Bartın University Journal of Faculty of Education
BUEFAD Volume 14, Issue 1

196-210
dergipark.org.tr/buefad
DOI: 10.14686/buefad.1392156

Comparison of Different Bandwidth Determination Methods in Kernel Equating

Vildan ÖZDEMİR [a*]

a Dr., Aksaray University, https://orcid.org/0000-0002-9051-8860 *vildanbagci@gmail.com

Research Article
Received: 17.11.2023
Revised: 22.3.2024
Accepted: 04.4.2024

Abstract

The study aims to compare the Gaussian Kernel, Logistic Kernel, and Uniform Kernel methods for determining the bandwidth parameter in the Kernel equating on TIMSS data. A bandwidth parameter needs to be determined when Kernel equating is used to equate two test forms. The bandwidth parameters determine the smoothness of the continuous score distributions, so their effect on equating results is critical. Gaussian Kernel, Logistic Kernel, and Uniform Kernel methods were used for bandwidth selection, and the results were compared according to the Percentage Relative Error (PRE), the Standard Error, and the Standard Error of Equating Difference (SEED). The findings of the study show that the three different approaches to minimizing the penalty function have similar results. Although the standard errors of the equated scores obtained with the uniform Kernel method were slightly smaller, the results were almost the same as the other two approaches. When the three equating methods were compared according to the percent relative error, the distribution obtained from Gaussian Kernel equating was more consistent with the population distribution.

Keywords: Kernel equating, bandwidth selection, continuousization

Kernel Eşitlemede Farklı Bant Genişliği Seçimi Yöntemlerinin Karşılaştırılması

Öz

Bu çalışma, Kernel eşitleme yönteminde bant genişliği parametresinin belirlenmesi için sunulan yöntemleri gerçek bir veri seti üzerinde karşılaştırmayı amaçlamaktadır. Kernel eşitlemenin süreklileştirme adımında eşitleme yapabilmek için bir bant genişliği parametresinin belirlenmesi gerekir. Bant genişliği parametresi, sürekli puan dağılımlarının düzgünlüğünü belirler, bu nedenle eşitleme sonuçları üzerindeki etkileri kaçınılmazdır. Bant genişliği seçiminde Gauss Kernel, Lojistik Kernel ve Tek Biçimli Kernel yöntemleri kullanılmıştır ve sonuçlar bağıl hata yüzdesi, standart hata ve eşitleme farkına ait standart hataya göre karşılaştırılmıştır. Çalışmanın bulguları, penalty/ceza fonksiyonunun minimize edilmesine yönelik üç farklı yaklaşımın benzer sonuçlar verdiğini göstermektedir. Tek Biçimli Kernel yöntemiyle elde edilen eşitleme puanlarının standart hataları biraz daha küçük olsa da sonuçlar diğer iki yaklaşımla neredeyse aynıdır. Üç eşitleme yöntemi bağıl hata yüzdelerine göre karşılaştırıldığında ise Gauss Kernel eşitlemesinden elde edilen dağılımın evren dağılımıyla daha tutarlı olduğu görülmektedir.

Anahtar Sözcükler: Kernel eşitleme, bant genişliği parametresi seçimi, süreklileştirme

To cite this article in APA Style:

Özdemir, V. (2025). Comparision of different bandwidth determination methods in Kernel equating. *Bartın University Journal of Faculty of Education, 14*(1), 196-210. https://doi.org/10.14686/buefad.1392156

© 2024 Bartın University Journal of Faculty of Education. This is an open-access article under the Creative Commons Attribution NonCommercial 4.0 license (https://creativecommons.org/licenses/by-nc/4.0/).

INTRODUCTION

In the process of ensuring the fairness of a test administered at different times or different versions of the same standardized test for test takers, test equating studies have emerged. Test equating is a statistical process that involves adjusting test scores to allow different test forms to be used interchangeably. The equating function has five important properties (Dorans & Holland, 2000):

The Same Construct: The tests being equated should measure the same construct,

The Equity: For test takers, it makes no difference whether they take any of the tests to be equated,

The Equal Reliability: Two tests cannot be equated if they measure the same construct but have different reliability,

The Symmetry: The inverse of the equating function equating scores in form X to scores in form Y should equate scores in form Y to scores in form X,

Population invariance: There should be no difference between the selection of sub-populations for X and Y tests, in other words, for the equating function equating X scores to Y scores, the populations of these forms should be invariant.

All these properties are to ensure that the test scores to be equated are used interchangeably. Checking these five properties is important for us to decide whether the equating is appropriate or not. On the other hand, meeting these five properties alone is not sufficient for test equating. This decision also depends on the purpose for which the test will be administered (Kolen & Brennan, 2014). For example, some tests are administered once a year and students are ranked based on the goals of the institution. If the test is administered to identify the highest-performing student, then test equating is not necessary. On the other hand, if several test forms are administered for a common purpose and the differences between the relative item difficulties of these test forms are not intended to affect student assessment, test equating should be used. However, Dorans and Holland (2000) stated that instead of taking these features as a theoretical basis for test equating, we should focus on the question of whether two tests can be equated.

In order to equate test scores, firstly it is necessary to select the appropriate test equating design and test equating method. One of these test equating methods is the Kernel equating described by Holland and Thayer (1989) and later developed by von Davier et al. (2004). Kernel equating differs from other traditional equating methods in that it uses different smoothing approaches to continuousize discrete score distributions.

Kernel equating is a family of equipercentile equating functions and a special case of the linear equating function, so it is a combined test equating approach. This method is so named because of the Kernel function used in nonparametric density estimation (Silverman, 1986; Tapia & Thompson, 1978). Kernel equating involves five steps (von Davier et al., 2004): pre-smoothing, estimation of score probabilities, continuizing the discrete score distributions, equating, and computing the standard error of the equating. In the continuousization step, it is aimed to continuousize the discrete test score distribution. For this, bandwidth parameters need to be chosen. Gaussian Kernel, logistic Kernel or uniform Kernel approaches can be used to continuousize the discrete functions (von Davier et al., 2004).

There are many studies in the literature comparing Kernel equating with other equating methods. In one of these studies, Livingston (1993) compared Kernel equating with other traditional equating methods and found that Kernel equating gives more precise results in terms of standard errors and is more effective than other equating methods with its explicit formula for standard error calculation. In another study comparing Kernel and other observed score equating methods in a real data set, it was found that the difference between Kernel and other traditional equating methods was very small in the equivalent groups design, and in the common-item non-equivalent groups (CINEG) design, Kernel equating gave similar results with equipercentile equating excluding low score ranges (Mao, et al., 2006). In another study conducted with SAT data, Kernel equating results were quite similar to other equating methods (Liu & Low, 2008). However, when the anchor score distributions of the two populations in different forms were similar, it was observed that even the equating methods with different assumptions gave the same or very similar results. In the study examining loglinear presmoothing in terms of equating bias, chained and post-stratification equating methods and Kernel equating were evaluated according to sample sizes, and it was concluded that presmoothing methods with fewer parameters were more biased and

standard error estimation was more precise and accurate in large samples (Moses & Holland, 2007). In the study investigating the effect of atypical extremes on test equating, Kernel estimation yielded more accurate results than traditional equating methods at the ends of the distribution (Cid & von Davier, 2015). There are many studies comparing Kernel equating with other equating methods. In general, Kernel equating has shown similar results with other equating methods. In addition to studies comparing the Kernel equating method with traditional equating methods, there are also studies on the continuization step of Kernel equating.

Liou, et al. (1996) examined the function of simplified formulas to calculate the standard error of the smoothed score distributions that are continuousized using Uniform and Gaussian Kernel functions. The simplified formulas gave good results for equating both observed and smoothed scores. In another study, Lee, and von Davier (2008) examined the impact of different Kernel functions on equating results. Using an equivalent group design, the results show that the characteristics of the tail function of Kernel functions have a large impact on the continuized score distributions. On the other hand, the equated scores obtained using different Kernel functions do not vary much except for the outliers. In another study evaluating the performance of various functions for Kernel density estimation, it was found that uniform Kernel estimation gave poor results compared to other Kernel methods (Soh, et al., 2013). In addition, bandwidth analysis shows that the performance deteriorates as the bandwidth increases.

It has been stated that the choice of the appropriate bandwidth parameter is important in the step of continuousizing the discrete score distribution. The continuousization step, which is considered the most important step of Kernel equating, is very important in terms of ensuring the similarity of the continuized discrete score distribution to the population distribution. Therefore, it is not surprising that studies focused on these issues.

Holland and Thayer (1987) stated that choosing the appropriate bandwidth for the data minimizes the sum of squares of the difference between the continuous distribution and the observed distribution. Moreover, Häggström and Wiberg (2014) emphasized that the choice of band is important because it has a direct impact on the equated scores. It was emphasized that the choice of bandwidth has more influence, especially for extreme scores. Livingston (1993) showed empirically that the bandwidth is not affected by bias in equating for small values. When the sample size is larger than 1000, the standard error formula works quite efficiently. Although the increase in the bandwidth increases the accuracy of the standard error estimation, it is notable that standard error estimations with large bandwidths may be biased in distributions whose score distributions are smoothed using a log-linear model. In another study, a data-adaptive bandwidth tended to be unstable in small samples by minimizing the square of the difference between the observed and continuous distributions (Liou et al., 1996). In this case, an extremely small bandwidth (B = 0.007, and N_X = 100) can be chosen for equating highly scattered distributions.

The bandwidth variable is mathematically complicated as it involves many calculations. In practice, a fixed bandwidth seems to be appropriate for minimizing the square of the differences, but it is clearly still in need of further investigation based on the results of the research. For this purpose, in this study different approaches to selecting the bandwidth parameter used to minimize the penalty function are discussed and presented. In this context, the comparison of three different bandwidth selection approaches (Gaussian, Logistic and Uniform) is considered to be useful in evaluating the accuracy of the continuization step. It is also thought to contribute to the field of "optimal bandwidth selection", which continues to be discussed in the literature. Thus, the research problem is "What is the effect of Uniform, Logistic and Gaussian bandwidth selection approaches on equating results in kernel equating?"

The fact that Kernel equating provides a clear formula for the standard error by using the information in the presmoothing step and allows comparison using the standard error of the difference between the two equating functions gives it an advantage over other equating methods. In this context, the study examined how the equating results of the scores obtained from equivalent test forms according to the Kernel equating method change according to different bandwidths.

METHOD

Equating Design

In the data collection process of the study, a CINEG design was used for test equating. This design is commonly used in exams where only one test form can be administered. For this design, different test forms called old form (Y) and new form (X), with a common set of items (anchor / A) are administered to each of the individuals

of two groups (G_1, G_2) from different populations (P, Q) as shown in Table 1. The common set of items should be as similar as possible to the test forms in terms of both statistical and content characteristics (Kolen & Brennan, 2014).

Table 1. Common-Item Non-Equivalent Groups Design

Population	Sample	**X**	**A**	**Y**
P	G_1	✓	✓	
Q	G_2		✓	✓

Note. P and Q represent different universe, A: anchor items, X : new form, Y : old form

In the study, since the eighth and ninth booklets of the TIMSS 2011 eighth grade mathematics subtest, which had 19 common items, were taken by the whole study group and these common items were included in the scoring together with the other items, a CINEG design with internal common items was used.

Study Group and Data Collection Tools

In the study, the eighth and ninth booklets of the TIMSS 2011 eighth grade mathematics subtest, which have 19 common items, were used. Accordingly, 494 students taking the eighth booklet and 502 students taking the ninth booklet, totally 996 students, constituted the study group. The eighth and ninth booklets of the TIMSS 2011 eighth grade mathematics subtest consisting of 34 multiple-choice questions were used as data collection tools.

Data Analysis

The data were analysed in three stages. In the first stage, moments were obtained by calculating the descriptive statistics of the old form (Y), new form (X) and anchor items (A) used in the study.

Distributions and Moments for the forms of X, Y and A

In this study, Form X scores were equated to Form Y scores. Table 2 shows the descriptive statistics of Forms X and Y, which consist of 19 common items.

Table 2. Descriptive Statistics for Forms X and Y

Form	Mean	S.D.	Skewness	Kurtosis	Min	Max.	N
X	11.051	7.546	0.981	0.008	0	33	494
Y	12.092	8.394	0.843	-0.414	0	34	502

According to the means in the Table 2, it can be said that Form X is more difficult than Form Y. However, according to the skewness and kurtosis values, it was seen that the distributions of the different groups of students who took both forms did not differ much from the normal distribution. When the standard deviation values of the forms X and Y were analysed, it was seen that the variance in Form Y was larger. This table is important in terms of comparing the moments of the score distributions after presmoothing and equating.

In the second stage, Kernel equating was used to obtain equated scores for different bandwidth selection approaches. These bandwidth approaches used in the research were described briefly.

The Gaussian Kernel Method.

This approach requires using a Gaussian Kernel function to continuousize the discrete score distribution in the third step of Kernel equating.

Define $\hat{f}(x_i)$ as a smoothed frequency distribution for the discrete score variable and Φ as the ordinate of a standard normal distribution. The continuous distribution of the random variable x* in the form in R(x_i, x^*) associated with the difference between x_i and x^* is as follows:

$$\hat{f}_{kernel}(x^*) = \frac{1}{constant} \sum_{i=0}^{K} \hat{f}(x_i) \Phi[\, R\,(x_i, x^*)]$$

At each discrete score point, the Kernel equating method uses a normally distributed Kernel to spread the score distribution over the range -∞, +∞. The wider the band parameter, the more intense the distribution at each discrete score point. Although the primary purpose of using a Gaussian Kernel is to make the distribution of scores continuous, it also provides a more uniform distribution of scores. The final distribution of the random variable of x^* is a continuous probability distribution for scores in the range -∞ to +∞. These continuous scores have the same mean and standard deviation as the distribution of discrete smoothed scores. However, scores may differ in kurtosis, skewness, and higher-order moments (Kolen & Brennan, 2014).

The Logistic Kernel Method.

The Logistic Kernel approach uses a logistic function in the third step of Kernel equating, which is the continuization of the discrete score distribution. The following logistic function is used to minimise the penalty function in this step (von Davier, 2010):

$$f_{h_x}^{(1)}(x; r) = \frac{1}{s(a_{xh_x})^2} \sum_i r_i \, k\big(R_{iX}(x)\big)\big[1 - 2K\big(R_{iX}(x)\big)\big].$$

The denser extremes and peaks of the logistic distribution led to larger cumulants than the normal distribution.

Uniform Kernel Method.

The Uniform Kernel approach requires the use of the uniform function in the third step of Kernel equating, which is the continuization of the discrete score distribution. For the optimal bandwidth parameter in the uniform Kernel, the distance between two consecutive possible scores ($2bh_x$) should be close to 1 (von Davier, 2010).

In the last stage of the data analysis, the standard errors of the equated scores were calculated, and the results obtained were compared. Kernel equating has a standard error computation method is provided based on the estimation of standard errors for score probabilities obtained using log-linear models (Anderson et al., 2013). This equation allows to calculate the standard error of the equating for all equating designs: $SEE_Y(x) = \sqrt{VAR(\hat{e}_Y(x))}$.

Another criterion used to compare different equating methods is the Percentage Relative Error (PRE) (von Davier, et al., 2004). PRE is a measure of equating bias. This value, which is obtained by calculating the difference between the moments of the distribution, is an indicator of the distance of the distribution of equated scores from the population distribution (Cid & von Davier, 2015). Before the PRE equation, the moments of Y and $e_Y(X)$ are: $\mu_p(Y) = \sum_k (y_k)^{p s_k}$ and $\mu_p(e_Y(X)) = \sum_j (e_Y(x_j))^{p r_j}$. Accordingly, $PRE_{(p)}$ for the pth moment is calculated as follows: $PRE_{(p)} = 100 \frac{\mu_p(e_Y(X)) - \mu_p(Y)}{\mu_p(Y)}$.

SPSS (version 21) and RStudio Desktop (version 1.4.1106) "*kequate*" package (Andersson et al., 2013) were used to analyze the data.

FINDINGS

In the pre-smoothing step, bivariate observed frequency distributions consisting of test scores and anchor item scores were obtained in accordance with the CINEG design, and both Form X and Form Y raw scores were smoothed according to these frequencies. The distribution of score probabilities was estimated according to the log-linear model. Models were evaluated according to the deviation of goodness-of-fit indices and AIC.

Table 3. Descriptive Statistics and Correlation Values of X, Y and Anchor Tests

	P				Q		
Test Scores	n	Mean	S.D.	Test Scores	n	Mean	S.D.
X	494	11.051	7.546	Y	502	12.092	8.394
A	494	5.988	4.750	A	502	6.436	5.039
Correlation	Form X	Anchor					

Form X	1	0.968
Form Y		0.969

Accordingly, Form X and Form Y showed adequate fit to the observed distribution for the P and Q populations based on the chi-squared values (p>0.05) at the points of $X, X^2, X^3, X^4, A, A^2, A^3, A^4, XA, X^2A^2$ and $Y, Y^2, A^2, A^3, A^4, YA, Y^2A, YA^2$, respectively.

The score probabilities were estimated by setting the weighting coefficients as w = 0.5. For the estimated distributions, bandwidths were determined according to Gaussian, Uniform and Logistic Kernel methods in the continuization step. The values that minimize the penalty function for the bandwidths were calculated. The values of h_x = 0,538, h_y= 0,557 for Gaussian; h_x = 1,0, h_y = 1,0 for uniform; and h_x = 0,395, h_y= 0,418 for logistic Kernel approach.

In the next step, the findings regarding the equated scores according to Gaussian, Uniform and Logistic Kernel equating methods were presented respectively. Figure 1 shows the relationship between the scores equated with the Gaussian Kernel equating method and the raw scores, as well as the difference in the standard error of equating in the CINEG design.

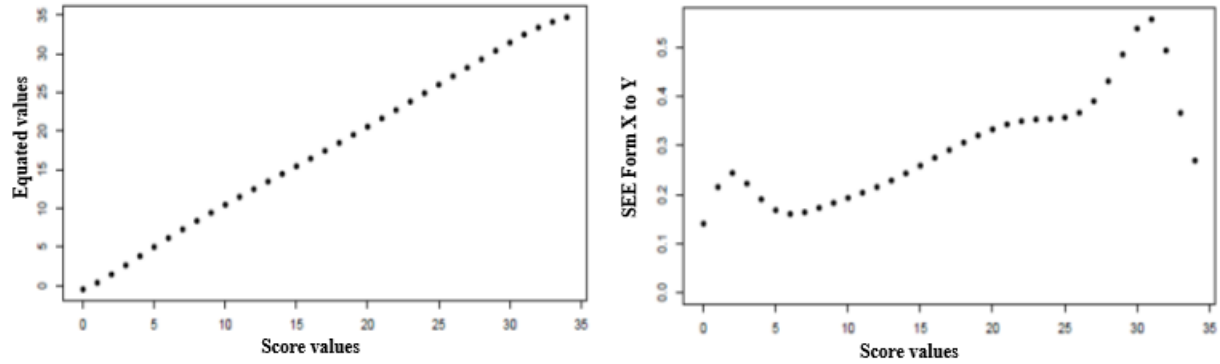

Figure 1. *Raw scores and equated scores according to Gaussian Kernel equating method.*

According to Figure 1, it can be said that the equated scores with Gaussian Kernel equating method are quite similar to the raw scores based on the linear relationship between them. When the standard error values in the next figure are examined, it is seen that the standard error is large at the lower values of the scale; however, it reaches maximum at the upper end. This indicates that the number of students who had very high scores on the test and those who had low scores may be small.

In the next step, the results were obtained according to the Uniform Kernel equating method. Figure 4 shows the relationship between the scores equated using the Uniform Kernel equating method and the raw scores and the difference in the standard error of the equating in the common-item non-equivalent groups design.

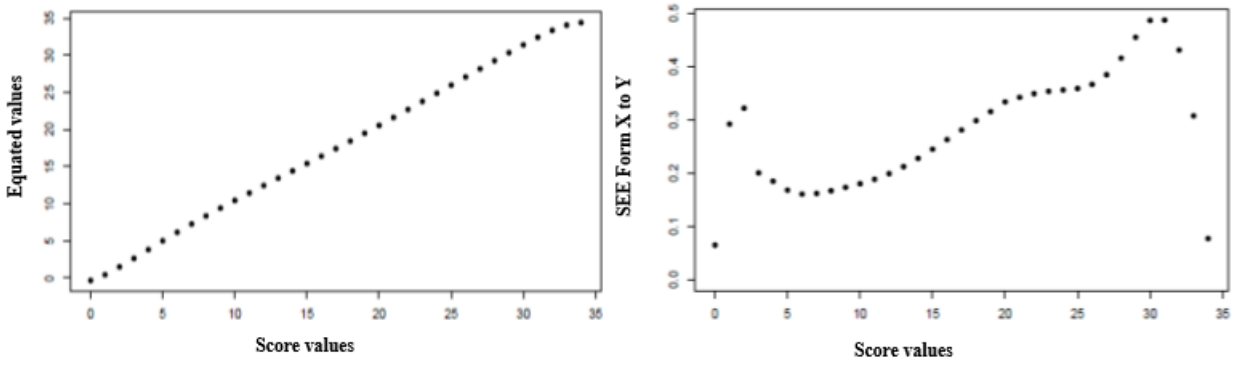

Figure 2. *Raw scores and equated scores according to Uniform Kernel equating method.*

The relationship in Figure 2 showed that the scores equated using the uniform Kernel equating method and the raw scores were quite similar. In addition, when the standard errors of the equated scores were examined based

on their distribution, it is seen that the standard errors were large in the range of 0-5 and 30-34 scores. In the range of 5-30, the standard error tends to increase.

Finally, equating results were obtained according to the Logistic Kernel equating method. Figure 5 shows the relationship between the scores equated using the Logistic Kernel equating method and the raw scores and the difference in the standard error of the equating in the CINEG design.

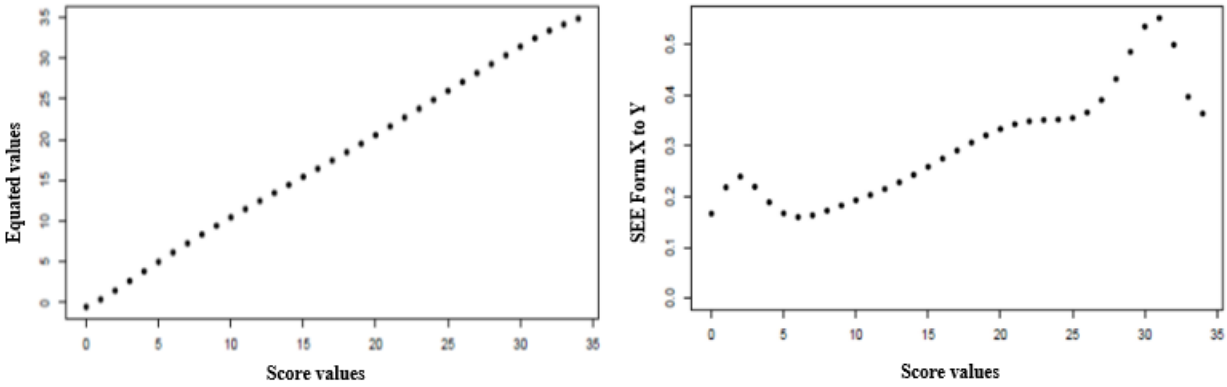

Figure 3. *Raw scores and equated scores according to Logistic Kernel equating method.*

The relationship between the scores obtained by logistic Kernel equating and the raw scores was linear, as shown in Figure 3, and thus the scores were quite similar. When the standard error of the logistic Kernel equating was examined, it was seen that the standard error is large at the lower and especially at the upper endpoints.

In the last step of Kernel equating, the standard error of the equating function can be estimated, as well as the standard error of the equating difference (SEED), which allows the comparison of different equating results. SEED values, which allow pairwise comparison of equating results, were compared in the following order: Logistic-Gaussian, Uniform-Gaussian, Uniform-Logistic.

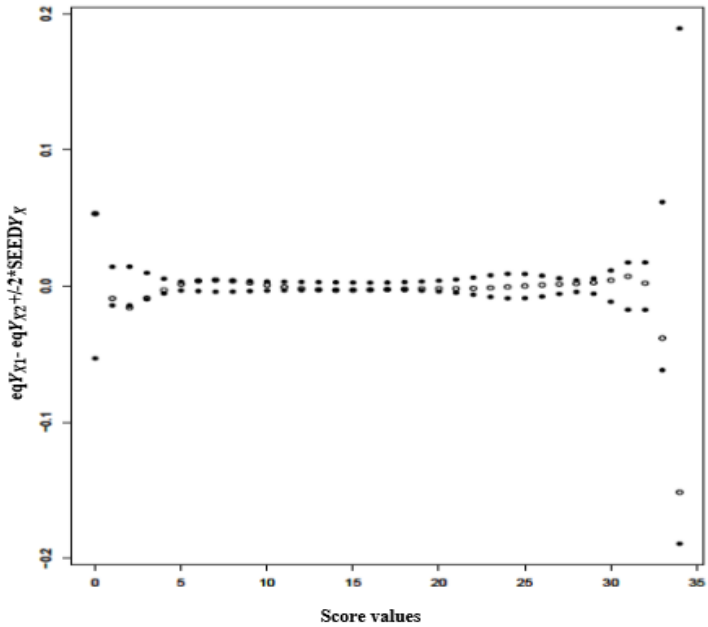

Figure 4. The difference between Gaussian Kernel and Logistic Kernel in relation to SEED for each score range.

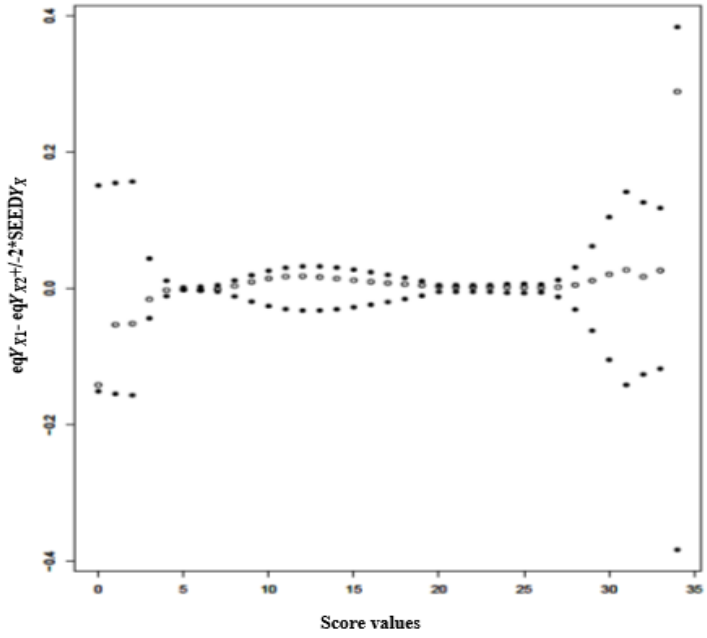

Figure 5. The difference between Gaussian Kernel and Uniform Kernel in relation to SEED for each score range.

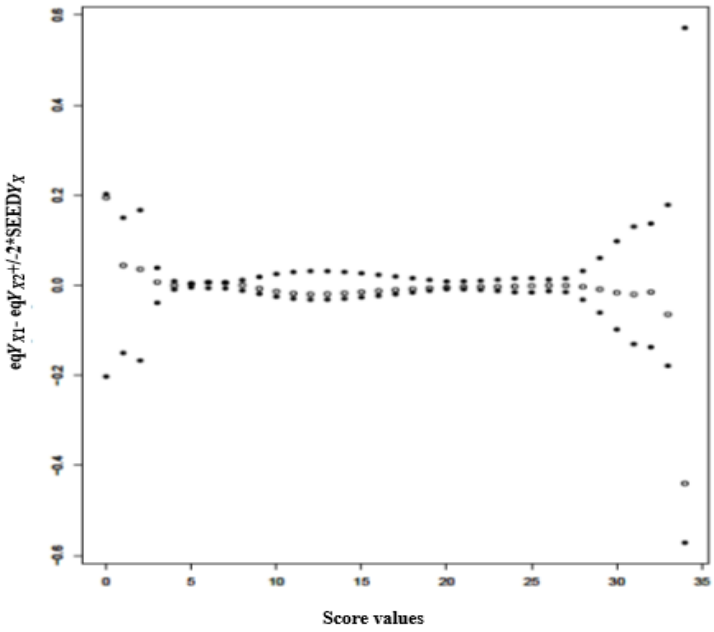

Figure 6. The difference between Logistic Kernel and Uniform Kernel in relation to SEED for each score range.

Figure 4, Figure 6 and Figure 6 showed that for each pairwise comparison, most of the difference between the equated values did not exceed the intervals of +2 and -2 standard error band. In this case, it was seen that there is a high level of consistency between the scores equated according to all three Kernel equating approaches. This is also clearly observed in Figure 7. Additionally, the equated scores obtained according to all three methods and standard errors of equated scores can be examined in Appendix 1 and Appendix 2.

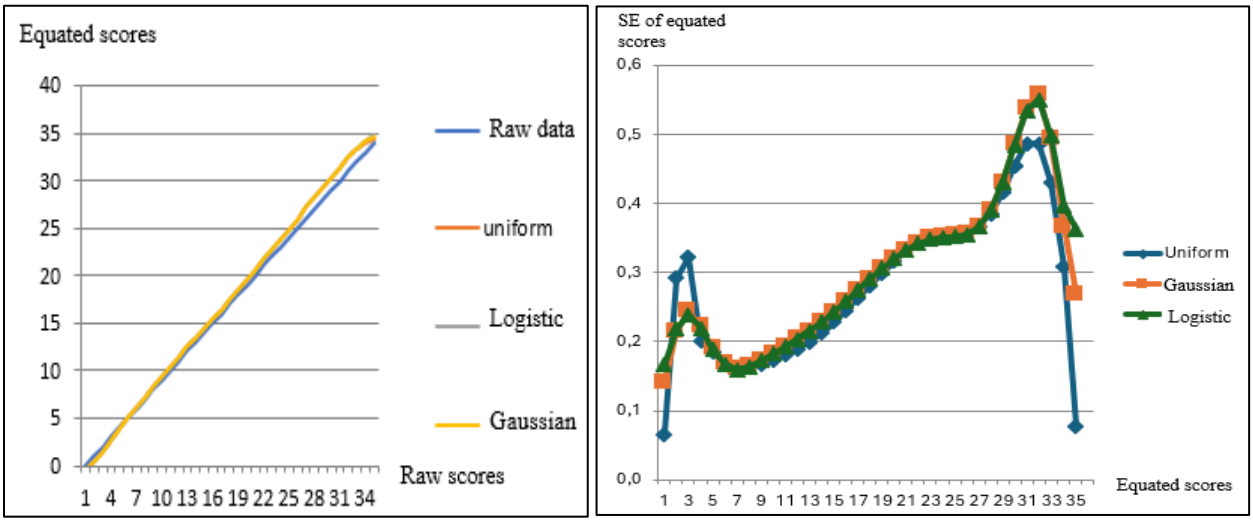

Figure 7. *Equated scores and standard errors for the three Kernel approaches*

The first graph in Figure 7 showed that the three different Kernel approaches in this study give almost the same result. However, according to the second graph, the standard errors differ for low and high values. While the Logistic and Gaussian Kernel approaches were more in line with each other in terms of standard error values, the standard error of the Uniform Kernel approach tended to be higher at low scores and lower than the other approaches at high scores.

Between the score range of 5 and 30, the standard errors of the equated scores according to all three approaches are almost the same. For extreme values, peaks were observed in the standard error values. In this case, it can be said that the number of individuals scoring in this range is low.

In the study, percentage relative error (PRE) was also calculated to compare the moments of distribution of the scores equated according to different Kernel approaches with the moments of distribution of the Y-form scores in the target population.

Table 4. Percentage Relative Error (PRE) of the Equated Score Distribution

	Equating Methods		
	Gaussian PRE ($\boldsymbol{Y_x}$)	Logistic PRE ($\boldsymbol{Y_x}$)	Uniform PRE ($\boldsymbol{Y_x}$)
1	0.003	0.006	-0.010
2	0.002	0.005	-0.077
3	0.016	0.023	-0.099
4	0.030	0.042	-0.120
5	0.044	0.061	-0.152
6	0.060	0.084	-0.194
7	0.079	0.115	-0.244
8	0.106	0.157	-0.300
9	0.140	0.212	-0.359
10	0.183	0.283	-0.420

The PRE values in Table 4 are close to each other for all three Kernel equating approaches. This supports the graphs in Figure 9. The deviation percentages of the first 10 moments from the target population vary between

0.002-0.183 for Gaussian Kernel equating, 0.005-0.283 for Logistic Kernel equating and 0.009-0.420 for Uniform Kernel equating. Considering the PRE values, it can be concluded that the equating results obtained from the bandwidth selected according to the Gaussian Kernel approach show a better fit to the distribution in the population.

DISCUSSION AND CONCLUSION

In this study, using a real data set, the equated scores according to different bandwidth selection strategies in the continuization step of Kernel equating are comparatively examined. Gaussian, Logistic and Uniform Kernel bandwidth selection approaches were used. PRE, Standard Error and SEED were used as criteria for comparing the equated results.

It was observed that the scores obtained with three different Kernel equating under the CINEG design with common items were quite similar to the raw score distribution in all three methods. When all three equating approaches were compared according to standard errors, it was observed that the error distributions were quite similar to each other, but the Logistic and Gaussian Kernel approaches gave closer results. However, as in Lee and von Davier (2008), these methods differed only at the endpoints. Häggström and Wiberg's study (2014) also emphasized that the choice of bandwidth has more impact, especially for endpoints. When the three equating approaches are compared according to the percentage relative error, it is seen that although the PRE values are close to each other, the Gaussian Kernel equating has a relatively lower PRE value.

The results of the study show that Gaussian Kernel equating has a lower percentage relative error, but in terms of standard errors, the standard error distributions obtained using Gaussian Kernel and Logistic Kernel equating are quite similar. This finding is also supported by von Davier (2011) who compared the results of Gaussian, Logistic and Uniform Kernel equating. In addition, another study by Liu et al. (1996) found no difference between Uniform and Gaussian Kernel methods, while Soh et al. (2013) found that Uniform Kernel estimation gives poor results compared to other Kernel methods.

According to the results of the study, it can be said that the effect of different bandwidths selected according to Gaussian Kernel, Logistic Kernel and Uniform Kernel approaches on the equating results is almost similar. As mentioned in many studies (Andersson, 2014; Liang & von Davier, 2014; Wang, 2008) where bandwidth selection approaches are developed, the choice of this parameter aims to minimize the penalty function in the continuization step and keep the fit distribution quite close to the distribution in the target population. In this respect, the choice of the bandwidth parameter is important for the continuization step of Kernel equating. In this context, in line with the findings of this study, researchers may be advised to use Gaussian Kernel or Logistic Kernel approach, which are less affected by the score distribution, in selecting the bandwidth parameter. However, bandwidth selection can also be determined based on cross validation techniques. In studies comparing Gaussian, Logistic and Uniform Kernel approaches based on minimizing the penalty function with techniques based on cross-validation, the results are similar with minor differences (Andersson & von Davier, 2014; Häggström & Wiberg, 2014; Wallin et al., 2021).

One of the limitations of this study is the use of PRE, Standard Error, and SEED as criteria for comparing bandwidth selection methods. It has been stated that PRE does not provide much information about bandwidth selection (Häggström & Wiberg, 2014). The criteria should be enhanced with mean squared error and other bias indicators so that the basis for selecting the appropriate bandwidth becomes stronger.

The bandwidth selection approaches in this study are based mainly on minimizing the penalty function. Due to the complexity of minimizing the penalty function, researchers have proposed simplified methods derived from the modified standard errors of equating (Andersson & von Davier, 2014) and cross-validation techniques (Liang & von Davier, 2014). These different approaches can be compared with approaches based on minimizing the penalty function.

In this study, the effects of three different approaches were examined under a CINEG design; studies comparing these methods under different designs can be conducted. In addition, different bandwidth selection methods can be compared with simulation data to examine the effect of sample size and score distributions in Kernel equating. Furthermore, how bandwidth selection methods perform at different test lengths could be examined.

Statements of Publication Ethics

The data of this study were taken from the TIMSS and PIRLS official web page, which is open to everyone. Therefore, it does not require ethics committee permission. During the research, the process was carried out by paying attention to the rules of research ethics.

Conflict of Interest

The corresponding author states that there is no conflict of interest.

REFERENCES

Andersson, M. 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. Doi: 10.18637/jss.v055.i06

Andersson, B. (2014). Contributions to kernel equating. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences, 106*(24).

Andersson, B., & von Davier, A. A. (2014). Improving the bandwidth selection in kernel equating. *Journal of Educational Measurement, 51(3), 223-238.* https://doi.org/10.1111/jedm.12044

Cid, J. A., & von Davier, A. A. (2014). Examining potential boundary bias effects in Kernel smoothing on equating an introduction for the adaptive and Epanechnikov Kernels. *Applied Psychological Measurement*, 39(3) 208–222. https://doi.org/10.1177/0146621614555901

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of educational measurement*, *37*(4), 281-306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Häggström, J. & Wiberg, M. (2014), Optimal bandwidth selection in observed-score kernel equating. *Journal of Educational Measurement*, 51: 201-211. https://doi.org/10.1111/jedm.12042

Holland, P. W., & Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions. *ETS Research Report Series*, (2), 1-40. https://doi.org/10.1002/j.2330-8516.1987.tb00235.x

Holland, P. W., & Thayer, D. T. (1989). The Kernel method of equating score distributions. ETS Research Report Series, (1), 1-45. https://doi.org/10.1002/j.2330-8516.1989.tb00333.x

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking.* Springer.

Lee, Y. H., & von Davier, A. A. (2008). Comparing alternative Kernels for the Kernel method of test equating: Gaussian, logistic, and uniform Kernels. *ETS Research Report Series*, *2008*(1), i-26. https://doi.org/10.1002/j.2333-8504.2008.tb02098.x

Liang, T., & von Davier, A. A. (2014). Cross-Validation an Alternative Bandwidth-Selection Method in Kernel Equating. *Applied Psychological Measurement*, *38*(4), 281-295. https://doi.org/10.1177/0146621613518094

Liou, M., Cheng, P. E., & Johnson, E. G. (1996). Standard errors of the Kernel equating methods under the common-item design. *ETS Research Report Series*, (1), 1-36. https://doi.org/10.1177/01466216970214005

Liu, J., & Low, A. C. (2008). A comparison of the Kernel equating method with traditional equating methods using SAT® Data. *Journal of Educational Measurement*, *45*(4), 309-323. https://doi.org/10.1111/j.1745-3984.2008.00067.x

Livingston, S. A. (1993). An empirical try out of Kernel equating. *ETS Research Report Series, (2),* 1-9.

Mao, X., Davier, A. A., & Rupp, S. (2006). Comparisons of the Kernel equating method with the traditional equating methods on Praxis™ Data. *ETS Research Report Series*, (2), 1-31. https://doi.org/10.1002/j.2333-8504.2006.tb02036.x

Moses, T., & Holland, P. (2007). Kernel and traditional equipercentile equating with degrees of presmoothing. *ETS Research Report Series*, (1), 1-39. https://doi.org/10.1002/j.2333-8504.2007.tb02057.x

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (26). CRC Press.

Soh, Y., Hae, Y., Mehmood, A., Ashraf, R. H., & Kim, I. (2013). Performance evaluation of various functions for Kernel density estimation. *Open J Appl Sci*, *3*(1), 58-64.

Tapia, R. A., & Thompson, J. R. (1978). *Nonparametric probability density estimation*. Johns Hopkins University Press.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). Kernel Equating versus Other Equating Methods. *The Kernel Method of Test Equating,* 87-95. Springer.

von Davier, A. (Ed.). (2010). *Statistical models for test equating, scaling, and linking*. Springer.

Wallin, G., Häggström, J., & Wiberg, M. (2021). How important is the choice of bandwidth in kernel equating?. *Applied Psychological Measurement*, *45*(7-8), 518-535.

Wang, T. (2008). The continuized log-linear method: An alternative to the Kernel method of continuization in test equating. *Applied Psychological Measurement.* https://doi.org/10.1177/0146621607314043

APPENDIX

Appendix 1. Results of Kernel Equating According to Different Bandwidth Selection Approaches

Raw Scores	Uniform	Logistic	Gaussian
0	-0.332	-0.527	-0.474
1	0.431	0.386	0.377
2	1.504	1.468	1.453
3	2.648	2.641	2.633
4	3.829	3.829	3.826
5	5.005	5.002	5.003
6	6.156	6.150	6.154
7	7.275	7.270	7.274
8	8.361	8.361	8.365
9	9.416	9.424	9.426
10	10.447	10.461	10.462
11	11.458	11.476	11.476
12	12.455	12.475	12.473
13	13.443	13.463	13.460
14	14.429	14.446	14.443
15	15.417	15.432	15.429
16	16.413	16.426	16.423
17	17.422	17.433	17.430
18	18.447	18.456	18.454
19	19.489	19.496	19.494
20	20.547	20.552	20.550
21	21.618	21.622	21.620
22	22.699	22.702	22.701
23	23.787	23.790	23.789
24	24.880	24.882	24.882

25	25.976	25.977	25.978
26	27.074	27.074	27.075
27	28.169	28.170	28.171
28	29.260	29.263	29.265
29	30.341	30.350	30.352
30	31.403	31.419	31.423
31	32.424	32.444	32.451
32	33.350	33.365	33.368
33	34.068	34.132	34.094
34	34.402	34.842	34.691

Appendix 2. Standard Errors of Equated Scores According to Different Bandwidth Selection Approaches

Raw Scores	**Uniform**	**Gaussian**	**Logistic**
0	0.066	0.141	0.167
1	0.292	0.216	0.218
2	0.322	0.244	0.239
3	0.201	0.223	0.220
4	0.185	0.191	0.189
5	0.169	0.168	0.167
6	0.161	0.161	0.160
7	0.162	0.164	0.163
8	0.167	0.173	0.172
9	0.174	0.183	0.183
10	0.181	0.194	0.193
11	0.189	0.204	0.203
12	0.200	0.216	0.215
13	0.213	0.229	0.228
14	0.228	0.243	0.243
15	0.245	0.259	0.259
16	0.263	0.275	0.275
17	0.281	0.291	0.291

18	0.299	0.306	0.306
19	0.316	0.321	0.321
20	0.334	0.333	0.333
21	0.343	0.343	0.343
22	0.349	0.350	0.348
23	0.354	0.353	0.351
24	0.356	0.355	0.352
25	0.359	0.357	0.355
26	0.367	0.367	0.366
27	0.385	0.391	0.390
28	0.416	0.431	0.431
29	0.455	0.486	0.485
30	0.487	0.539	0.535
31	0.487	0.557	0.551
32	0.431	0.494	0.499
33	0.308	0.367	0.396
34	0.078	0.270	0.364