaklaşım olan Kernel eşitleme yöntemlerinin geleneksel eşitle yöntemleri ve Madde Tepki Kuramı eştilem yöntemleri ile karşılaştırıldığı çalışmalar bulunmakdadır. Bu çalışmanın amacı ise, Türkiye'nin de yer aldığı TIMMS fen datasındaki Kernel eşitleme yöntemlerine göre eşitlenmesidir.
Araştırmanın Amacı: Bu araştırmanın amacı, TIMMS fen datasındaki 1. Ve 14. Kitapçıklarının Kernel eşitleme yöntemlerinden zincirleme ve son tabakalama eşitleme yöntemlerine göre eşitlenerek, en iyi eşitleme yönteminin belirlenmesidir.

Araştırmanın Yöntemi: TIMSS 2015 araştırmasının yapıldığı dönemde Türkiye'de toplam 1.108.572 4. sınıf öğrencisi, 1.187.893 de 8. sınıf öğrencisi bulunmaktadır. 6456, 4. sınıf öğrencisi ve 6079, 8. Sınıf öğrencisi TIMMS uygulamasına katılmıştır. Araştırmanın örneklemini ise Türkiye'deki TIMMS uygulamasına katılan 8 sınıf öğrenciler arasından, bu uygulama esnasında 1. ve 14. kitapçıkları alan 865 öğrenci oluşturmaktadır. Veri analizi için TIMMS 2015 uygulanmasına katılan Türkiye'deki 8. sınıf öğrencilerin fen okuryazarlığı maddelerine verdiği cevap örüntülerinden oluşan veri setinden yararlanılmıştır. Bu çalışmada TIMMS uygulamasında yer alan 14 kitapçıktan 1 ve 14 nolu kitapçıklarda yer alan maddeler kullanılmıştır. 4 nolu

kitapçıkta 39, 14 nolu kitapçıkta 38 madde yer almaktadır. Yanlış ve kayıp veriler 0 ve kısmi puanlanan ve doğru cevapların hepsi 1 olarak kodlanarak analiz edilecek veri hazırlanmıştır. Verilerin analizinin birinci aşamasında, Kernel zincirleme ve Kernel son tabakalama eşit yüzdelikli ve doğrusal eşitleme yöntemlerine göre kitapçıklar eşitlenmiştir. Daha sonra eşitleme yöntemleri DTM, PRE, SEE, SEED ve RMSD kriterlerine göre değerlendirilmiştir.

Araştırmanın Bulguları: Kernel zincirleme eşit yüzdelikli, zincirleme doğrusal, son tabakalama doğrusal ve son tabakalama eşit yüzdelikli eşitleme yöntemlerine göre kitapçıklar eştilendiğind eilk olarak PRE değerleri elde edilmiştir. KE zincirleem eşit yüzdelikli ve son tabakalama eşit yüzdelikli eşitleme yöntemlerine datanın daha iyi uyum sağladığı elde edilmiştir. Eşitleme yöntemleri karşılaştırıldığında, eşit yüzdelikli eşitleme yöntemlerinin ve doğrusal eşitleme yöntemlerinin birbiriyle benzer sonuçlar ürettiği ve aralarındaki farkın DTM'den küçük olduğu elde edilmiştir. Eşitleme yöntemlerine göre SEE değerleri karşılaştırıldığında, orta puan ölçeğinde bu değerlerin birbirlerine yakın olduğu görülmektedir. Uç puanlarda ise Kernel eşit yüzdelikli eşitleme yöntemleri düşük, doğrusal eşitleme yöntemleri ise yüksek standart hatalara sahip olduğu elde edilmiştir. Eşitleme yöntemlerine göre SEED değerleri karşılaştırıldığında, eşitleme yöntemleri arasındaki farkın DTM'den küçük olduğu ve ±2 SEED çizgisi arasında kaldığı bulunmuştur. Eşitleme yöntemlerine karışan random hatayı değerlendirebilmek için RMSD katsayısı hesaplanmıştır. En az random hata içeren eşitleme yöntem son tabakalama eşitleme yönteminde iken en fazla random hata içeren yöntemin zincirleme doğrusal eşitleme yönteminde olduğu elde edilmiştir.

Araştırmanın Sonuçları ve Önerileri: Kernel eşitleme yöntemleri ortalama SEE açısından karşılaştırıldığında, doğrusal eşitleme yöntemlerinin eşit yüzdelikli yöntemlere göre daha yüksek ortalama SEE sahip olduğu bulunmuştur. Bu bulgu Choi (2009) ve Liou ve diğerlerinin (1997) bulgularıyla tutarlı olmadığı görülmektedir. Elde edilen bu sonuç diğer çalışmalarda simülasyon data veya geniş örneklem büyüklüğünün kullanılmasından kaynaklı olabilir. Ayrıca KE doğrusal eşitleme yöntemlerinde uç puanlarda orta puanlara göre daha yüksek standart hata verdiği bulunmuştur. Bu bulgu literatürdeki çalışmaları desteklemektedir. RMSD katsayıları karşılaştırıldığında en az random hata içeren yöntem son tabakalama eşitleme yöntemi iken en fazla random hata içeren yöntemin zincirleme doğrusal eşitleme olduğu görülmüştür. Elde edilen bu sonuçlardan hareketle, gelecek çalışmalarda farklı kriterler kullanılarak farklı eşitleme yöntemleri kullanılabilir ve bu çalışmanın sonuçlarıyla karşılaştırılabilir.

Anahtar Sözcükler: eşitleme, eşit yüzdelikli, doğrusal, SEED, SEE