ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(4); 348-357

Comparison of Kernel Equating and Kernel Local Equating in Item Response Theory Observed Score Equating

Merve YILDIRIM SEHERYELİ * Hasibe YAHSİ SARI ** Hülya KELECİOĞLU ***

Abstract
The present study aims to compare the Kernel equating and Kernel local equating methods in observed score equating. Functions and error estimates regarding the difference between raw and equated scores and the scores equated by Stocking-Lord and Haebara true-score equating methods in Kernel local equating and Kernel equating were examined in Item Response Theory Observed Score Equating. Therefore, 5, 10, and 15 external anchor items were used, and scores were obtained from two forms based on the 2PL model. R (version 3.5.3.) programming software was used for IRT assumptions, item parameters, calibration, and equating analyses. The results revealed that Stocking-Lord and Haebara true-score equating methods yielded similar results. Moreover, if the equating method is the same, estimation errors decreased when the number of anchor items increased. The mean scores obtained by Kernel equation 5 and 15 anchor items were lower than Kernel local equating, while means of Kernel equating of 10 anchor items were higher. As the number of items increased, estimation errors decreased, and Kernel local equating revealed the lowest errors in the medium score scale. Kernel equating can be used based on the related ability level if the individual's ability distribution is known.

Key Words: Test equating, Kernel equating, Kernel local equating, item response theory, local equating.

INTRODUCTION

Measurement tools are used for many purposes, such as measuring cognitive, affective, or psychomotor characteristics of individuals, getting to know individuals, placing them in any institution or school. The validity and reliability of the measurements of these measurement tools are a need for better measurement. To increase test reliability and therefore test validity, different test forms measuring the same feature are used especially in exams with wide participation and high risk, such as selection and placement exams, whose results greatly affect the future of individuals. These different test forms must have the same degree of difficulty for individuals to be evaluated fairly (Haladyna & Downing, 2004). However, this is not always possible in practice. In this case, the scores obtained from the test forms that do not have the same difficulty level should be brought to the same scale and the scores should be equated so that the forms can be used interchangeably. These statistical processes are possible with the help of test equating (Kolen & Brennan, 2004). Equating brings the scores on the test forms to the same scale, allowing them to be used interchangeably and the scores to be compared (Hambleton & Swaminathan, 1985). Thus, bias towards the measurement tools used in different test forms can be eliminated.

Equating has certain steps to be followed. The first step is to determine the data collection design. There are five data collection designs, which include equivalent groups design, single group design, counterbalanced design, non-equivalent groups with anchor test design, and non-equivalent groups with covariates design (González & Wiberg, 2017). The second step is to determine the equating method. These methods differ based on classical test theory (CTT) and item response theory (IRT).

* Ress. Assist., Hasan Kalyoncu University, Faculty of Education, Gaziantep-Turkey, yldrm.mrv.7806@gmail.com, ORCID ID: 0000-0002-1106-5358
** PhD. Student, Hacettepe University, Faculty of Education, Ankara-Turkey, hsbyahsi@gmail.com, ORCID ID: 0000-0002-0451-6034
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

To cite this article:
Yıldırım Seheryeli, M., Yahsi Sarı, H., & Kelecioğlu, H. (2021). Comparison of Kernel equating and Kernel local equating in item response theory observed score equating. *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 348-357. doi: 10.21031/epod.900843

Received: 22.03.2021
Accepted: 30.09.2021

Diao (2018) categorizes CTT test equating methods into four: identity equating, mean equating, linear equating and equipercentile equating. CTT-based equating methods differ among themselves based on true-score equating and observed score equating. Before starting the equating process in IRT, separate and concurrent calibrations are conducted to scale parameters. Item parameters of the two forms are estimated at the same time in concurrent calibration. In separate calibration, on the other hand, forms are scaled separately and calibrated using common items. Calibration methods include the moment method (mean-mean, mean-sigma) and characteristic curve transformation (Haebara and Stocking-Lord) (Kolen & Brennan, 2004). At the last stage of equating, standard errors are calculated, and properties of equating are checked. These properties are symmetry, same specifications, equity, observed score, and group invariance (González & Wiberg, 2017; Kolen & Brennan, 2004).

There are equating studies based on non-equivalent groups with covariates in IRT. Also, the effect of skewness of ability distributions, multidimensionality, violation of group invariance in different levels on equating errors have been examined in some studies (Gök & Kelecioğlu, 2014; Öztürk-Gübeş, 2019; Öztürk-Gübeş & Kelecioğlu, 2015; Tanberkan-Suna, 2018; Uysal, 2014). The general results of the studies showed that IRT true score equating method performed best in providing test fairness, while IRT observed score equating method performed best in decreasing measurement errors.

Equapercentile and linear equating methods are used in observed-score equating (von Davier, 2008). These methods include equapercentile equating methods, linear equating methods, IRT observed score equating, local equating, non-linear equating, and Kernel equating (von Davier, 2013).

Kernel equating is defined as an equapercentile equating method that transforms discrete score distribution into a continuous distribution (von Davier, Holland & Thayer, 2004). Kernel equating was first defined by Holland and Thayer (1981). The Kernel equating methods can be applied as post-stratification equapercentile, post-stratification linear, chained equapercentile and chained linear. The equating methods based on CTT uses linear estimates for the continuation of the score distributions, while Gauss Kernel method is used in Kernel equating (von Davier et al., 2004). There are five steps in observed score Kernel equating: pre-smoothing, estimating score distributions for the target population, computing the equating function, continuizing the discrete score distributions, and computing the standard error of equating (von Davier, 2013). Pre-smoothing helps the data become consistent. Kernel equating smoothes data transformation and provides a small standard error. Also, it is less affected by the change in the sample compared to other methods. Kernel equating is used with equivalent groups design, single group design, counterbalanced design, and non-equivalent groups with anchor test design (von Davier et al., 2004). There are various studies that used Kernel equating (Akın Arıkan, 2017; Andersson & Wiberg, 2014; Choi, 2009; Liou, Cheng, & Johnson, 1997; Norman Dvorak, 2009; Wiberg, van der Linden, & von Davier, 2014). Akın Arıkan (2017) found that the extreme scores yielded greater standard errors as the group ability distributions varied in Kernel equating. In IRT true score equating; on the other hand, middle and high scores had the greatest error. Kernel equating methods had lower standard errors in the medium score scale and had higher standard errors in extreme scores where score frequency was lower compared to the IRT true score equating in all conditions. Moreover, lower errors were obtained through the IRT true score equating method than the Kernel equating methods regarding the extreme scores.

Local equating also became popular along with Kernel equating in observed score equating (von Davier et al., 2004). It was first introduced by Lord (1980) in his definition of equating (as cited in van der Linden, 2000). All traditional equating methods use the same equating transformations for all populations of test participants. van der Linden (2000) revealed that equating should be done separately for each ability level. Local equating offers a common ground for different transformations for each ability level. In local equating, if both test forms are appropriate for the item response theory (IRT) and can be used with any equating design, the IRT observed score could be defined as the local Kernel equating (Wiberg et al., 2014).

Wiberg et al. (2014) proposed three different observed score Kernel local equating methods by combining local equating and Kernel equating. The methods for local Kernel equating on-equivalent groups with anchor test design are: IRT observed score equating, anchor test score Kernel equating and local Kernel equating with ability estimated by anchor test. These new methods are compared to previous methods in terms of measures such as bias and relative error percentage. The item response theory observed score local Kernel equating method, which is used for all common equating methods, yielded bias, relative error, and Kernel standard equalization error, even when the measurement precision of the test was reduced. Kernel local equating methods generally showed low bias in the non-equivalent groups with anchor test design. In addition, the anchor was highly stable against variations in the accuracy and length of the test.

Many studies used Kernel equating (Akın Arıkan, 2017; Andersson & Wiberg, 2014; Choi, 2009; Liou et al., 1997; Norman Dvorak, 2009; Wang, Zhang ve You, 2020; Wiberg et al., 2014). These studies revealed that Kernel equating and similar traditional equating methods can be compared with equivalent groups design in equivalent and non-equivalent groups with anchor test design when estimating standard errors in equapercentile equating (Choi, 2009); and that R program was used for IRT observed score Kernel local equating (Andersson & Wiberg, 2014). Wiberg et al. (2014) used three different observed score Kernel local equating methods by combining local equating and Kernel equating in their study. Studies have shown that with Kernel local equating, equating functions can be obtained at each ability level, and thus estimation errors can be minimized (González & Wiberg, 2017; Wiberg et al., 2014). The present study compares the Kernel local equating with Kernel equating to examine the bias in the equating processes and the contribution of the methods to the test validity.

In kernel equating methods, in cases where the ability distributions between groups are different, extreme scores yield high standard errors (Akın Arıkan, 2017). Wiberg et al. (2014) concluded that in the common item nonequivalent groups, It is predicted that Kernel local equating methods will yield a lower standard error in cases where the ability distributions of individuals are known and the test fairness is ensured to make more accurate equating. In this study, the results of Kernel local equating are compared with Kernel equating under various conditions. Since there are few studies on this subject (Akın Arıkan, 2017; Wiberg et al., 2014) and there are no studies that examine Kernel equating and Kernel local equating together, the study aims to compare the results of Kernel equating and Kernel local equating.

Purpose of the Study

In the present study, θ values with values decreasing one by one between -6 and 0 (low), θ = 0 (middle) and θ values with values increasing one by one between 0 and +6 (high) ability levels of the scores obtained from two different forms based on 2PL model and different anchor item numbers in IRT observed score Kernel equating and IRT observed score Kernel local equating were included. Stocking-Lord and Haebara were used for data transformation and the equating results were compared. To this end, different anchor item numbers (10, 20, and 30) were used and after data were transformed with Stocking-Lord and Haebara methods, equating functions and errors were examined with observed score Kernel equating and observed score Kernel local equating.

METHOD

Research Design

In this study, data were artificially produced in order to examine the change of errors in cases where different anchor items were used in the equating methods and these items were not included in the total score. Therefore, this study is a simulation study.

Data Production

The items in the X and Y forms and the anchor materials were produced under the conditions in Table 1 according to the 2PL Model. The items in both data sets were produced using the “kequate” package

(Andersson, Bränberg & Wiberg, 2020) in the R Studio interface of the R (version 3.5.3) programming software with 20 items different and the number of anchor items 10, 20 and 30.

The forms consisting of X form and anchor items are named as P and Y forms and forms consisting of anchor items as Q forms. A parameters are uniformly distributed with values ranging from 0.50 - 2.0, the *b* parameters are *N*(0; 1) and the ability parameters are *N*(0.50; 1) for the P form, and *N*(0; 1) for the Q form.

According to Baker (trans. 2016), *a* parameters range between -2.80 and +2.80 in practice, while *b* parameters range between -3.00 and +3.00. In addition, he specified the cutoff point as 0.35 for the low level of the parameter *a*, 0.64 for the medium level, and 1.35 for the high level. In order to have medium and high level *a* parameters, they were taken between 0.50 and 2.00.

Wang, Lee, Brennan, and Kolen (2008) stated that the similarity distributions are important in terms of equating results and they considered the difference over 0.25 as *very wide*. Therefore, it was expected that the difference was taken as 0.50 to reveal the difference between Kernel equating and Kernel local equating errors more clearly.

Kolen and Brennan (2004) stated that the ratio of the number of anchor items to the total number of items in the test should be at least 20%. Therefore, while 30 items were different in all data sets (X and Y forms), the number of anchor items was determined as 5, 10 and 15. The number of iterations was determined as 100. A total of 600 (2 x 3 x 100) data sets were obtained. The average and ranges of the difficulty and discrimination parameters of the sample distributions obtained from the P and Q forms according to the number of anchor items are given in Table 1.

Table 1. The Mean and Ranges of the Difficulty and Discrimination Estimates Obtained From the P and Q Forms According to the Anchor Item Numbers

		P forms				Q forms			
Total Number of Items	Number of Anchor Items	Mean b	Range b	Mean a	Range a	Mean b	Range b	Mean a	Range a
35	5	-0.505	4.179	1.261	1.622	-0.013	4.454	1.253	1.614
40	10	-0.534	4.502	1.248	1.653	0.002	4.448	1.253	1.677
45	15	-0.537	4.625	1.255	1.656	-0.031	4.547	1.244	1.624

Table 1 shows that for each anchor item number, the means of *b* parameters related to the P forms are approximately 0.50 lower than the Q forms. The means of *a* parameter are approximately the same.

Data Analysis

In IRT Kernel equating (post-stratification equating) parameters were calibrated based on Stocking-Lord and Haebara methods. Then, an external anchor design was used in which the anchor items were not included in the total score. The distribution of equated scores, means of equating errors, and functions related to the difference between the raw score and the equated scores were compared regarding both calibration methods. Similar studies in kernel local equalization for cases where ability levels are low (decreasing one by one between -6 and 0) (L: Low), zero (M: Medium), and high (increasing one by one between 0 and +6) (H: High) was also repeated. “psych” (Revelle, 2021), “mirt” (Chalmers et al., 2021), “kequate” (Andersson et al., 2020), “ltm” (Rizopoulos, 2018) packages were used in the R Studio interface of the R (version 3.5.3) programming software for all analyzes.

RESULTS

This chapter presents the results of IRT Kernel observed score equating and Kernel observed score local equating. In this regard, score distributions, equating functions and distribution of equating errors were examined.

Equated Score Distributions and Functions of Score of Difference

In cases when the numbers of anchor items were 5, 10 and 15, Stocking-Lord and Haebara methods were examined. Ability levels for Kernel local equating are low (decreasing one by one from 0), medium (0), and high (increasing one by one from 0). Table 2 shows the equated score distributions obtained with the IRT observed score Kernel equating and Kernel local equating. These values were the result of calculating the means of the values obtained with 100 iterations.

Table 2. Equated Score Distributions Obtained With IRT Observed Score Kernel Equating and Kernel Local Equating

		Kernel Equating				Kernel Local Equating				
Calibration	Number of Anchor	Min.	Max.	Mean	S.D.	θ level	Min.	Max.	Mean	S.D.
Stocking-Lord	5	0.140	29.905	15.054	8.989	L	0.141	29.907	15.071	9.000
						M	0.167	29.906	15.039	9.022
						H	0.173	29.905	15.023	9.008
	10	-0.026	29.917	14.827	9.067	L	-0.022	29.782	14.817	9.039
						M	-0.048	29.787	14.792	9.060
						H	-0.045	29.916	14.802	9.086
	15	0.112	29.876	15.008	9.005	L	0.112	29.944	15.044	9.039
						M	0.069	29.948	15.009	9.070
						H	0.073	29.877	14.973	9.033
Haebara	5	0.141	29.912	15.067	8.993	L	0.142	29.919	15.084	9.005
						M	0.181	29.919	15.058	9.021
						H	0.186	29.912	15.042	9.007
	10	-0.032	29.921	14.824	9.073	L	-0.026	29.782	14.812	9.043
						M	-0.051	29.787	14.790	9.062
						H	-0.049	29.921	14.802	9.090
	15	0.117	29.871	15.002	8.997	L	0.118	29.938	15.040	9.033
						M	0.064	29.942	15.001	9.069
						H	0.068	29.871	14.963	9.030

Note. L: Low. M: Medium. H: High

Kernel equating results in Table 2 shows that when 10 anchor items were used in both calibrations, the equated scores were estimated with a low mean score. Again, in both estimation methods, the condition in which scores are estimated with a higher mean is the case where the number of anchor items is 5. The number of anchor items shows that the mean and standard deviations of the scores equated according to the methods do not differ much. Figure 1 shows the function graph regarding the differences between the equated scores and the raw scores taken from the test.

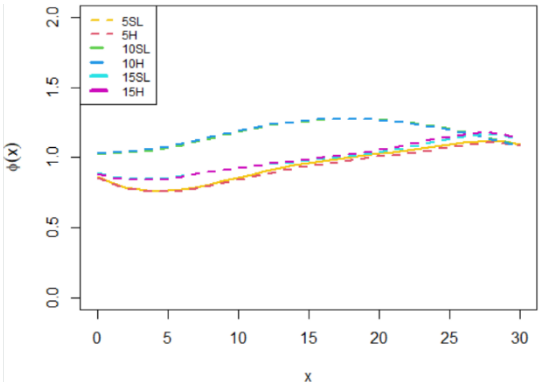

Figure 1. Function Graphs Regarding the Differences of Equated Scores and Raw Scores Obtained With IRT Kernel Equating

Figure 1 shows that the distribution of the difference scores is almost the same when the number of anchors is the same and the calibration method is different. When using 5 anchors, the differences decreased up to approximately 5 raw points, while the differences gradually increased after 5 raw points. When 10 anchor items were used, the difference scores increased as the raw scores increased and started to decrease after about 17 raw scores. When 15 anchor items were used, the difference scores increased as the raw scores increased, but the rate of change was relatively low. On the other hand, when 10 anchor items are used up to 26 raw points, the differentiation from the raw scores in both methods is higher compared to the other anchor items. When the raw score is greater than 26, it is seen that the differentiation from the raw score is more when 15 anchor items are used.

Table 2 demonstrates the Kernel local equating results and reveals that the mean scores of the equated scores at low, medium, and high ability levels are the highest in 5 anchor items and the lowest in 10 anchor items when Stocking-Lord (S-L) and Haebara (H) methods are used. In the case where 5 anchor items were used, the highest mean score was obtained in the low ability level with Haebara method, while the lowest mean score at the high ability level was obtained with the S-L method. When 10 anchor items were used, the highest mean score was obtained at the low ability level with S-L, and the lowest mean score at the medium ability level was obtained with the Haebara method. In the case where 15 anchor items were used, the highest mean score was obtained at the low ability level with S-L, and the lowest mean score at the high ability level was obtained with the Haebara method. In all conditions, the lowest mean score was obtained with the middle ability level when 10 anchors and the Haebara method was used, and the highest mean score was obtained with the low ability level when 5 anchors and the Haebara method were used.

When both methods are compared, in the case that 5 anchor items were used, the mean score obtained with Kernel equating was lower than the mean score obtained with Kernel local equating based on low ability level. The closest mean score was obtained when the Haebara method is used with the middle ability level. In the case where 10 anchor items are used, the mean scores obtained with Kernel equating are higher under all equating conditions. The closest mean score was obtained when the S-L method is used with the low ability level. In the case where 15 anchor items are used, the mean score obtained with Kernel equating is lower than the mean score obtained with Kernel local equalization with the low ability level. The closest mean score is the case in which the S-L method is used in the equating made according to the middle ability level. Figure 2 shows the function graph regarding the differences between the equated scores and the raw scores obtained from the test.

Figure 2 reveals that the distribution of the difference scores is almost the same when the number of anchors is the same and the calibration method is different. In the case of using 5 anchor items, the difference scores are higher in the equalizations made according to medium and high ability levels up to 14 raw points, while the difference scores are higher in the equations made according to the high ability level and Kernel equating over 17 raw points. In cases where 10 and 15 anchor items are used, up to 12 raw points, the difference scores are higher in the equalizations made according to medium and high ability levels, while the difference scores are higher in the equalizations made at medium and low ability levels over 23 raw points. In the equating made according to the middle ability level, the range of difference scores in each anchor item condition is the smallest.

Error Distributions

Equating errors were calculated for all conditions. These values were the result of calculating the means of the values obtained with 100 iterations. Table 3 shows the distribution of equating errors obtained with the observed score Kernel equating and Kernel local equating.

Table 3 shows that the error means of equated scores under all conditions are estimated higher in the Stocking-Lord method than in the Haebara method. The difference between these error means was found to be approximately .004 when 5 anchor items were used, .002 when 10 anchor items were used, and .003 when 15 anchor items were used. The distribution of the errors shows that as the number of anchor items increases, the errors are closer to each other and become more homogeneous. The smallest of these errors occurred in the calibration made according to both methods when 15 anchor items are used; the greatest was obtained in the calibration performed according to the Stocking-Lord method when 5 anchors were used.

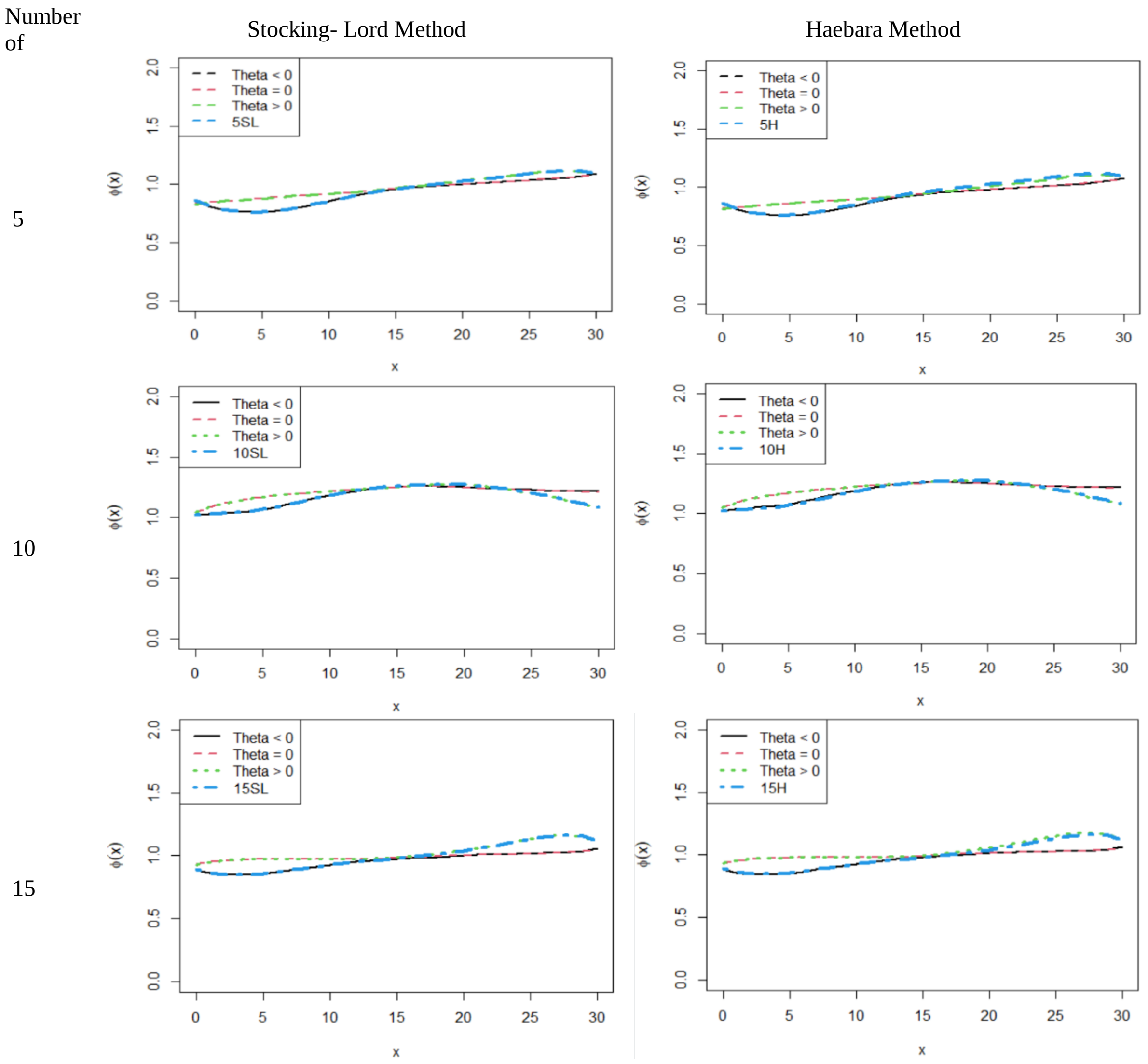

Figure 2. Equated Scores Obtained With Kernel Equating and Kernel Local Equating Based on IRT in Conditions Where the Number of Anchor Items are 5, 10 and 15, Respectively, and Function Graphs Regarding The Differences Of Raw Scores

The results of the Kernel local equating show that the errors of the equated scores were estimated .003 times higher with the Stocking-Lord method compared to the Haebara method. As the number of anchor items increased, equating errors decreased in both methods and both ability levels. In cases when 5, 10, and 15 anchor items were used, the smallest error was obtained with equalizations of the middle ability level. Similarly, in the equatings made according to low ability level, estimates were made with relatively high errors. In addition, in the equalizations made according to the middle ability level, errors are relatively more homogenous. The smallest of these errors was when the 15 anchor items were used according to the middle ability level with the Haebara method. The greatest error, on the other hand, was when 5 anchor items were used with the low ability level based on the Stocking-Lord method.

It was revealed that Kernel equating errors were greater than those of Kernel local equating when comparing Kernel equating and Kernel local equating in all conditions in all ability levels.

Table 3. Error Distributions Obtained From IRT Observed Score Kernel Equating and Kernel Local Equating

		Kernel Equating				Kernel Local Equating				
Calibration	Number of Anchor	Min.	Max.	Mean	S.D.	θ level	Min.	Max.	Mean	S.D.
Stocking-Lord	5	0.188	0.496	0.375	0.080	L	0.188	0.490	0.354	0.082
						M	0.189	0.367	0.312	0.047
						H	0.181	0.418	0.333	0.058
	10	0.147	0.365	0.278	0.056	L	0.144	0.362	0.263	0.060
						M	0.141	0.263	0.228	0.032
						H	0.139	0.302	0.243	0.040
	15	0.134	0.312	0.243	0.045	L	0.129	0.310	0.229	0.049
						M	0.127	0.232	0.201	0.027
						H	0.127	0.265	0.215	0.033
Haebara	5	0.186	0.491	0.371	0.079	L	0.187	0.482	0.350	0.080
						M	0.188	0.363	0.308	0.047
						H	0.180	0.418	0.330	0.059
	10	0.146	0.362	0.276	0.056	L	0.143	0.358	0.260	0.059
						M	0.140	0.261	0.225	0.032
						H	0.138	0.303	0.241	0.041
	15	0.134	0.309	0.240	0.045	L	0.129	0.305	0.226	0.048
						M	0.126	0.228	0.198	0.027
						H	0.127	0.264	0.213	0.034

Note. L: Low, M: Medium, H: High

DISCUSSION and CONCLUSION

In this study, the P and Q forms based on the 2PL model with different anchor item numbers (5, 10, 15) were evaluated for different ability levels [θ < 0 (low), 0 (moderate) and θ >0 (high)]. Equating results of Stocking-Lord and Haebara methods were examined.

The present study used simulated data in which the anchor items were not included in the individual scores, unlike the studies of Öztürk-Gübeş and Kelecioğlu (2015), Pektaş and Kılınç (2016), Tanberkan-Suna (2018), in which the real data were used. Akın Arıkan (2017), used simulated data as well; however, she only compared the Haebara method in IRT true score equating and Kernel equating methods. Öztürk-Gübeş (2019), on the other hand, investigated the effect of multidimensionality on test equating and not included the change in the item numbers. Moreover, Wang et al. (2020) compared equapercentile equating, Kernel equating, and IRT Kernel equating methods.

Errors and function graphs were examined related to the difference between raw and equated scores in IRT observed score Kernel equating non-equivalent anchor test design when anchor items and calibration methods differ. The results revealed that there are differences and similarities between the equated scores, the distribution of the difference scores and errors in non-equivalent groups with anchor test design with Stocking-Lord and Haebara methods. Equated scores were estimated with a higher mean score when 5 anchor items were used in both calibration methods. In all the conditions, equated scores are lower than each score that can be obtained from the test. In cases when the anchor item numbers were the same, errors of the equated scores based on Haebara method were estimated lower. As the number of anchor items increased, the errors of the estimates in both methods were closer to one another. Wang et al. (2020) also obtained similar results where the number of items was 30 and 45 in the simulation. This finding is not supported by the findings of Uysal (2014), in which he found that error estimates with the Stocking- Lord method were lower than the Haebara method.

In addition, the present study investigated the functions and errors regarding the difference scores and equated scores when the item numbers and calibration methods differed. Both Stocking-Lord and

Haebara methods yielded similar results in cases that the same number of anchor items were used and the equatings of the scores and errors were conducted according to the low, middle, and high ability levels. The mean scores of low, middle, and high ability levels were the greatest with the 5 anchor items and the smallest with the 10 anchor items with both methods. In all conditions, the lowest mean score was obtained with the Haebara method and 10 anchor items according to middle ability level. The highest mean score, on the other hand, was obtained with the Haebara method and 5 anchor items according to low ability level.

When both methods are compared, mean scores obtained with Kernel local equating with 5 and 15 anchor items according to low ability level were estimated higher than Kernel equating. When 10 anchor items were used, the results of Kernel equating were the highest in all the conditions. Also, the graphs about the relationship between raw and equated scores showed that the range of difference scores were the narrowest when Kernel local equating were used regardless of the calibration method. The reason for this could be the fact that errors were estimated lower with the help of different equating functions based on the ability level and raw scores.

The lowest errors were estimated when both methods were used with 5, 10, and 15 anchor items. Moreover, errors were homogenous in the equatings based on middle ability level. The reason behind this result could be that the simulation data were simulated with normal distribution in the middle ability level ($b = 0$). This finding was supported by Wiberg et al. (2014), which suggests three different observed score Kernel local equating methods by combining local equating and Kernel equating and found that Kernel local equalization methods are quite stable against the changes in the accuracy and length of the anchor test in the non-equivalent groups anchor test design. The Kernel local equating errors were lower than Kernel equating errors when the two methods were compared. This finding is not supported by the results of the study of Wiberg et al. (2014) in which they found that the Kernel local equating method yielded higher standard errors than Kernel equating.

As a result, it was found that IRT observed score Kernel equating and Kernel local equating Stocking-Lord and Haebara methods can both be used and to keep the errors low, the number of anchor items should be kept higher. Also, Kernel local equating should be used with the ability level most appropriate to the ability distribution of the individuals. In future studies, different Kernel equating methods, different calibration types, and different data collection designs can be used to compare the observed score with the true score equating in cases where the anchor item is internal and external. Also, Kernel equating, and Kernel local equating methods can be examined using the equivalent groups design. In addition, equating errors can be examined by dividing ability levels in IRT Kernel local equating. The present study made use of the simulation data; a similar study can be conducted with real data set.

REFERENCES

Akın Arıkan, Ç. (2017). *Kernel eşitleme ve madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması* (Yayımlanmış Doktora Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Andersson, B., & Wiberg, M. (2014). *IRT observed-score kernel equating with the R package kequate.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.372.8712&rep=rep1&type=pdf

Andersson, B., Bränberg, K., & Wiberg, M. (2020). *Package 'kequate'*. Retrieved from https://mran.microsoft.com/snapshot/2020-03-08/web/packages/kequate/kequate.pdf

Baker, F. B. (2016). *Madde tepki kuramının temelleri* [The basics of item response theory]. (N. Güler, Ed., & M. İlhan, Çev.). Ankara: Pegem Akademi. (1985)

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim K. H., Falk C. F., …, and Oguzhan, O. (2021). *Package 'mirt'*. Retrieved from https://cran.r-project.org/web/packages/mirt/mirt.pdf

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.

Diao, H. (2018). *Investigation repeater effects on small-sample equating: Include or exclude?* (Doctoral thesis). University of Massachusetts-Amherst.

Gök, B., & Kelecioğlu, H. (2014). Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 10*(1), 120-136. https://dergipark.org.tr/tr/download/article-file/161036 adresinden erişilmiştir.

González, J., & Wiberg, M. (2017). *Applying test equating methods: Using R*. Switzerland: Springer International Publishing. Retrieved from http://www.mat.uc.cl/~jorge.gonzalez/index_archivos/EquatingRbook.htm

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in highstakes testing. *Educational Measurement: Issues and Practice.*, *23*(1), 17-27. doi: 10.1111/J.1745-3992.2004.TB00149.X

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Baston: Kuluwer-Nijhoff Publisihing.

Holland, P. W., & Thayer, D. T. (1981). *Section pre-equating the graduate record examinations* (ETS Research Report Series). *1981*(2), i-62.

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

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

Norman Dvorak, R. L. (2009). *A comparison of kernel equating to the test characteristic curve method* (Unpublished doctoral dissertation). University of Nebraska-Linkoln.

Öztürk-Gübeş, N. (2019). Test eşitlemede çok boyutluluğun eş zamanlı ve ayrı kalibrasyona etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 34*(4), 1061-1074. doi: 10.16986/HUJE.2019049186

Öztürk-Gübeş, N., & Kelecioğlu, H. (2015). Farklı test eşitleme yöntemlerinin eşitlik özelliği ölçütüne göre karşılaştırılması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 48*(1), 299-214. doi: 10.1501/Egifak_0000001358

Pektaş, S., & Kılınç, M. (2016). PISA 2012 matematik testlerinden iki kitapçığın gözlenen puan eşitleme yöntemleri ile eşitlenmesi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 1*(40), 432-444. https://dergipark.org.tr/tr/download/article-file/264191 adresinden erişilmiştir.

Revelle, W. (2021). *Package 'psych'*. Retrieved from https://cran.rstudio.org/web/packages/psych/psych.pdf

Rizopoulos, D. (2018). *Package 'ltm'*. Retrieved from https://cran.r-project.org/web/packages/ltm/ltm.pdf

Tanberkan-Suna, H. (2018). *Grup değişmezliği özelliğinin farklı eşitleme yöntemlerinde eşitleme fonksiyonları üzerindeki etkisi* (Yayımlanmış Doktora Tezi). Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Uysal, İ. (2014). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması* (Yayımlanmamış Yüksek Lisans Tezi). Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Enstitüsü, Bolu.

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika*, *65*(4), 437-456. Retrieved from https://link.springer.com/content/pdf/10.1007/BF02296337.pdf

von Davier, A. A. (2008). New results on the linear equating methods for the non-equivalent groups design. *Journal of Educational and Behavioral Statistics*, *33*(2), 186-203. doi: 10.3102/1076998607302633

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, *78*(4), 605-623. doi: 10.1007/s11336-013-9319-3

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

Wang, S., Zhang, M., & You, S. (2020). A Comparison of IRT Observed Score Kernel Equating and Several Equating Methods. *Frontiers in psychology, 11*, 308. doi: 10.3389/fpsyg.2020.00308

Wang, T., Lee, W. C., Brennan, R. J., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. *Applied Psychological Measurement*, *32*(8), 632-651. doi: 10.1177/0146621608314943

Wiberg, M., van der Linden, W. J., & von Davier, A. A. (2014). Local observed-score kernel equating. *Journal of Educational Measurement*, *51*(1), 57-74. doi: 10.1111/jedm.12034