International Journal of Assessment Tools in Education
2021, Vol. 8, No. 2, 222–238
https://doi.org/10.21449/ijate.815961
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

# Automated Essay Scoring Effect on Test Equating Errors in Mixed-format Test

**Ibrahim Uysal** [1,*], **Nuri Dogan** [2]

[1]Department of Educational Sciences, Faculty of Education, Bolu Abant İzzet Baysal University, Bolu, Turkey
[2]Department of Educational Sciences, Faculty of Education, Hacettepe University, Ankara, Turkey

**ARTICLE HISTORY**
*Received: Oct. 24, 2020*
*Revised: Dec. 31, 2020*
*Accepted: Feb. 07, 2021*

**Keywords:**
Test equating,
Automated scoring,
Classical test theory,
Item response theory,
Mixed-format tests.

**Abstract:** Scoring constructed-response items can be highly difficult, time-consuming, and costly in practice. Improvements in computer technology have enabled automated scoring of constructed-response items. However, the application of automated scoring without an investigation of test equating can lead to serious problems. The goal of this study was to score the constructed-response items in mixed-format tests automatically with different test/training data rates and to investigate the indirect effect of these scores on test equating compared with human raters. Bidirectional long-short term memory (BLSTM) was selected as the automated scoring method for the best performance. During the test equating process, methods based on classical test theory and item response theory were utilized. In most of the equating methods, errors of the equating resulting from automated scoring were close to the errors occurring in equating processes conducted by human raters. It was concluded that automated scoring can be applied because it is convenient in terms of equating.

## 1. INTRODUCTION

Test developers often have a dilemma in choosing the item format to be included on the tests. Reasons for this include suitability for the measurement of cognitive features, cost of application and scoring, the effect of item types used in tests on teaching, and psychometric properties. With practicality in mind, tests can be designed to include only multiple-choice items, only constructed-response items, or both multiple-choice and constructed-response items (Martinez, 1999; Rodriguez, 2002). Martinez (1999) states that a single-format test is not suitable for all purposes and situations, while Messick (1993) states that using different test item formats together will benefit from the strengths of each format and compensate for weaknesses. Therefore, it is essential to use both multiple-choice and constructed-response items, especially in large-scale tests. Because with constructed-response items, students have opportunity to organize and apply what they learn in a deeper way (Tankersley, 2007). However, it is difficult, time-consuming, and costly to score constructed-response items in large-scale testing applications. Due to the scoring difficulties of constructed-response items, test developers searched for and introduced the concept of automated scoring (Page, 1966).

CONTACT: İbrahim UYSAL ✉ ibrahimuysal06@gmail.com Department of Educational Sciences, Faculty of Education, Bolu Abant İzzet Baysal University, Bolu, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

Using automated essay scoring systems in tests will ensure efficient use of funds, reduce scoring time, and efforts (Attali & Burstein, 2006; Chen et al., 2014). The use of this system will eliminate the necessity to use many raters. Besides, scoring bias can be prevented. Reliability problems arising from differently trained raters will be overcome, as will generalizability (Adesiji et al., 2016). However, the effectiveness of automated scoring systems in applications such as test equating, which is important in ensuring justice between individuals taking different test forms or participating in the test at different times, has not been adequately investigated in the literature. Applying automated scoring without such research can cause serious problems (such as making wrong decisions about individuals). When automated scoring conditions change, equating error is also likely to change. In this respect, it is necessary to determine the acceptable automated scoring limits for test equating. The current study was designed based on these problem situations.

This study is important in determining whether automated scoring and training/test data rates in automated scoring increase test equating errors and whether the equating errors that occur because of automated scoring are different from the equating errors that occur with human raters. Thus, test equating after automated scoring can be performed under relevant conditions. When the literature was examined, a test equating study that Almond (2014) conducted on constructed-response items by automatically scoring common items in a sample of 500 people was found. In this study, the linear logistic equating method, a variant of Tucker linear equating, was used. Also, there was only one test equating study using automated scoring in mixed-format tests. This study, conducted by Olgar (2015), contains 30 multiple-choice items and one open-ended item in tests. The studies carried out by Almond (2014) and Olgar (2015) used the linear logistics equating method. The current study focused on equating tests with a large number of constructed-response items with automated scoring.

Moreover, this study was not based on a single test equating method but was carried out using both classical test theory (CTT) and item response theory (IRT) based test equating methods. It was seen that test equating methods based on IRT were not used in test equating studies carried out with automated scoring. So, to investigate which method works better in equating with automated scoring, both CTT and IRT were used in the study.

In the literature, similar studies compared the equating methods based on CTT and IRT in mixed-format tests and between nonequivalent groups using a common item pattern (Hagge & Kolen, 2011; Hagge et al., 2011; He, 2011; Lee et al., 2012; Liu & Kolen, 2011; Wolf, 2013). In the current study, CTT-based equating methods (Tucker linear, chained linear, chained equipercentile, frequency equipercentile), and IRT-based true score equating methods (mean-mean, mean-sigma, Stocking-Lord and Haebara) were used. Most of the literature studies (Hagge & Kolen, 2011; Hagge et al., 2011; He, 2011; Liu & Kolen, 2011; Wolf, 2013) compared CTT-based chained equipercentile and frequency estimation methods and IRT-based true and observed score equating methods. Among these studies, Hagge and Kolen (2011) and Hagge et al. (2011) used the Haebara method, Wolf (2013) used simultaneous scaling and He (2011) and Liu and Kolen (2011) used the Stocking-Lord method in IRT-based true score equating. In their research, Lee et al. (2012) compared Tucker, Levine observed score, Levine true score, chained equipercentile, frequency estimation, Stocking-Lord, and IRT observed score equating methods.

In the current study, in cases where equipercentile equating, based on CTT, was used, pre-smoothing with the bivariate log-linear function was applied. Similar to this study, Hagge et al. (2011), Lee et al. (2012), and Wolf (2013) pre-smoothed with the log-linear function. On the other hand, Liu and Kolen (2011) used pre-smoothing while obtaining the results for the population to make a comparison in the equating process. In addition, they changed synthetic population ratios of equating methods other than chained equating methods. Similarly, Hagge

and Kolen (2011), Hagge et al. (2011), and Wolf (2013) changed the synthetic population ratio to 1 in their study. However, these studies did not evaluate the effect of the synthetic population ratio but showed the results based on the new group that took the test. While Hagge and Kolen (2011) and Liu and Kolen (2011) conducted their research on real data, Wolf (2013) worked on simulated data. Of these researchers, Liu and Kolen (2011) included only multiple-choice items in tests as common items, while Hagge and Kolen (2011) and Wolf (2013) used mixed-format tests as common items in tests.

More constructed-response items should be included in large-scale tests to measuring more complex skills such as higher-order, critical thinking and reasoning, better evaluating items involving multiple steps in the solution process. But these items should also be easily and accurately scored. Therefore, the current study is important. In addition, test equating studies on restricted constructed-response items with automated scores are not enough. This study has two purposes: i) to evaluate the effect of constructed-response items scored by automated scoring systems in the test equating process on equating errors, ii) to examine the change of equating errors in the change of the conditions in the automated scoring systems.

## 2. METHOD

### 2.1. Design

The study was correlational, as it aims to determine the effect of automated scoring of constructed-response items on test equating in mixed-format tests by comparing it with test equating performed by human raters. Creswell (2012) stated that it is possible to see how a difference in one variable affects the other variable in correlational studies.

### 2.2. Sample

The data for this study were obtained from the eighth-grade Turkish test that is part of the Academic Skills Monitoring and Evaluation (ABİDE) project implemented by the Ministry of National Education (MoNE) in 2016. Data for 1000 students who answered the $A_1$ and $B_1$ booklets on the Turkish test were selected randomly. After selecting and cleaning data, 607 students from the $A_1$ booklet and 584 students from $B_1$ booklet were studied. Details were given in the data analysis section. Spence (1996) stated that at least 500 individuals must answer each test form for test equating studies. The number of students answering the $A_1$ and $B_1$ booklets in this research met this criterion.

### 2.3. Data Collection Tools

Multiple-choice and constructed-response items are included in ABİDE tests, which aim to examine students' higher-order thinking skills using different types of items. Two human rater groups scored Constructed-response items, and a third rater group was consulted in case of a dispute between the first two raters' groups. The focus of the research was the data obtained from two Turkish test forms ($A_1$ and $B_1$) with 18 items. 9 items in the $A_1$ test and 10 items in the $B_1$ test were constructed-response items. Constructed-response items were scored as either 0-1 or 0-1-2. Nine items were common in $A_1$ and $B_1$ tests (MoNE, 2017).

Since the tests used in the study contain common items, they were equated using the common-item nonequivalent group (CINEG) design. However, some criteria must be met to equate the tests using a CINEG design. Angoff (1984) stated that even if the test length increases, the proportion of common items in the test should not be less than 20%. In this application, the proportion of common items was 50%. Considering the data characteristics, it is necessary to use dichotomously and polytomously scored item types together in common items in tests. As a matter of fact, Tate (2000) proposed the use of both types of items as common items in mixed-format tests. The reason for this is that the common items should represent the entire test. In the $A_1$ and $B_1$ booklets, five of the nine common items were constructed-response and four were multiple-choice.

Cramer's V coefficient calculated the consistency between raters for each constructed-response item included in the tests in the ABİDE study. Cramer's V ranged from .83 to .98 for items included in the Turkish test in $A_1$ booklet, and from .87 to .99 for items included in the Turkish test in $B_1$ booklet. Internal consistency coefficients for test scores were stated as .73 for booklet A and .76 for booklet B (MoNE, 2017).

### 2.4. Data Analysis

The data were entered based on the balanced distribution of the categories regarding the scores obtained from the constructed-response items. This was done to avoid the problem of prevalence regarding constructed-response items in the data. Indeed, this is important in automated scoring. Taking into account 9 items for $A_1$ booklet and 10 items for $B_1$ booklet 697 data entries from $A_1$ booklet and 701 data entries from booklet $B_1$ were made. Then, within the researchers' criteria, students responding to half or more of the constructed-response items and multiple-choice items in the test were selected. After this process, the missing data rates were calculated for each constructed-response and multiple-choice item. The data were cleaned so that the missing data rate remained below 5%. It was anticipated that a large number of blank answers will show higher interrater reliability coefficients in automated scoring. As there were few data in some categories, individuals scoring in these categories were retained in the response data as much as possible. Then, the scores given by the two groups of human raters (group 1 and group 2) were examined. Due to the missing data, a group of students were also excluded from the study. In the last case, 90 students using the $A_1$ booklet and 117 students using the $B_1$ booklet were excluded. Thus, the data preparation process was completed, and the automated scoring process was started with 607 data from the $A_1$ booklet and 584 from the $B_1$ booklet.

In the study, an automated scoring system was created using the Python program on the Linux operating system. Automated scoring was done using supervised machine learning algorithms by mapping the computer's scoring features through human raters. Five methods were used in automated scoring: SVM (support vector machine), LR (logistic regression), MNB (multinominal naive Bayes), LSTM (long-short term memory), and BLSTM (bidirectional long-short term memory). Two libraries were used in the software prepared through Python. 90% of the data was used to train the system and 10% to test the system. Random sampling method was applied with cross validity. Ten-fold cross-validation was used. Turkish test constructed-response items belonging to "Monitoring, Research and Development Project for Measurement and Evaluation Applications" implemented by MoNE were used while developing the software. This test is different from the ABİDE tests used in this research. It is given to fifth-grade students (10–11 years old) and includes five constructed-response items. Five constructed-response items were used while preparing the software. Three of the five constructed-response items are scored as 0-1, while two are scored as 0-1-2. Two human rater groups scored each student's answer, and a third rater group was applied in case of dispute. Rubrics were used in scoring processes. Table 1 shows the sample results of 0-1 scored item 16 and 0-1-2 scored item 20. While 0-1 scored item 16 was tested with 303 data, 0-1-2 scored item 20 was tested with 637 data. Since item 20 was scored in three categories, it was found appropriate to experiment on more data.

**Table 1.** *Agreement percentages between automated and human scoring.*

| | Number of data | Number of cathegory | SVM (%) | LR (%) | MNB (%) | LSTM (%) | BLSTM (%) |
|---|---|---|---|---|---|---|---|
| Item 16 | 303 | 2 | 98.0 | 98.3 | 96.1 | 99.0 | 99.0 |
| Item 20 | 637 | 3 | 85.5 | 82.4 | 75.1 | 87.3 | 88.7 |

Note: Agreement percentages above 80% indicate an acceptable fit (Hartmann, 1977).

Table 1 shows that the percentages of agreement obtained for item 16 were relatively high. The methods that showed the highest agreement percentage for this item were LSTM and BLSTM. Therefore, the agreement percentages obtained for item 20 are sufficient. The method that showed the best agreement in item 20 was the BLSTM method. The fact that the percentages of agreement obtained for all methods were at the expected level showed that the system created would be sufficient to score the current study's constructed-response items.

The entry of the student answer sheets in JPEG format for constructed-response items was done manually. This is because students' handwriting was difficult to read and because optical character recognition (OCR) systems cannot be used on account of the use of adjacent handwriting. In addition, it was to eliminate errors that may arise from OCR programs. In order to completely match the manually entered data with student answers, the data were checked by a team of six people and errors were corrected. Student answers were directly conveyed and were not subject to any correction.

The automated scoring system was trained in the automated scoring phase using the human raters' final scores. In this way, it was taught how to score by human raters and the scoring features were mapped to the system. Test data, which were not used in the training of the system, were scored automatically. The amount of data used to test the system wass a factor studied in the research. The data rates used to test the system were determined as 10%, 20% and 33%. Therefore, the amount of data used in training the system was 90%, 80% and 67% respectively. These values indicated that 61, 121 and 200 of the 607 data for the $A_1$ booklet were used to test the system, respectively, while 546, 486 and 407 data, respectively, were used to train the system. From the $B_1$ booklet, 584 data, 58, 117 and 193 are used to test the system, respectively; 526, 467 and 391, respectively, were used to train the system. The amount of data to be used for training the system was reduced as much as possible, and the effect of this on automated scoring and indirect effect on test equating examined. While calculating the results, 10-fold cross-validation was used for the 10% test data rate, 5-fold cross validity was used for the 20% test data rate, and 3-fold cross validity was used for the 33% test data rate. In this way, training and test data were differentiated and all data from both booklets were converted into test data. As a result, the system obtained 607 data scored for the $A_1$ booklet and 584 scored for the $B_1$ booklet.

Automated scoring was performed for 10%, 20% and 33% test data rates using the BLSTM method, which shows the best fit, and equating was started. In order to make comparisons, the test forms were equated by using the final scores of the human raters for each test form. In the equating process, methods based on CTT and IRT were used. The test data's statistics and reliability values to this research were examined before the equating process. The statistics and reliability coefficients of the $A_1$ and $B_1$ booklet for human raters and automated scoring (BLSTM 10%, BLSTM 20% and BLSTM 33%) are given in Table 2. The reliability coefficient was examined in two ways. In the first case, reliability was determined by Cronbach's alpha coefficient (Cronbach, 1951) and in the second case by McDonald's omega coefficient (McDonald, 1999) based on factor analysis. While the alpha coefficient was used because it gave the lower bound estimate of reliability, the omega coefficient was chosen because it had less and more realistic assumptions (Bendermacher, 2010; Dunn et al., 2014).

Table 2 shows that the average score generated by human rating was slightly lower than the average score calculated after automated scoring. When using human raters, the standard deviation was slightly higher than automated scoring. Omega and Cronbach's alpha reliability coefficients were found to be close to each other under both human rating and automated scoring. However, when using human raters, both Cronbach's alpha and omega coefficients were slightly higher.

**Table 2.** *Test statistics on $A_1$ and $B_1$ booklets.*

| | Human Raters | | BLSTM %10 | | BLSTM %20 | | BLSTM %33 | |
|---|---|---|---|---|---|---|---|---|
| | $A_1$ | $B_1$ | $A_1$ | $B_1$ | $A_1$ | $B_1$ | $A_1$ | $B_1$ |
| Number of Item | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| Sample Size | 607 | 584 | 607 | 584 | 607 | 584 | 607 | 584 |
| Mean | 13.152 | 14.101 | 13.259 | 14.300 | 13.283 | 14.361 | 13.273 | 14.346 |
| Standart Deviation | 4.530 | 4.964 | 4.331 | 4.777 | 4.333 | 4.765 | 4.313 | 4.760 |
| Median | 14 | 15 | 13 | 15 | 14 | 15 | 14 | 15 |
| Minimum | 1 | 0 | 2 | 0 | 2 | 0 | 2 | 1 |
| Maximum | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Skewness | -.249 | -.466 | -.208 | -.520 | -.218 | -.538 | -.209 | -.518 |
| Reliability (Alfa) | .766 | .797 | .746 | .784 | .746 | .783 | .747 | .786 |
| Reliability (Omega) | .868 | .893 | .857 | .885 | .856 | .882 | .858 | .884 |

Chained linear (LC), Tucker linear (LT), chained equipercentile (EC), and frequency estimation (EF) equating methods based on CTT were chosen. Synthetic population value was changed to $w_1 = 1$ (WS = 1) and the effect of this situation was investigated. When the synthetic population was determined as $w_1 = 1$, the group that takes the new test form in the common item design in nonequivalent groups was determined as the synthetic universe (Kolen, & Brennan, 2014). When the synthetic population value was not changed, the synthetic population was determined according to the number of samples in the groups (to be $w_1 + w_2 = 1$). However, since chained equating did not support the synthetic population, synthetic population ratios had not been changed in methods using chained equating (Kolen, & Brennan, 2014). In addition, presmoothing (PSM) was performed for equipercentile equating methods. For the EF method, PSM is performed and the synthetic population ratio was changed. With these changes, the effects of synthetic population parameters and/or PSM on the equating results were also evaluated. "equate" (Albano, 2016) package in R (R Development Core Team, 2018) was used while equating test forms according to CTT methods. PSM was carried out using PROC IML (Moses & von Davier, 2006) code in SAS 9.4 (SAS Institute, 2015). The reason for performing this procedure outside the R program was that the total scores obtained from the $A_1$ booklet or the $B_1$ booklet and the total scores obtained from the common tests should be subtracted because some of the frequencies associated with the score combinations were zero (Moses et al., 2004). However, the "equate" package in the R software did not allow this.

PSM was performed using polynominal bivariate loglinear function distribution due to the use of nonequivalent group design. The best model was chosen for each form by comparing 11 different models in the polynominal bivariate loglinear function distribution. The equating was carried out by using 10000 replications with the bootstrap technique.

The mean-mean (MM), mean-sigma (MS), Haebara (HB) and Stocking-Lord (SL), which are true score equating methods based on separate calibration in IRT, were used. Before equating, IRT assumptions were examined. The first assumption was unidimensionality. Factor analysis for mixed tests for each test form was carried out for both human scorers and automated scoring conditions using the MPLUS (Muthén & Muthén, 2012) program. Due to the use of mixed-

format tests, polycoric and tetracoric correlations were utilized. The weighted least square mean and variance adjusted (WLSMV) were used as the estimation method in the factor analysis. WLSMV estimation method is known as one of the most suitable methods when using polycoric and tetracoric correlations (Barendse et al., 2015). In addition, parallel analysis (Timmerman & Lorenzo-Seva, 2011) was carried out through the Factor 10.5 program (Lorenzo-Seva & Fernando, 2006) in order to decide the number of dimensions. Parallel analysis results showed that each test form has a single factor structure for both automated scoring (with 10%, 20% and 33% test data rates) and human raters.

Five models were compared to determine which IRT model fit the data for each test form. Since there were re constructed-response items rated binary and there was no possibility to respond to these items by chance, all binary items were examined based on one parameter model (1PLM) and two-parameter model (2PLM). Models reviewed include 1) 1PLM and partial credit model (PCM), 2) 1PLM and generalized partial credit model (GPCM), 3) 1PLM and graded response model (GRM), 4) 2PLM and GPCM, 5) 2PLM and GRM. When comparing models, the differences between -2log likelihood values and degrees of freedom were calculated, and these values were compared with the chi-square table. If the value obtained was greater than the value determined for the 5% error in the chi-square table, a higher model had been adopted. When comparing models with the same degrees of freedom, standard error averages related to theta estimation were used. EAP method was used to estimate ability parameters. Accordingly, models with lower standard errors were used to estimate the ability and item parameters. Model comparisons were made for all of the human raters' final scores and the rating done by the automated scoring systems and it was concluded that the 2PLM and GPCM methods were more appropriate overall. Ability and item parameters were estimated using XCalibre 4.1 (Yoes, 1996). The XCalibre program estimates the discrimination and difficulty parameters with a lower error (RMSE) than BILOG (Mislevy & Bock, 1997; Weiss & Minden, 2012). Test equating was performed by transferring the ability parameters and item parameters estimated in the XCalibre program to the IRTEQ program.

Standard error of equating (SEE), bias (BIAS), and root mean squared error (RMSE) were calculated to be used in comparisons after test equating with methods based on CTT and IRT. The random error (SEE) was designed based on the standard deviation of the equated scores and results from the sample. Bias, that is, systematic error, was based on the difference between the estimated equation and the criterion (real) equation relationship. Bias results from reasons such as the common items do not represent the test form in terms of content and statistical properties in nonequivalent groups, the serious differences between the groups and the difference of common items from one application to another. Bias was not a coefficient directly affected by the sample. RMSE is a combination of bias and standard error (Kolen & Brennan, 2014; LaFlair et al., 2017). The bias value was not directly used in comparing the performance of the methods due to the high level of negative and positive values can neutralize each other (Zu & Liu, 2010). Absolute BIAS values have not been studied since the negative BIAS value indicates that the skills are predicted to be lower than they are and the positive indicates that the skills are predicted higher than they are (Pang et al., 2010). The methods were compared over SEE and RMSE, which is a combination of SEE and BIAS. While choosing the best method, RMSE values were used due to the combination of systematic and random error.

SEE, BIAS, and RMSE values were calculated through the “equate” package (Albano, 2016) after the equating process in CTT and the MSEXCEL module after the IRT equating process. By choosing the same error coefficients, CTT and IRT equating methods were compared. To make it easier to compare with the CTT, theta was used to calculate the IRT errors. Below are the equations used to calculate BIAS (equation 1), RMSE (equation 2) and SEE (equation 3) in the CTT (Gonzalez & Wiberg, 2017). L is the number of bootstraps performed, l are the

samples, $\hat{\varphi}(x_i)$ is the estimated equated scores, $\varphi(x_i)$ is the real equated scores, and $\bar{\hat{\varphi}}(x_i)$ is the estimated equated mean scores:

$$BIAS(x_i) = \frac{1}{L}\sum_{l=1}^{L}[(\hat{\varphi}_1(x_i) - \varphi_1(x_i)] \quad (1)$$

$$RMSE(x_i) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}[(\hat{\varphi}_1(x_i) - \varphi_1(x_i)]^2} \quad (2)$$

$$SEE(x_i) = \sqrt{RMSE(x_i)^2 - BIAS(x_i)^2} \quad (3)$$

The following equations can be used when calculating SEE (equality 4), BIAS (equality 5) and RMSE (equality 6) values based on IRT. The resources of Deng and Monfils (2017) and Keller and Keller (2011) were used for equations. $\Theta_i$ is the ability of the individual *i*, $\widehat{\Theta}_i$ is the ability of the individual i estimated by the equating method used, and N is the sample size:

$$SEE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{\Theta}_i - \Theta_i - BIAS)^2} \quad (4)$$

$$BIAS = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\Theta}_i - \Theta_i) \quad (5)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{\Theta}_i - \Theta_i)^2} \quad (6)$$

After the equating errors were obtained for three automated scoring conditions, they were compared with the human raters. It was then decided to perform a difference test to determine the status of showing significant difference in the errors (RMSE) of the rater type in the equating process. Accordingly, the average of three conditions related to automated scoring was calculated. Normality was then tested for each group. A Shapiro-Wilks test was used while testing normality. The results showed that the RMSE values of the equating process performed through human raters did not distributed normally ($W(sd = 13) = .860$, $p < .05$), and the RMSE values of the equating process performed through automated scoring system were normally distributed ($W(sd = 13) = .914$, $p = .210 > .05$). As a result, since one of the groups did not provide the assumption of normality, the difference test was carried out with the Mann-Whitney U test, a nonparametric technique. To determine the effect of the scoring type on the RMSE, the effect size was calculated through Cliff's Delta coefficient (Cliff, 1996). The Cliff's Delta coefficient used to compare two groups ranges from -1 to +1. If the coefficient is closer to -1 or +1 the effect size is increased and if closer to 0 effect size is decreased (Cliff, 1993). For this purpose, R "effsize" package (Torchiano, 2020) was used.

After calculating the effect size, the correlation between the errors of the human raters' equating and the errors of the automated scoring equating were examined. According to the normality tests, the relationship was examined using Spearman's rho correlation since one of the variables did not meet the normality assumption.

## 3. RESULT / FINDINGS

Table 3 shows the errors related to the test equating process. Equating was made with human scores for both forms and equating errors displayed in the "human" column. Equating using machine scores was performed for both forms and equating errors are shown in the "BLSTM" column. Table 3 shows the equating errors using the scores obtained with 10%, 20%, 33% test data rates via the BLSTM method. In Table 3, the lowest error methods are shown in bold and the highest error are shown in italics for each rater and type of error condition.

When the human raters were re taken into consideration in Table 3, that the lowest random error (SEE) was .050 obtained in the MS method based on IRT. MM method followed this with .061. When using methods based on IRT, the highest SEE (.083) showed in SL and HB methods. When using human raters, the method that showed the lowest SEE (.197) in CTT based equating methods was the Tucker linear in which the synthetic population ratio was determined as 1 (LT[WS = 1]). This value was followed by (.198) the LT equating method in which the random universe ratio was not changed and the random universe ratio was determined based on the sample numbers. The method with the highest SEE (.357) was the PSMEC equating method, which was pre-smoothed with a bivariate logarithmic linear function. In the case where human raters were used, the highest SEEs were obtained in equipercentile equating methods. In this condition, methods based on IRT generally showed lower SEEs than methods based on CTT.

When test equating results made after automated scoring performed with a 10% test data rate and the BLSTM method were evaluated in terms of random error, the lowest random error (.047) was found in MS method. This value (.047) was lower than that of human raters (.050). This value (.047), which was obtained at the 10% test data rate, was followed by the MM method with .079. When using methods based on IRT, HB method showed the highest SEE (.110). When automated scoring was performed at a rate of 10% test data, LT[WS = 1] was the method that shows the lowest SEE (.200) in test equating methods based on CTT. This value was followed by the LT equating method with .201. The method with the highest SEE (.407) is the EC. In the equating performed after automated scoring with the 10% test data rate and BLSTM method, the highest SEEs were obtained in equipercentile equating methods. In this condition, methods based on IRT generally showed less SEEs than methods based on CTT. The SEEs calculated for all methods were close to the SEEs of equating with human raters. In two conditions, automated scoring (using BLSTM method with 10% test data rate) led to test equating with fewer errors.

When test equating results made after automated scoring performed with a 20% test data rate were evaluated in terms of random error, the lowest random error (.006) was found in the MS method. The value obtained was quite close to 0 (.006) and was much lower than the SEE (.050) obtained when human raters are used. This value (.006), which was obtained at the 20% test data rate, was followed by the MM method with .098. When using methods based on IRT, HB method showed the highest SEE (.127). When automated scoring was performed at a rate of 20% test data, LT[WS = 1] was the method that shows the lowest SEE (.196) in equating methods based on CTT. This value is followed by the LT equating method with .197. The method with the highest SEE (.405) was the PSMEC equating method. In the equating performed after automated scoring with the 20% test data rate and BLSTM method, the highest SEEs were obtained in equipercentile equating methods in general. In this condition, methods based on IRT generally showed lower SEEs than methods based on CTT. The SEEs calculated for all methods are close to the SEEs of equating with human raters. In four conditions, automated scoring (using BLSTM method with 20% test data rate) led to test equating with fewer errors.

When test equating results made after automated scoring performed with a 33% test data rate wereevaluated in terms of random error, the lowest random error (.012) was found in the MS method. This value obtained is quite close to 0 (.012) and is much lower than the SEE (.050) obtained when human raters were used. This value (.012), whichwas obtained at the 33% test data rate, was followed by the MM method with .071. When using methods based on IRT, the HB method showed the highest SEE (.137). When automated scoring was performed at a rate of 33% test data, LT[WS = 1] was the method that shows the lowest SEE (.200) in test equating methods based on CTT. This value was followed by the LT equating method, with an SEE of .202. The method with the highest SEE (.398) is the EC equating method.

**Table 3.** *Errors related to equating methods based on CTT and IRT.*

| | | SEE | | | | BIAS | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Human | BLSTM | | | Human | BLSTM | | | Human | BLSTM | | |
| | | | %10 | %20 | %33 | | %10 | %20 | %33 | | %10 | %20 | %33 |
| CTT | LC | .211 | .213 | .209 | .215 | **.003** | **.002** | **.002** | **.003** | .211 | .213 | .209 | .215 |
| | LT | .198 | .201 | .197 | .202 | **.003** | **.002** | **.002** | **.003** | .198 | .201 | .197 | .202 |
| | LT (WS=1) | .197 | .200 | .196 | .200 | **.003** | **.002** | **.002** | .004 | .197 | .200 | .196 | .200 |
| | EC | .351 | *.407* | .396 | *.398* | .061 | *.216* | *.159* | *.142* | .357 | *.461* | *.427* | *.423* |
| | EF | .330 | .336 | .347 | .336 | .062 | .032 | .052 | .071 | .336 | .337 | .351 | .344 |
| | EF (WS=1) | .330 | .362 | .371 | .348 | .059 | .048 | .158 | .062 | .335 | .365 | .403 | .353 |
| | PSMEC | *.357* | .328 | *.405* | .350 | .044 | .042 | .087 | .041 | *.359* | .331 | .414 | .352 |
| | PSMEF | .321 | .341 | .360 | .307 | .023 | .021 | .084 | .021 | .322 | .342 | .369 | .307 |
| | PSMEF (WS=1) | .333 | .349 | .371 | .317 | .023 | .021 | .078 | .021 | .334 | .349 | .379 | .318 |
| IRT | MM | .061 | .079 | .098 | .071 | -.010 | .022 | .039 | .010 | **.062** | **.083** | **.106** | **.072** |
| | MS | **.050** | **.047** | **.006** | **.012** | .064 | .128 | .127 | .079 | .081 | .136 | .127 | .080 |
| | HB | .083 | .110 | .127 | .137 | *-.079* | -.108 | -.087 | -.127 | .114 | .154 | .154 | .187 |
| | SL | .083 | .100 | .118 | .119 | *-.079* | -.098 | -.078 | -.118 | .114 | .140 | .141 | .167 |

Note: In terms of SEE, BIAS and RMSE, the lowest coefficient is shown in bold and the highest coefficient in italics in each condition.

In the equating performed after automated scoring with the 33% test data rate and BLSTM method, the highest SEEs were obtained in equipercentile equating methods. In this condition, methods based on IRT generally showed lower SEEs than methods based on CTT. The SEEs calculated for all methods were close to the SEEs of equating with human raters. In four conditions, automated scoring (using BLSTM method with 33% test data rate) made test equating with fewer errors.

When the random errors obtained in all equating processes were evaluated, the errors were very close to each other. In the equating performed by automated scoring, in some cases, lower SEE values were obtained than in the equating performed by human raters. IRT based methods had lower SEE values than methods based on CTT, even if human raters were used or automated scoring was performed. Considering all the equating processes, the lowest SEE value (.006) was obtained using the MS method with BLSTM in automated scoring based on a 20% test data rate. The highest SEE value (.407) was obtained by the EC equating method in all test equating processes performed using BLSTM in automated scoring based on a 10% test data rate.

Systematic error (BIAS) sizes obtained in the equating process with human raters vary between .003 and .079. BIAS values obtained after equating with scores obtained through the BLSTM method based on a 10% test data rate vary between .002 and .216. BIAS values obtained after equating with scores obtained through the BLSTM method based on a 20% test data rate vary between .002 and .159. BIAS values obtained after equating with scores obtained through the BLSTM method based on a 33% test data rate vary between .003 and .142.

When the human raters were taken into consideration, as shown in Table 3, the lowest RMSE was .062 obtained by the MM method based on IRT. This value was followed by .081 with the MS method. When using IRT methods, the highest RMSE (.114) was found in the SL and HB methods. These results mean that moment methods (MM and MS) show lower RMSEs than characteristic curve methods (SL and HB) based on IRT. When using human raters, the method that shows the lowest RMSE (.197) in CTT based equating methods is the LT[WS = 1]. This value is followed by .198 with the LT equating method. The method with the highest RMSE (.359) was the PSMEC equating method. In the case where human raters are used, the highest RMSEs were obtained in equipercentile equating methods. In this condition, methods based on IRT generally showed less RMSEs than methods based on CTT.

When test equating results made after automated scoring performed with a 10% test data rate were evaluated in terms of RMSE, the lowest RMSE (.083) was found in the MM method. This value (.083) was close to the lowest RMSE value (.062) obtained when human raters are used. This value (.083), which was obtained at the 10% test data ratewas followed by MS method with .136. When using methods based on IRT, HB method showed the highest RMSE (.154). When automated scoring was performed at a rate of 10% test data, LT[WS = 1] was the method that shows the lowest RMSE (.200) in test equating methods based on CTT. This value was followed by the LT equating method with .201. The method with the highest RMSE (.461) was the EC equating method. In the equating performed after automated scoring with the 10% test data rate and BLSTM method, the highest RMSEs were obtained in equipercentile equating methods in general. In this condition, methods based on IRT generally showed less RMSEs than methods based on CTT. The RMSEs calculated for all methods were close to the RMSEs calculated from equating with human raters. In one condition (PSMEC), automated scoring (using BLSTM method with 10% test data rate) led to test equating with fewer RMSE.

When test equating results conducted after automated scoring performed with a 20% test data rate and BLSTM were evaluated in terms of RMSE, the lowest RMSE (.106) was found in the MM method. This value (.106) was close to the lowest RMSE value (.062) obtained when human raters were used. This value (.106), which was obtained at the 20% test data rate, was followed by the MS method with .127. When using methods based on IRT, HB method showed

the highest RMSE (.154). When automated scoringwas performed at a rate of 20% test data, LT[WS = 1] was the method that shows the lowest RMSE (.196) in equating methods based on CTT. This value was followed by the LT equating method with .197. The method with the highest RMSE (.427) was the EC equating method. In the equating performed after automated scoring with the 20% test data rate and BLSTM method, the highest RMSEs were obtained with equipercentile equating methods. In this condition, methods based on IRT generally showed lower RMSEs than methods based on CTT. The RMSEs calculated for all methods are close to the RMSEs calculated by equating with human raters. In three conditions, automated scoring (using BLSTM method with 20% test data rate) performed test equating with fewer RMSEs.

When test equating results made after automated scoring performed with a 33% test data rate are evaluated in terms of RMSE, the lowest RMSE (.072) was found in the MM method. This value (.072) was very close to the lowest RMSE value (.062) obtained by human raters. This value (.072), whichwas obtained at the 33% test data rate, was followed by the MS method with .080. When using methods based on IRT, the HB method showed the highest RMSE (.187). When automated scoring was performed at a rate of 33% test data, LT[WS = 1] shows the lowest RMSE (.200) in equating methods based on CTT. This value was followed by the LT equating method with .202. The method with the highest RMSE (.423) was the EC equating method. In the equating performed after automated scoring with the 33% test data rate and BLSTM method, the highest RMSEs were obtained with equipercentile equating methods. In this condition, methods based on IRT generally showed lower RMSEs than methods based on CTT. The RMSEs calculated for all methods are close to the RMSEs of equating with human raters. In four conditions, automated scoring (using BLSTM method with 20% test data rate) performed test equating with fewer RMSEs.

Figure 1 shows RMSE values of the equating performed by human raters and automated scoring based on 10%, 20% and 33% test data rates. The chart was drawn in the range of 0 to 1, since in the literature it was noted that RMSE values below 1% are not important (Pang et al., 2010).

**Figure 1.** *RMSE values of the methods according to the rater type.*

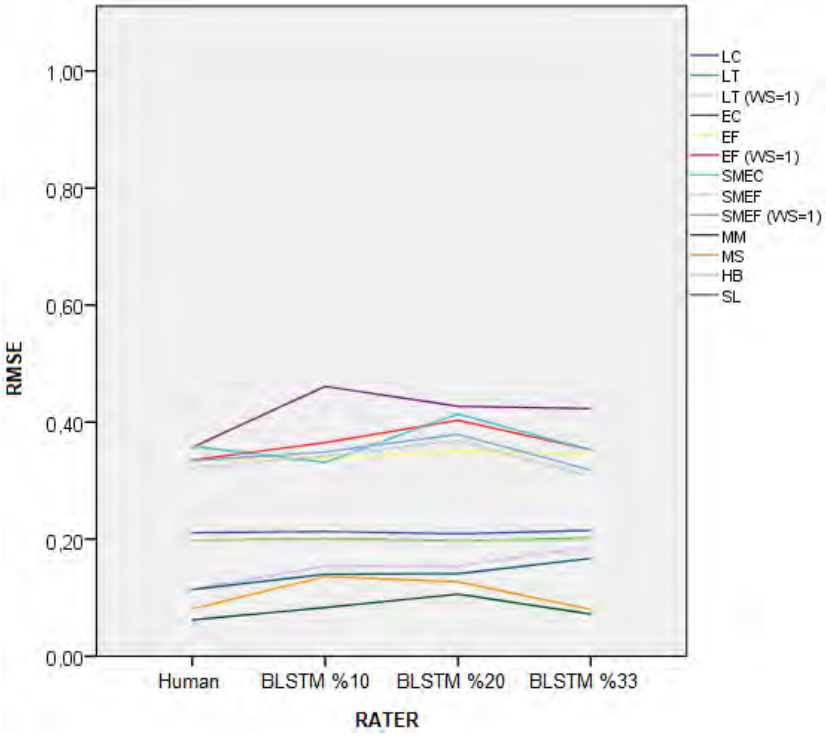

Figure 1 shows that the RMSE values obtained with all equating methods are close to each other. In the equating performed with automated scoring, in some cases, lower RMSE values were obtained than in the equating performed with human raters. IRT based methods had lower RMSE values than methods based on CTT, even if human raters were used or automated scoring was performed. Considering all the equating processes, the lowest RMSE value (.062) was obtained in MM method with the using human raters. In equating with automated scoring scores, the lowest RMSE value (.072) was obtained with the MM method. When IRT test equating methods were compared for each condition, it can be indicated that moment methods showed less error (RMSE) than characteristic curve methods. The highest RMSE value (.359) was obtained in the PSMEC equating method in all test equating process performed using human raters. In automated scoring, the highest RMSE value (.461) was obtained with the EC equating method. In general, equipercentile equating methods equate tests with more RMSE. Changing the synthetic population ratio to 1 generally reduced RMSE values in linear methods. However, in equipercentile equating methods and when pre-smoothing was applied in equipercentile equating methods RMSE values generally increased. Changing the ratio of synthetic population to 1 did not create very large decreases or increases in RMSE coefficients. The pre-smoothing process decreased RMSE values in some cases but increased it in other cases.

The average of errors resulting from test equating performed with the scores obtained by automated scoring with the test data rates of 10%, 20% and 33% were calculated. Then, the significant difference between these averages and the errors of the equating obtained through human raters was examined. Equating methods, variations in synthetic population ratios and/or pre-smoothing versions of these methods have been investigated to determine whether there is a difference between human raters and automated scoring averages. A Mann-Whitney U test was used because the normal distribution assumption was not met for each group. The results are shown in Table 4.

**Table 4.** *Difference test regarding RMSE values obtained as a result of human raters and automated scoring.*

| | Rater | N | Mean Rank | Sum of Ranks | U | *p* |
|---|---|---|---|---|---|---|
| RMSE | Human Scoring | 13 | 12.000 | 156.000 | 65.000 | .336 |
| | Automated Scoring | 13 | 15.000 | 195.000 | | |

Table 4 shows that the RMSE values (median = .211) of 13 equating methods obtained through human raters did not differ significantly from the mean RMSE values (median = .212) of 13 equating methods obtained through automated scoring (U = 65,000, $p$ = .336 > .05). Accordingly, the use of human raters or automated scoring did not have a significant effect on the RMSE values obtained as a result of the equating process. The effect size was investigated through the Cliff's Delta coefficient and -.18 was found. This effect size is small (Cliff, 1993). The relationship between the errors of the equating (RMSE) performed by human raters and the averages of the equating errors (RMSE) performed by automated scoring was evaluated with the correlation of Spearman rank differences and at a high and significant level relationship was found ($r$ = .96, $p$ = .00 < .05).

## 4. DISCUSSION and CONCLUSION

Three equating procedures were performed in the study according to the test data rates used in automated scoring. The equating process was carried out for human scorers as well as for automated scoring. In the equating process for human raters, the final scores of the human raters for the $A_1$ and $B_1$ booklets were used. In the equating process for automated scoring, the scores

obtained by the automated scoring of the constructed-response items in both test forms were used. Constructed-response items and objectively scored items are not subjected to equating separately. Methods based on CTT and IRT have been used as the equating method.

This study had found that the errors (RMSE) obtained in all methods and different combinations of methods in automated scoring conditions and in the condition where human raters were found similar. In some cases, lower RMSE values were found in the equating performed through automated scoring than human raters' equating processes. It was observed that pre-smoothing decreased RMSE values in some cases but increased in other cases. Hagge et al. (2011) determined that the pre-smoothing reduced the standard error of chained equipercentile equating and frequency estimation methods. This study changed the ratio of synthetic population decreased RMSE values in linear equating methods, while it increased RMSE values in equipercentile equating methods. However, it should be noted that equating errors presented here were based on automated scoring conditions. The result of the equating showed that methods based on IRT equate tests with lower errors (in terms of SEE and RMSE) compared to methods based on CTT either in automated scoring conditions or when human raters were used. Hagge and Kolen (2011) and Liu and Kolen (2011) stated that methods based on IRT showed lower errors than the methods based on CTT according to the root mean squared error in conditions like this study. Liu and Kolen (2011) also found that IRT true score equating methods had lower SEE values than frequency estimation and chained equipercentile equating methods. Although the same criterion is not considered, Lee et al. (2012) stated that IRT true score equating performed better than Tucker linear, chained equipercentile, frequency estimation, pre-smoothed chained equipercentile, and pre-smoothed frequency estimation methods in terms of primary level equality. Wolf (2013) also found that in terms of primary level equality, IRT true score equating performed better than frequency estimation and chained equipercentile equating. Hagge et al. (2011) stated that IRT based methods had lower SEE values than CTT based methods. However, these studies weren’t equating based on automated scoring. When methods based on IRT were compared for each condition, moment methods equate with less error than characteristic curve methods. This situation may be related to linearity besides the number of common items and test length. The highest RMSE and SEE values are found in equipercentile equating methods.

Regarding RMSE and SEE, the highest errors were obtained in the chained equipercentile and pre-smoothed chained equipercentile equating methods. Hagge and Kolen (2011) and Hagge et al. (2011) also stated that the method with the highest SEE value wass chained equipercentile equating. However, He (2011) stated that the chained equipercentile equating method performed better than frequency estimation method according to primary level equality criterion. The difference between this study and He (2011) is thought to be due to the sample size. In automated scoring, the average RMSE values of different test data rates for each equating method were calculated and the statistical differences of these values from the errors of equating performed by human raters were examined. As a result, it was determined that there was no significant difference between the errors and that the errors showed a high level of compliance. Olgar (2015) used the open-ended items as common items by scoring them automatically and stated that even though the common items were multiple-choice items or open-ended items scored automatically with multiple-choice items, the results were similar. He even found that the including automatically scored open-ended items in common items yielded better results in some cases. Almond (2014) stated that in tests consisting only of constructed-response items, linear logistic equating can be used as an alternative by automatically scoring common items with generic e-rater.

In cases where automated scoring is made, based on the results of this study, methods based on IRT in equating procedures are recommended. This study was carried out on approximately 1200 people. In subsequent studies, the effect of automated scoring on the equating process can

be examined using larger samples. This study determined the effect of changing the synthetic population ratio on equating errors under automated scoring conditions. In future studies, when there is a difference between the number of groups to be equated, the effect of the synthetic population ratio to .5 can be evaluated. This study also discussed the effect of pre-smoothing under automated scoring conditions. In further research, pre- and post-smoothing can be compared, and different pre- and post-smoothing methods can be examined under different patterns.

## Acknowledgments

This paper was produced from part of the first author's doctoral dissertation prepared under the supervision of the second author. Thanks to Behzad Naderalvojoud for his support in the creation of the automated essay scoring software used in this research.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

## Authorship contribution statement

**İbrahim UYSAL**: Investigation, Software, Methodology, Formal Analysis, Visualization, Resources, and Writing the original draft. **Nuri DOĞAN**: Investigation, Software, Methodology, Supervision, and Validation.

## ORCID

Ibrahim UYSAL https://orcid.org/0000-0002-6767-0362
Nuri DOGAN https://orcid.org/0000-0001-6274-2016

## 5. REFERENCES

Adesiji, K. M., Agbonifo, O. C., Adesuyi, A. T., & Olabode, O. (2016). Development of an automated descriptive text-based scoring system. *British Journal of Mathematics & Computer Science, 19*(4), 1-14. https://doi.org/10.9734/BJMCS/2016/27558

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1-36. https://doi.org/10.18637/jss.v074.i08

Almond, R. G. (2014). Using automated essay scores as an anchor when equating constructed-response writing tests. *International Journal of Testing, 14*(1), 73-91. https://doi.org/10.1080/15305058.2013.816309

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Educational Testing Service.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-30. http://www.jtla.org.

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(1), 87-101. https://doi.org/10.1080/10705511.2014.934850

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3*(2), 77-85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Chen, H., Xu, J., & He, B. (2014). Automated essay scoring by capturing relative writing quality. *The Computer Journal, 57*(9), 1318-1330. https://doi.org/10.1093/comjnl/bxt117

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*(3), 494-509. https://doi.org/10.1037/0033-2909.114.3.494

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Routledge.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *The Journal of Applied Psychology, 78*(1), 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Pearson.

Deng, W., & Monfils, R. (2017). *Long-term impact of valid case criterion on capturing population-level growth under Item Response Theory equating* (Research Report 17-17). Educational Testing Service. https://doi.org/10.1002/ets2.12144

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2-18. https://doi.org/10.1037/a0024338

Gonzalez, J., & Wiberg, M. (2017). *Applying test equating methods: Using R*. Springer.

Hagge, S. L., & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. In M. J. Kolen & W-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1, pp. 95-135). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Hagge, S. L., Liu, C., He, Y., Powers, S. J., Wang, W., & Kolen, M. J. (2011). A comparison of IRT and traditional equipercentile methods in mixed-format equating. In M. J. Kolen & W-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1, pp. 19-50). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

He, Y. (2011). Evaluating equating properties for mixed-format tests [Doctoral dissertation, University of IOWA]. https://ir.uiowa.edu/etd/981/

Kaiser, H. F. (1970). A second-generation little jiffy. *Psychometrika, 35*(4), 401-415. https://doi.org/10.1007/BF02291817

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement, 34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking* (2nd ed.). Springer.

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different Item Response Theory scaling methods. *Educational and Psychological Measurement, 71*(2), 362–379. https://doi.org/10.1177/0013164410375111

LaFlair, G. T., Isbell, D., May, L. D. N., Arvizu, M. N. G., & Jamieson, J. (2017). Equating in small-scale language testing programs. *Language Testing, 34*(1), 127-144. https://doi.org/10.1177/0265532215620825

Lee, E., Lee, W-C., & Brennan, R. L. (2012). *Exploring equity properties in equating using AP® examinations* (Report No. 2012-4). CollegeBoard.

Liu, C., & Kolen, M. J. (2011). A comparison among IRT equating methods and traditional equating methods for mixed-format tests. In M. J. Kolen & W-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1, pp. 75-94). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*(1), 88-91. https://doi.org/10.3758/BF03192753

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218. https://doi.org/10.1207/s15326985ep3404_2

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed-response, performance testing, and portfolio assessment* (pp. 61-73). Lawrence Erlbaum Associates, Inc.

MoNE. (2017). Monitoring and evaluation of academic skills (ABİDE) 2016 8th grade report. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_11/30114819_iY-web-v6.pdf

Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (Report No. 06-05). Educational Testing Service.

Moses, T., von Davier, A. A., & Casabianca, J. (2004). *Loglinear smoothing: An alternative numerical approach using SAS* (Research No. 04-27). Educational Testing Service.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Olgar, S. (2015). *The integration of automated essay scoring systems into the equating process for mixed-format tests* [Doctoral dissertation, Florida State University]. http://diginole.lib.fsu.edu/islandora/object/fsu%3A253122

Page, E. B. (1966). The imminence of grading essays by computers. *Phi Delta Kappan, 47*(5), 238–243. http://www.jstor.org/stable/20371545

Pang, X., Madera, E., Radwan, N., & Zhang, S. (2010). *A comparison of four test equating methods* (Research Report). Eqao.

R Development Core Team. (2018). *R: A language and environment for statistical computing* (version 3.5.2) [Computer software]. R Foundation for Statistical Computing.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students: Validity, technical adequacy and implementation* (pp. 213-231). Lawrence Erlbaum Associates.

SAS Instutite. (2015). *Statistical analysis software* (version 9.4) [Computer software]. SAS Institute.

Spence, P. D. (1996). *The effect of multidimensionality on unidimensional equating with item response theory* [Doctoral dissertation, University of Florida]. https://www.proquest.com/docview/304315473

Tankersley, K. (2007). *Tests that teach: Using standardized tests to improve instruction*. Association for Supervision and Curriculum Development.

Tate, R. (2000). Performance of a proposed method for the linking of mixed-format tests with constructed-response and multiple-choice items. *Journal of Educational Measurement, 37*(4), 329-346. http://www.jstor.org/stable/1435244

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220. https://doi.org/10.1037/a0023353

Torchiano, M. (2020). *effsize: Efficient Effect Size Computation* (Version 0.8.1) [Computer software]. https://CRAN.R-project.org/package=effsize

Weiss, D. J., & Minden, S. V. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG* (Technical Report). Assessment Systems Corporation.

Wolf, R. (2013). *Assessing the impact of characteristics of the test, common-items, and examinees on the preservation of equity properties in mixed-format test equating* [Doctoral dissertation, University of Pittsburgh]. https://core.ac.uk/download/pdf/19441049.pdf

Yoes, M. E. (1996). *User's manual for the XCALIBRE marginal maximum-likelihood estimation program* [Computer software]. Assessment Systems Corporation.

Zu, J., & Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement, 47*(4), 395-412. https://doi.org/10.1111/j.1745-3984.2010.00120.x