Eurasian Journal of Educational Research

www.ejer.com.tr

A Comparison of Kernel Equating and Item Response Theory Equating Methods*

Çiğdem AKIN-ARIKAN[1] Selahattin GELBAL[2]

ARTICLE INFO

Article History:

Received: 20. Jan.2020

Received in revised form: 28. Aug.2020

Accepted: 16.Dec.2020

DOI: 10.14689/ejer.2021.93.9

Keywords

Equating, kernel, IRT, error

ABSTRACT

Purpose: This study aims to compare the performances of Item Response Theory (IRT) equating and kernel equating (KE) methods based on equating errors (RMSD) and standard error of equating (SEE) using the anchor item nonequivalent groups design. **Method:** Within this scope, a set of conditions, including ability distribution, type of anchor items (internal-external), the ratio of anchor items, and spread of anchor item difficulty, were observed in 24 different simulation conditions. **Findings:** The results showed that ability distribution, type of anchor items, the ratio of anchor items, and spread of anchor item difficulty affected the performance of the equating methods. It was also observed that kernel chained equating methods (KE CE) were less affected by the difference in group mean ability. Moreover, in the case of increased average differences in ability between groups, a high range of score scale yielded higher standard errors in KE methods, while a medium-high range of scale scores exhibited higher standard errors in IRT equating. Using external anchor items led to lower SEE and RMSD than using internal anchor items, and both errors decreased as the ratio of anchor items increased. When internal anchor items were used with similar average group ability distribution, mini and midi anchor tests gave similar results. On the other hand, a midi anchor test performed better with increased average differences in group ability distribution for external anchor items. At the end of the scale scores, the IRT equating method had a lower rate of errors. **Implications for Research and Practice**: KE methods can be used while IRT assumptions are not met.

*This study is a part of the doctoral dissertation "Comparison of kernel equating and item response theory equating methods" prepared in the Educational Measurement and Evaluation Program, Hacettepe University, Turkey.

[1] Ordu University, TURKEY, e-mail: akincgdm@gmail.com, ORCID: 0000-0001-5255-8792

[2] Hacettepe University, TURKEY, e-mail: sgelbal@gmail.com, ORCID: 0000-0001-5181-7262

Introduction

Tests are used for many different purposes, such as collecting information about individuals, measuring traits such as interests, abilities, and attitudes, and selecting individuals. In most testing situations, especially for large-scale educational assessments, the need for different forms of the same test for testing examinees on different administrations is an important requirement to be met for testing security. However, multiple alternate test forms may differ in difficulty. This may lead to discrepancy among examinees taking different test forms in that those answering the easy version get higher scores than those answering the difficult version. Test forms need to be equated due to differences between test difficulties (Kolen & Brennan, 2004; von Davier, Holland, & Thayer, 2004). Equating refers to associating with or converting test scores into scores obtained from another test form (Hambleton & Swaminathan, 1985).

Before beginning test equating, decisions should be made about the equating design to be used. Test equating designs are single group, equivalent group, and nonequivalent group with anchor tests (NEAT), which was selected as equating design in this study (Livingston, 1993; von Davier et al., 2004). For the NEAT design, different test forms which have anchor (common) items are applied to two groups with nonequivalent ability distribution. Test forms are equated by attempting to remove differences between ability distributions in the groups through the anchor items (Kolen & Brennan, 2004).

The NEAT design can be used in different equating methods. These methods are based on classical test theory (CTT), item response theory (IRT) and kernel equating. Methods based on CTT are linear (e.g., Tucker, Levine true and observed scores, Braun Holland) and equipercentile (frequency estimation, chained) equating methods. Linear and equipercentile chained equating (CE), linear and equipercentile post-stratification equating (PSE), and Levine linear equating methods are included in kernel equating. Equating methods based on IRT are classified as IRT true scores and IRT observed score equating. In this research, information is provided about these methods when used for the kernel equating and IRT true-score equating.

Equating Methods

Kernel equating

In the equipercentile equating method, scores corresponding to the same percentage rank are considered equal. For this, first, the cumulative frequency of each form is calculated. Second, scores corresponding to the same percentage scores according to these cumulative frequencies are equated. In equipercentile equating, examinees obtaining the equated scores in the same percentile rank are assumed to have the same ability level (Kolen, 1988). It is nearly impossible for these individuals to have the same ability level in real applications. Thus, kernel equating was originally developed to solve this problem that occurs in equipercentile equating (von Davier et al., 2004). The main reason for this problem is the discrete distribution of scores.

Holland and Thayer (1981) found a solution for this problem by transforming discrete distributions into continuous distributions using kernel equating. In kernel equating, discrete distributions are equated through continuous distributions (Livingston, 1993; Ricker & von Davier, 2007). Kernel equating uses the Gaussian kernel approach (Lee & von Davier, 2010; von Davier et al., 2004). Although kernel is an equipercentile equating method, it includes linear equating methods at the same time (Andersson & von Davier, 2014; von Davier, 2008). Selection of bandwidth (parameter h) as one of the parameters used in kernel equating states use of an equipercentile or linear equating method. If the ideal bandwidth is used, the equating results approximate to equipercentile equating; on the other hand, if large bandwidth is used, they approximate to the linear equating method (Ricker & von Davier, 2007; von Davier et al., 2006). Kernel equating consists of five stages: pre-smoothing, estimation of score distributions, continuization, equating, and calculating the standard error of equating.

Equating Based on Item Response Theory

IRT explains the abilities of individuals with mathematical models. Lord (1953) stated that true and observed scores do not mean the same as ability score, and ability score is independent of the test, whereas true and observed scores are dependent on the test (as cited in Hambleton & Jones, 1993). Item parameters with two response categories are estimated using one-parameter logistic (1PL) model, two-parameter logistic (2PL) model and three-parameter logistic (3PL) model (Embretson & Reise, 2000).

After deciding the IRT model, the parameters of item and ability are estimated. In NEAT design, A (slope) and B (intercept) linking coefficients are obtained using the parameters (a and b) of the anchor items, which are answered in both groups. Using these coefficients, the θ value in the test form is converted to the θ value for the other test form. For converting estimations obtained from a test form to estimations obtained from the other test in IRT, two methods are used of separate and concurrent calibration. Separate calibration methods are divided into two as characteristic curve and moment methods. The characteristic curve methods were developed to reduce the difference between item characteristic curves for anchor items and Haebara, one of the characteristic curve methods, is used in this research (Kolen & Brennan, 2004). In this method, the differences between item characteristic curves for a given ability level are calculated by summing up the squared differences between item characteristic curves for each item.

Purpose of this Study

Before using the kernel equating method, it is important to compare the results of equating methods frequently used in test applications and determine whether the results have similarities and differences (Mao, von Davier, & Rupp, 2006). Knowledge about the strengths and weaknesses of equating methods facilitates the selection of an appropriate equating method required for test programs. Also, it is important to know how the choice of one method over another method affects the decision since important decisions are made about individuals based on the equating results obtained from large-scale and high-risk tests (Kim & Cohen, 2002). Due to the popularity of test equating implementations, it is considered important to reveal which equating method provides the best results according to certain conditions.

Since equating methods differ concerning the theories and assumptions on which they are based, the choice of an equating method is of great importance for both test developers and examinees. There are few studies in the literature comparing IRT and kernel equating methods (e.g., Chen, 2012; Godfrey, 2007; Meng, 2012; Norman Dvorak, 2009). In these studies, only kernel equipercentile equating methods were compared with IRT equating methods of concurrent calibration, Stocking and Lord's Method, and Mean/Sigma transformation methods. In this study, kernel linear and equipercentile equating methods were compared with IRT true score equating methods of Haebara. Besides, the choice of anchor items is very important for equating tests under NEAT design. Many researchers (e.g. Budescu, 1985; Kolen, 1988; Kolen, 2007; Petersen, Kolen, & Hoover, 1989) argue that the anchor item set should be a small version of the test. However, in practice, it is difficult to create similar forms of difficulty distribution in anchor tests. Sinharay and Holland (2006a, 2006b, 2007) argue that there is no evidence that the anchor test must have the same difficulty distribution as the total test since it is quite a restrictive condition; yet, better equating results could be obtained if the content representativeness is the same as the test while the spread of anchor item difficulty is smaller than that of the total test. Moreover, Sinharay and Holland (2007) stated that if a spread of anchor item difficulty is used in external common tests, lower equating errors will be obtained. As a result, it is considered important that different equating methods note the spread of anchor item difficulty and the effects of common test types.

Due to these reasons, it is considered important to compare the performance of kernel post-stratification equipercentile (ideal bandwidth), kernel post-stratification linear (large bandwidth), kernel chained equipercentile (ideal bandwidth), and kernel chained linear (large bandwidth) equating methods with the IRT true-score equating method, which is frequently used so that advantages and disadvantages of the equating methods are explored along with conditions. For this purpose, answers were sought for the problem below.

When tests are equated according to kernel post-stratification linear, kernel chained linear, kernel post-stratification equipercentile, kernel chained equipercentile, and IRT true-score equating methods;

a) How does equating error change with ability distribution, the ratio of anchor item, and spread of anchor item difficulty for internal anchor tests?

b) How does equating error change with ability distribution, the ratio of anchor item, and spread of anchor item difficulty for external anchor tests?

a) How does standard equating error change with ability distribution, the ratio of anchor item, and spread of anchor item difficulty for internal anchor tests?

b) How does standard equating error change with ability distribution, the ratio of anchor item, and spread of anchor item difficulty for external anchor tests?

Method

Research Design

In this research, simulation data were compared for the effects of the ratio of anchor items, the spread of anchor item difficulty, ability distribution (average group differences in ability), type of anchor items (internal or external), and various equating methods on equating error. As a result, this research is descriptive as it investigates the effects of equating methods in the study conditions in detail and reveals which method provides better results in which conditions. Descriptive research includes studies defining a situation as much as possible without deficiency and with great care (Fraenkel, Wallen, & Hyun, 2012).

Simulation Conditions

Holland, Dorans, and Petersen (2006) stated that the quality of main tests, characteristics of anchor test, sample size, ability distributions of groups, and selection of the equating method are the main considerations for successful equating. In this study, the effects on equating error of ability distribution, the ratio of anchor item, type of anchor test, and spread of item difficulty conditions were investigated. NEAT design test forms (X and Y) and anchor test forms were generated for this purpose. In doing so, form X refered to the old test form, while Y is the (new) test form to be equated. The groups given the form X and form Y were called Group 1 and Group 2, respectively. The NEAT design is shown in Table 1.

Table 1

Equating Design

Group	X	Y	Anchor Test
Group 1	✓		✓
Group 2		✓	✓

Sample Size and Test Length: The sample size was dealt with as a single condition. Each test form was applied to equal numbers of individuals of 1,500 and a total of 3,000 individuals were analyzed in this study. Kolen and Brennan (2004) have emphasized that for IRT, equating a sample size of 1,500 requires the 3PL model under the NEAT design. Kolen and Brennan (2004) emphasized the need for at least 30-40 items as test length for equating tests. In this study, a single total test length of 50 items was used. As a result, the test length was sufficient to determine test equating.

Ability Distribution: The abilities of groups in the NEAT design are not equivalent. As a result, the ability distribution of both groups was produced differently. In this study, the data were generated so that the new form group could have a standard normal distribution with ability distribution of ($\theta \sim N$ (.05,1)) and ($\theta \sim N$(.5,1)), while the group receiving the old form could have ($\theta \sim N$ (0,1)). The relevant literature suggests that groups should have similar ability distribution as a prerequisite for equating. Wang, Lee, Brennan, and Kolen (2008) regard the difference between the mean ability distributions of groups in the range of .05 and .10 as large, but they refer to it as too large if it falls in the range of .25 and greater.

Anchor Test: According to test length, three common item rates were used; 20%, 30%, and 40%. Angoff (1971) and Budescu (1985) suggested that a minimum of 20% of the item numbers on the test should be anchor items. Also, Kolen and Brennan (2004) have proposed that at least 20% of the total test length should be used as the ratio of common items in tests with a length of 40 items or longer.

Type of Anchor Test: The anchor test is divided into two as internal and external anchor test and both types were used in this study. The internal anchor test adds scores from the common test to the total scores, while the external anchor test does not add to the total test scores.

Spread of Difficulty Levels of Anchor Items: The other factor in the study is the spread of difficulty levels of anchor items. Sinharay and Holland (2006a) propose that if the anchor test has the same scope and similar statistics as the total test, it is called a mini-test, and if the anchor test items have different difficulties and all item difficulties are medium, it is called a midi test. Mini and midi anchor tests with different distributions of anchor item difficulty were used in this study.

Data Generation and Analysis of Data

The item responses produced with the 3PL model used the R program (R Core Team, 2016). In the first stage, ability distributions for the groups were produced from standard normal distribution according to the determined conditions. In this study, the same number of examinees was used for tests. In the second stage, the two test forms using the NEAT design and anchor items were produced. The item discrimination parameter (a) of the test and anchor test forms were generated from uniform distribution (e.g., between *U*(.5-2); parameter c was generated from uniform distribution (e.g., between *U*(.05-.2); lastly, parameter b, which refers to item difficulty parameter was generated from standard normal distribution (e.g., *N*(0, 1)). In the case of a mini anchor test, it was generated to ensure the same difficulty parameter as the total test, which was an identical mean and standard deviation. In the case of a midi test, it was derived from the same mean as the total test but the standard deviation was .2. Item and examinee parameters were generated by using the "*irtoys*" package (Partchev, 2016) with R program. The third stage was the test equating process. The "*kequate*" package (Andersson, Branberg, & Wiberg, 2013) was used to equate test forms with kernel equating methods. There are three steps to obtain equating results.

Firstly, the creation of frequency distributions and then fitting to a generalized linear model and in the last step performing the equating. The bandwidth selection for linear and equipercentile methods were selected by "kernel" package with Gaussian kernel smoothing. To conduct item and ability parameter estimations, the "*ltm*" package (Rizopoulos, 2015) was used for IRT. Estimations were made by using Marginal Maximum Likelihood for item parameter estimation (MMLE) and Expected a Posterior (EAP) methods for estimation of ability parameters. IRT true-score equating with Haebara method, which is a separate calibration method, was performed by using the "*plink*" package (Weeks, 2010). In this study, a total of 24 conditions (*2 ability distribution × 3 ratio of anchor item × 2 type of anchor item × 2 spread of anchor item difficulty*) were investigated according to five different equating methods and each analysis was replicated 100 times.

Two different evaluation criteria of RMSD (root-mean-square difference) and standard error of equating (SEE) were used to assess the accuracy of the equating results. Mao et al. (2006) obtained the RMSD index by adapting the RMSE index. RMSD index reflects how biased or accurate the equating results are against the equating criterion (Qu, 2007). In this study, the equipercentile equating method in the EG design was used as the equating criterion.

$$\mathrm{RMSD} = \sqrt{\bar{\mathrm{d}}^2 + \mathrm{sd}_{\mathrm{d}}^2} \qquad (1)$$

($\overline{\mathrm{d}}$): The mean of the difference between the criterion equating and equating method for each equated score,

sd: Standard deviation of the difference obtained.

SEE which gives the random error, is equal to the square root of the ratio of the sum of the squares of the difference of the mean value obtained with each estimated value to the number of replication.

$$\mathrm{SE}[\hat{\mathrm{e}}_{\mathrm{Y}}(\mathrm{x_i})] = \sqrt{\frac{1}{\mathrm{R}}\sum_1^{\mathrm{R}}[\hat{\mathrm{e}}_{\mathrm{Y}}(\mathrm{x_i}) - \overline{\hat{\mathrm{e}}_{\mathrm{Y}}}(\mathrm{x_i})]^2} \ (2)$$

$\hat{\mathbf{e}}_{\mathbf{Y}}(\mathbf{x_i})$:Equated score obtained for each replication

$\overline{\hat{\boldsymbol{e}}_{\boldsymbol{Y}}}(\boldsymbol{x_i})$: Mean of equated scores obtained through replication

R : Number of replications

Results

To understand the findings, the results of the equating error and standard error of equating under the conditions covered in this research were structured separately for internal and external anchor tests. The average bandwidths (hx and hy) were between varied from .607 to .689 for KE Chaied equipercentile, .645 to .577 for KE PSE equipercentile, 8461.722 to 9720.298 for KE PSE equipercentile and 8051.256 to 9771.645 for KE CE equipercentile.

Findings for Equating Error

The graphs of the equating error reflecting the conditions discussed here are given in Figure 1 and Figure 2. CE-EQ refers to the kernel (KE) chained equipercentile equating, CE-L is KE chained linear equating, PSE-EQ is KE post-stratification equipercentile equating, PSE-L is KE post-stratification linear equating, and IRT is IRT true-score equating method.

Findings of equating error for internal and external anchor tests

Figure 1 and Figure 2 display graphs of equating errors obtained when the ratio of anchor items is 20%, 30%, and 40% for internal and external anchor tests when the ability distribution of groups is similar and different, respectively.

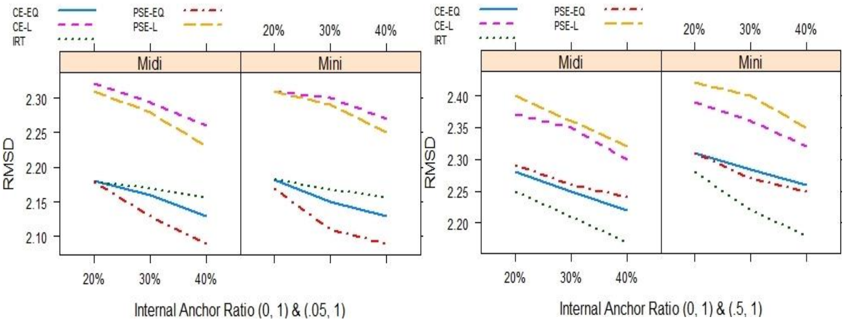

Figure 1. *RMSD Values with Similar (first figure) and Different (second figure) Group Mean Ability Distribution on Internal Anchor Test*

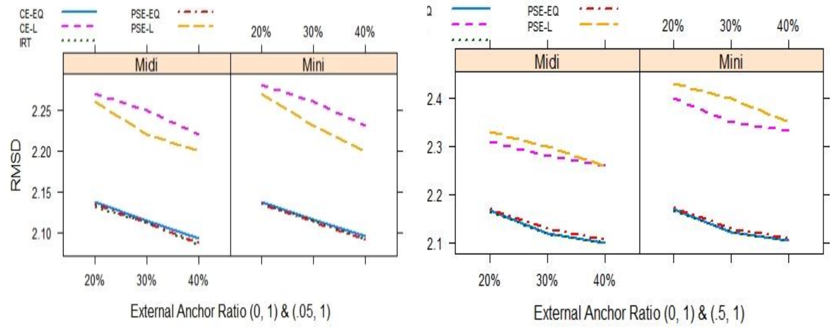

Figure 2. *RMSD Values from Similar and Different Group Mean Ability Distribution on External Anchor Tests *CE-EQ: chained equipercentile equating, CE-L: chained linear equating, PSE-EQ: post-stratification equipercentile equating, PSE-L: post-stratification linear equating, IRT: IRT true-score equating method.*

In Figures 1 and 2, the total error decreased as the ratio of anchor items increased in the internal and external anchor tests with all equating methods. The rate further decreased for linear equating methods with similar ability distribution as the ratio of anchor items increased in the internal anchor test. In all conditions, the lowest error was found for IRT equating when the ability distribution between the groups was different.

As for the external anchor test, the error rate decreased more as a result of using the linear equating methods but still had a higher error. When the external anchor test used the midi test, the error was found to be even lower than for the mini test condition, especially when ability distribution was similar. However, regular error values could not be obtained when the internal anchor test was a midi test. In the case where the distribution of ability between groups was similar for both internal and external anchor tests, the total error appeared to be lower than with different ability distribution. The linear equating methods (large h) gave higher errors than the equipercentile equating methods. Moreover, the IRT equating method had minimum or nearly minimum error in the external anchor test under all conditions. It provided a lower error rate for different group ability distributions with the internal anchor test for 30% and 40% as the ratio of anchor items. The lowest rate was recorded for the PSE-EQ equating method when group ability distribution was similar and the CE-EQ equating method when the ability distribution was different.

Findings for Standard Error of Equating

To solve this sub-question, a comparison of the equating methods was performed using the standard error of equating (SEE). For convenience, the equating results for conditions covered in the research were prepared separately for each of the internal and external anchor tests. Figures 3 and 4 display the graphs for standard error of equating under the conditions discussed.

Findings of the standard error of equating for internal anchor test

In the internal anchor test, the anchor items ratio was adjusted to 20%, 30%, and 40% for two different ability distributions. The results for standard error of equating obtained from both distributions are represented in graphs in Figure 3.

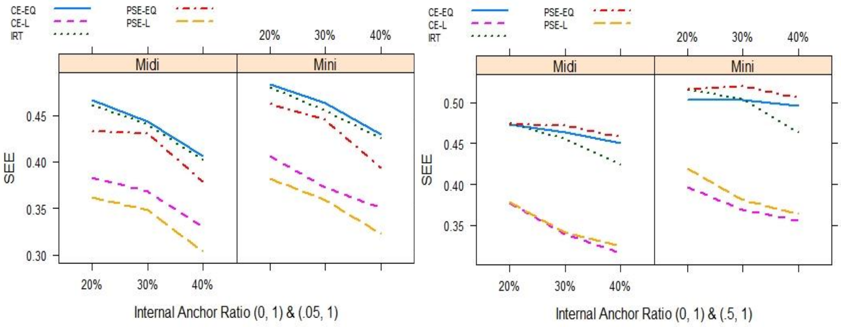

Figure 3. *SEE Values with Similar and Different Mean Group Ability Distribution for the Internal Anchor Test*

When the graphs above are examined, the standard error rate decreases as the ratio of anchor item increases for all of the equating methods. The decrease seems sharper, particularly for IRT equating when the ratio of anchor items increases from 30% to 40%. For all the equating methods, the standard errors obtained with similar group ability distribution appear to be lower than in the case of the different group ability distribution. At all ratios for anchor items and ability distributions, the KE linear equating method yielded a lower standard error rate than the KE equipercentile equating methods. Also, the difference between the standard error values of the linear and equipercentile equating methods was closer in the case of different ability distributions for the midi anchor test compared to the other conditions.

While the post-stratification linear equating method yielded lower standard errors for similar group ability distribution, the chained linear equating method proved a lower standard error value. It can be said that in all cases, the IRT equating method produced higher standard error than linear equating methods but this error was equal to or lower than for equipercentile equating methods. In the case where the distribution of ability between groups was similar, IRT equating standard error rates remained below those of CE-EQ equating but above the standard error rate for the PSE-EQ equating method. In the case of the different ability distribution, the two equipercentile equating methods yielded lower standard error rates for the midi anchor test condition; the rates were close to those of PSE-EQ with 20% and 30% ratio of anchor items for the mini anchor test while the error rate was lower with 40% ratio of anchor items.

Findings for standard error of equating for external anchor test

Figure 4 shows graphs for the standard error of equating for the external anchor test with the ratio of anchor items of 20%, 30%, and 40% when the distribution of group ability is similar and different in mini and midi anchor tests, respectively.

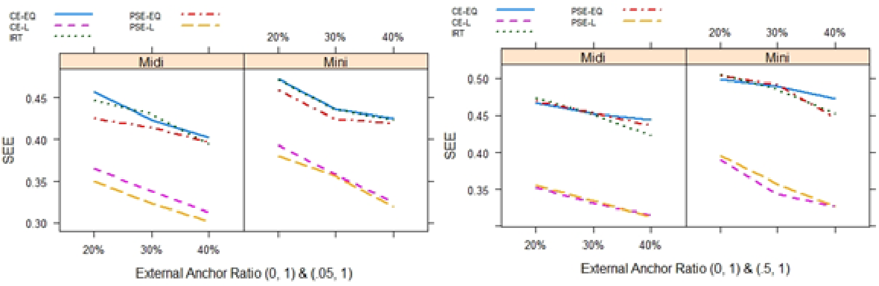

Figure 4. *SEE Values in Similar and Different Group Mean Ability Distributions for the External Anchor Test*

As seen in Figure 4, the equating error decreased as the ratio of the anchor items increased in the external anchor test, as for the internal anchor test, with all of the equating methods. This difference was more clearly seen for linear equating methods, especially. The standard error values increased when the distribution of ability was different in both anchor test conditions. In all conditions, the kernel linear equating methods gave a lower standard error rate than the kernel equipercentile equating methods. When group ability distribution was similar, the chained equating methods generated higher standard errors than the post-stratification equating methods. The standard errors for chained equating methods were smaller with different ability distributions. In the mini test condition where distribution of ability was different between the groups, the difference between standard error rates for the linear and equipercentile equating methods with 30% anchor items was larger than for the remaining conditions. For both models of ability distribution, a lower standard error was obtained for the midi test condition than for the mini test condition. When Figures 3 and 4 are compared, the standard error values obtained from the external anchor test remained below those obtained from the internal anchor test.

Discussion, Conclusion, and Recommendations

In this study, the kernel equating methods and IRT equating method results were compared under various conditions and the methods were investigated based on equating error. On the basis of equating methods, generally, KE CE and IRT performed better than KE PSE according to RMSD and KE linear equating methods yielded lower standard error rates than the KE equipercentile and IRT equating methods. In other words, the selection of parameter h is thought to have reduced the standard error values for KE. This result seems to be similar to the findings by Choi (2009) and Mao (2006). Ricker and von Davier (2007) pointed out that in cases where a linear criterion equating method is not employed for calculating RMSD (such as equipercentile equating), higher RMSD values occur since the linear equating methods are based on linear equating functions. In addition, the IRT equating resulted in a lower error rate than KE for outliers. The reason for this is that the Gaussian kernel method is used for the continuization of cumulative score distributions and it leads to higher standard errors for outliers. Therefore, IRT may be preferred over KE as there are likely to be outliers when considering actual test applications.

Both standard and total errors for the external anchor test were lower than for the internal anchor test. In other words, test length affects the error. In the case of an external anchor test, the tendency to increase the correlation between the total test and the anchor test may account for fewer errors because the total test contains more items. We obtained results that are similar to those of Kim (2014), yet contradictory with the findings from von Davier et al. (2006). This may be because the study by von Davier et al. (2006) was conducted with a set of real data. Moreover, Budescu (1985) stated that the standard error decreases as the ratio of anchor items increases. In the same direction, our results demonstrate that the error (standard and total) decreases as the ratio of anchor items increases for all of the equating methods. These findings seem to comply with other examples reporting decreased standard and total errors against increased ratios of anchor items (e.g. Hou, 2007; Kim, 2014; Meng, 2012; Sinharay & Holland, 2006b; Wang et al., 2008). In NEAT, as the ratio of anchor items increases, information derived from those items increases as well, which causes error rates to reduce as a result. Another finding obtained in the present study is that standard and total errors for the midi anchor test condition were less than for the mini anchor test since the midi responds correctly, which may result in a decrease in the equating error. This finding can be supported by Antal, Proctor, and Melican (2014), Fitzpatrick and Skorupski (2016), Kim (2014), Sinharay and Holland (2006b, 2007), and Sinharay, Haberman, Holland, and Lewis (2012). Thus, if the test is not too long, it is recommended to use external anchor tests with the midi anchor test.

The error increased when the distribution of ability was different between groups under the conditions considered for all of the equating methods. The previous researches found similar findings (e.g., Godfrey, 2007; Kim, 2014, Powers and Kolen, 2011; Sinharay and Holland, 2006a; Sinharay and Holland, 2007). Apart from that, the CE methods were affected more by different ability distributions between groups than the PSE methods. As the ability distribution difference between groups increased, the probability of contingency score distribution of test X on the type of anchor test A is identical to contingency score distribution of the test Y on the anchor test A, which leads to increased equating errors with PSE. Holland, von Davier, Sinharay, and Han (2006) reported that the CE-EQ and the PSE-EQ methods give better results with similar group ability distribution, but CE-EQ performs better when the distribution is unequal. As differences between the ability of groups are wide, the error rate for a high range of score scales was seen to increase as a result of all the equating methods. This finding is similar to the findings of Godfrey (2007), who noted deviations of equating methods from the criterion equating method (equipercentile equating) at outlier values. In this respect, differentiation of ability distribution between groups has a significant effect on the error rate, pushing it upwards. The reason may be that in the case of the similar group ability distribution, the group with a higher average responds to more items correctly, while the lower group behaves in the opposite manner. When the ability distribution is different, IRT and CE-PSE equating methods can be preferred.

In the present study, which compared the KE methods to the IRT equating method in different conditions, the findings showed that the KE methods provided results as satisfactory as IRT results under certain circumstances. Godfrey (2007), Meng (2012), and this study also used simulated IRT-model-based data and Norman Dvorak (2009) used a 'neutral' data generation approach. When the 'neutral' data generation approach was used, KE performed better than IRT; in other cases, in general, the results for IRT are advantageous. Because of that, the data generation approach night affect the results significantly. In future studies, KE and IRT can be compared with different conditions using different data generation approaches. Also, in this study, Gaussian kernel was used for bandwidth selection; in future research, logistic, uniform, and cross-validation approaches may be used to compare the results of equating methods. In addition, these equating methods can be compared for different test lengths and ability distributions.

In the light of the findings of both the present and previous studies, it is suggested that equated scores obtained from equated test forms differ based on equating method. When IRT assumptions are not met, KE can be used. In addition, among the kernel equating methods, equipercentile equating methods can be preferred to linear methods. Guided by the goal of the test to be applied, in testing situations should decide on the equating method to be used by taking the strengths and weaknesses of each method into account. It should also be recalled that the equating method to be selected may not necessarily yield better results than the others under all circumstances.

References

Andersson, B., & Davier, A. A. (2014). Improving the bandwidth selection in Kernel equating. *Journal of Educational Measurement, 51*(3), 223-238.

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 508–600). Washington, DC: American Council on Education.

Antal, J., Proctor, T. P., & Melican, G.C. (2014). The effect of type of common item construction on scale drift. *Applied Measurement in Education, 27*, 159–172.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the type of common item. *Journal of Educational Measurement, 22*, 13-20.

Chen, H. (2012). A comparison between linear IRT observed-score equating and Levine observed-score equating under the generalized Kernel equating framework. *Journal of Educational Measurement, 49*(3), 269-284.

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 thesis). University of Illinois at Urbana-Champaign., Illinois, USA. Available from ProQuest Dissertations and Theses database (UMI No. 3391908)

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates.

Fitzpatrick, J., & Skorupski, W. P. (2016). Equating with midi tests using IRT. *Journal of Educational Measurement, 53*(2), 172-189.

Godfrey, K. E. (2007). *A comparison of Kernel equating and IRT true score equating methods* (Unpublished doctoral thesis). The University of North Carolina, Greensboro, USA. Retrieved from https://libres.uncg.edu/ir/uncg/f/umi-uncg-1439.pdf

Hagge, S. L., & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1).* (CASMA Monograph Number 2.1) (pp. 95–135). Iowa City, IA: CASMA, The University of Iowa.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement, 12*, 38-47.

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

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*, 133-183.

Holland P. W., Dorans N. J., Petersen N. S. (2006). Equating test scores. In Rao C. R., Sinharay S. (Eds.), *Handbook of statistics* (Vol. 26, pp. 169-203). Oxford, UK: Elsevier.

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.

Hou, J. (2007). *Effectiveness of the hybrid Levine equipercentile and modified frequency estimation equating methods under the common-item nonequivalent groups design* (Unpublished doctoral thesis). University of Iowa, Iowa, USA. Retrieved from https://ir.uiowa.edu/etd/339/.

Fraenkel, J. R., Wallen, N. E. & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). New York: McGraw-Hill

Kim, H. Y. (2014). *A comparison of smoothing methods for the common item nonequivalent groups design.* (Unpublished doctoral thesis). University of Iowa. Retrieved from the University of Iowa at http://ir.uiowa.edu/etd/1344.

Kim, S. H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*(1), 25-41.

Kolen, M. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31– 55). New York: Springer.

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

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

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

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23-39.

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

Mao, X., von Davier, A. A., & Rupp, S. (2006). *Comparisons of the Kernel equating method with the traditional equating methods on PRAXIS™ data* (ETS Research Rep. No. RR-06-30). Princeton, NJ: ETS.

Meng, Y. (2012). *Comparison of Kernel equating and Item Response Theory equating methods* (Unpublished doctoral thesis). University of Massachusetts, Amherst, USA. Available from https://search.proquest.com/docview/1033227222

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

Partchev, I. (2016). Package 'irtoys'. (Version 0.2.0) [https://cran.r-project.org/web/packages/irtoys/irtoys.pdf, Acsess date: October 2016.]

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 221-262). New York: American Council on Education/Macmillan.

Powers, S. J., & Kolen, M. J. (2011). Evaluating equating accuracy and assumptions for groups that differ in performance. In M. J. Kolen, & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equatin*g (vol. 1) (CASMA Monograph Number 2.1). Iowa City: CASMA, The University of Iowa.

Powers, S., Hagge, S. L., Wei, W., He, Y., Liu, C., & Kolen, M. J. (2010). Effects of group differences on mixed-format equating. In M. J. Kolen & W. Lee (Eds.*), Mixed-format tests: Psychometric properties with a primary focus on equatin*g (vol. 1) (CASMA Monograph Number 2.1). Iowa City: CASMA, The University of Iowa.

Qu, Y. (2007). *The effect of weighting in Kernel equating using counter-balanced designs* (Unpublished doctoral thesis). Michigan State University, Michigan, USA.

R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. (http://www.R-project.org).

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

Rizopoulos, D. (2015). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software, 17*(5), 1-25.

Sinharay, S., & Holland, P. W. (2006a). *The correlation between the scores of a test and an type of common item* (ETS SR-06-04). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Holland, P. W. (2006b). *Choice of type of common item in equating* (ETS RR-06-35). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make type of common item s mini-forms of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44,* 249–275.

Sinharay, S., Haberman, S., Holland, P. W., & Lewis, C. (2012). *A note on the choice of an anchor test in equating* (ETS Research Report 12-14). Princeton, NJ: Educational Testing Service

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.

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 pseudo-tests constructed from real test data* (ETS RR-06–02). Princeton, NJ: Educational Testing Service.

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

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

Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1-33.

Zhu, W. (1998). Test equating: What, why, how? *Research quarterly for exercise and sport, 69*(1), 11-23.

Kernel ve Madde Tepki Kuramı Eşitleme Yöntemlerinin Karşılaştırılması

Atıf:

Akın Arıkan, Ç., Gelbal, S. (2021). Kernel ve madde tepki kuramı eşitleme yöntemlerinin karşılaştırılması, *Eurasian Journal of Educational Research*, 179-198, DOI: 10.14689/ejer.2021.93.9

Özet

Problem Durumu: Birçok test uygulamasında, özellikle geniş ölçekli ve yüksek riskli testlerde, test güvenliği ve bireylerin farklı günlerde test edilebilmesi için aynı testin farklı formlarının geliştirilmesi önemli bir gereklilik olarak uygulayıcıların karşısına çıkmaktadır. Ancak farklı formların geliştirilmesiyle, bu formlardaki maddeler güçlükleri açısından farklılaşabilmektedirler. Bu durum da kolay test formunu alan bireyin yüksek puan, zor testi alan bireyin daha düşük puan almasına neden olabilmektedir. Farklı formları alan bireyleri karşılaştırmaya duyulan gereksinimden dolayı, test güçlükleri arasındaki farkı ayarlamak için test formları eşitlenmektedir (Kolen ve Brennan, 2014; von Davier, Holland ve Thayer, 2004). Eşitleme, test puanlarının diğer test formundan elde edilen puanlarla ilişkilendirilmesine veya dönüştürülmesine denir (Hambleton ve Swaminathan, 1985). Bu araştırma kapsamında kernel eşitleme ve MTK eşitleme yöntemleri kullanılmıştır.

Kernel eşitleme, kesikli puan dağılımlarının sürekli dağılımlara dönüştürerek puan dağılımlarının eşitlendiği bir eşit yüzdelikli gözlenen puan eşitleme yöntemidir (von Davier vd, 2006). Eşit yüzdelikli eşitlemede, aynı yüzdelik sırasına denk gelen puanların eşit olduğu kabul edilir. Bunun için ilk olarak her bir formun yığılmalı frekansı hesaplanarak tablolaştırılır ve bu yığılmalı frekanslara göre aynı yüzdelik puanlara karşılık gelen puanlar eşitlenir. Eşit yüzdelikli eşitlemede, aynı yüzdelik sırasındaki eşitlenmiş puanlara sahip bireylerin aynı yetenek düzeyinde olduğu kabul edilir (Kolen, 1988). Ancak gerçek uygulamalarda bu bireylerin aynı yetenek düzeyinde olması oldukça güçtür. Kernel eşitleme, eşit yüzdelikli eşitlemede ortaya çıkan bu problemi çözmek için geliştirilmiştir. Bu problemin ortaya çıkmasının nedeni ise puan dağılımlarının kesikli olmasıdır. Holland ve Thayer (1981) bu probleme, kesikli dağılımları kernel eşitleme ile süreklileştirerek çözüm getirmiştir. Kernel eşitlemede, kesikli dağılımlar sürekli hale getirilerek, sürekli dağılımlar üzerinden puanlar eşitlenir (Livingston, 1993; Ricker ve von Davier, 2007).

Madde Tepki Kuramı (MTK) bireylerin yeteneklerini matematiksel modellerle açıklar. Lord (1953), gerçek ve gözlenen puanın yetenek puanıyla aynı anlama gelmediğini, yetenek puanının testten bağımsız iken, gerçek ve gözlenen puanının teste bağımlı olduğunu belirtmiştir (akt: Hambleton ve Jones, 1993). MTK gerçek puan eşitleme üç aşamadan oluşur. Geleneksel eşitleme metotlarında olduğu gibi ilk aşama veri toplama deseninin seçilmesidir, ikinci aşama uygun MTK modeline karar verilmesi ve son aşama ise uygun model ile kestirilen madde parametrelerinin ortak ölçeğe

yerleştirilmesidir (Cook ve Eignor, 1991; Zhu, 1998). MTK gerçek puan eşitlemede, eğer toplam puan kullanılacaksa, puanlar aynı ölçeğe yerleştirildikten sonra toplam puanların eşdeğerleri elde edilir. Parametre kestirimleri aynı ölçek üzerinde ise, MTK gerçek puan ve MTK gözlenen puan eşitleme metotları X formuna ait toplam puanlar ile Y formuna ait toplam puanları ilişkilendirmek için kullanılır (Kolen, 2007). MTK gerçek puan eşitlemede, bir forma ait belirli bir θ değeriyle ilişkilendirilen gerçek puan ile diğer formdaki aynı θ değeriyle ilişkilendirilen gerçek puan arasında ilişki olduğunu varsayar ve bu ilişkiyi eşitleme için kullanır.

Araştırmanın Amacı: Eşitleme yöntemlerinin güçlü ve zayıf yönlerinin bilinmesi, test programlarının gereksinimine göre uygun eşitleme yönteminin seçimini kolaylaştırır. Ayrıca geniş ölçekli ve yüksek riskli testlerde elde edilen eşitleme sonuçlarına göre bireyler hakkında önemli kararlar verildiği için bir yöntemin diğer yönteme göre tercih edilmesinin verilecek kararı nasıl etkilediğinin bilinmesi önemlidir (Kim ve Cohen, 2002). Kernel eşitlemenin simülasyon çalışmaları ile uygulamada sıklıkla kullanılan MTK gerçek puan eşitleme yöntemi karşılaştırılarak, kernel eşitlemenin avantaj ve dezavantajlarının ortaya çıkarılması ile hangi durumlarda kullanılmasının daha uygun olduğunun belirlenmesinin önemli olduğu düşünülmektedir. Bu amaç doğrultusunda, aşağıdaki probleme cevap aranmıştır.

"Testler, kernel son tabakalama doğrusal, kernel zincirleme doğrusal, kernel son tabakalama eşit yüzdelikli, kernel zincirleme eşit yüzdelikli ve MTK gerçek puan eşitleme yöntemlerine göre eşitlendiğinde yetenek dağılımı, ortak madde tipi, ortak madde oranı ve ortak madde güçlük dağılımına göre eşitlemenin hatası nasıl değişmektedir?

Araştırmanın Yöntemi: Araştırmada, beş farklı eşitleme yönteminin performansı; yetenek dağılımı (2 koşul), ortak madde tipi (2 koşul), ortak madde oranı (3 koşul) ve ortak madde güçlük dağılımı (2 koşul) olmak üzere toplam 24 koşulda incelenmiştir. Bu koşullar altında kernel son tabakalama eşit yüzdelikli (ideal h), kernel son tabakalama doğrusal (geniş h), Kernel zircirleme eşit yüzdelikli (ideal h), kernel zincirleme doğrusal (geniş h) ve MTK gerçek puan (Haebara) eşitleme yöntemleri karşılaştırılmıştır. Hem kernel hem de MTK gerçek puan eşitleme yöntemlerinde kullanılan test formlarındaki maddeler için verilerin türetilmesi aşamasında belirtilen değerlerle madde ve birey parametre değerleri simüle edildikten sonra R programında "irtoys" paketi (Partchev, 2016) kullanılarak MTK 3 PL modele uyumlu iki kategorili (1-0) cevaplar türetilmiştir. Kernel eşitleme yöntemleri ile test formlarının eşitlenmesi için "kequate" paketi (Andersson, Branberg & Wiberg, 2013) kullanılmıştır. MTK gerçek puan eşitleme için test formlarının cevaplarına ait 3 PL modele uygun olarak madde ve yetenek parametre kestirimleri ise R programında "ltm" paketi (Rizopoulos, 2015) ile yapılmıştır. Madde parametre kestirimleri için Marjinal En Çok Olabilirlik, yetenek parametrelerinin kestirimi için ise Beklenen Sonsal Dağılım yöntemleri kullanılılarak kestirimler yapılmıştır (Rizopoulos, 2015). Daha sonra kestirilen parametreler, MTK gerçek puan eşitleme için ayrı kalibrasyon yöntemlerinden Haebara yöntemiyle aynı ölçeğe yerleştirilerek gerçek puan eşitleme yapılmıştır. Madde parametrelerine ait kalibrasyonlar ve gerçek puan eşitleme için "plink" paketi (Weeks, 2010) kullanılmıştır. Eşitleme sonuçlarının doğruluğunu değerlendirmek için

RMSD (Root mean square difference) ve eşitlemenin standart hatası (SEE) kullanılmıştır.

Araştırmanın Bulguları: Bu çalışmada kernel eşitleme yöntemleri ile MTK eşitleme yöntemi gruplar arası yetenek dağılımı, ortak madde oranı ve ortak madde güçlük dağılımı değişkenleri açısından karşılaştırılmıştır. Bütün eşitleme yöntemlerinde ele alınan koşullara göre gruplar arası yetenek dağılımı farklı olduğunda standart hata ve toplam hatanın arttığı görülmektedir. Bir başka deyişle, bütün eşitleme yöntemleri daha düşük performans göstermiştir. Çalışmadan elde edilen bulgulara göre dış ortak testte standart ve toplam hatanın iç ortak teste göre daha az olduğu bulunmuştur. Dış ortak testte, toplam test daha fazla maddeye sahip olduğundan, toplam test ile ortak test arasındaki korelasyonun artma eğiliminde olması daha az hatanın elde edilmesine neden olmuş olabilir. Çalışmadan elde edilen diğer bulgu da midi ortak test koşulunda standart ve toplam hatanın mini ortak test koşuluna göre daha az olmasıdır. Eşitleme yöntemleri bazında; doğrusal eşitleme yöntemlerinin, eşit yüzdelikli eşitleme ve MTK eşitleme yöntemlerine göre daha düşük standart hata verdiği bulgusuna ulaşılmıştır. Yani, h parametre seçimi standart hata değerlerinin azalmasına neden olmuştur.

Araştırmanın Sonuçları ve Öneriler: Bu çalışmada, kernel eşitleme yöntemlerinin bazı koşullarda MTK eşitleme kadar iyi sonuçlar verdiği bulunmuştur. Bu çalışmadan elde edilen bulgular ve önceki çalışmalar ışığında, test formlarının eşitlenmesi sonucunda elde edilen eşitlenmiş puanlar eşitleme yöntemlerine göre farklılık göstermektedir. Testin amacı doğrultusunda, eşitleme yöntemlerinin güçlü ve zayıf yönleri dikkate alınarak, eşitleme yöntemine karar verilmelidir.

Anahtar Kelimeler: Eşitleme, kernel, MTK, hata