Journal of Educational Science and Technology
Volume 5 Number 2 August 2019 page 100-109
p-ISSN:2460-1497 and e-ISSN: 2477-3840
DOI: https://doi.org/10.26858/est.v5i2.9258

Equating Method for Prevent Discrimination in Classroom

Deni Iriyadi[1], Dali Santun Naga[2], Wardani Rahayu[3]
[1] Mathematics, Measurement and evaluation, Universitas Negeri Jakarta, Indonesia
Email: deni.iriyadi@unj.ac.id
[2] Informatics Engineering, Universitas Tarumanagara, Indonesia
Email: dalinaga@gmal.com
[2] Mathematics, Universitas Negeri Jakarta, Indonesia
Email: wardani.rahayu@unj.ac.id

(***Received***: May-2019; ***Reviewed:*** June-2019; ***Accepted:*** July-2019; ***Published:*** August-2019)

ABSTRACT
This study aims to determine effectivenes Simplified Cirlce Arc method used at the school level to prevent discrimination against student grades. This study National Examination data on mathematics subjects from of the Center for Educational Assessment in DKI Jakarta and Tangerang regions. Using the Rasch Model analysis, data obtained for 2135 in the DKI Jakarta (X) and 2271 in the Tangerang area (Y). The data was obtained after conducting Rasch analysis with Mean Square Outfit (MNSQ) of 0.5 <MNSQ <1.5. Replication is done 50 times for each form data distribution from each region. Results of replication then RMSE value is calculated. The results showed that equal form with normal data distribution, statistically the average RMSE with the Simplified Circle Arc method smaller than the average RMSE result of equalization with the Nominal Weight Mean method which indicates that the Simplified Circle Arc method more accurate than Nominal Weight Mean method. Likewise, with equal equations with positive skeweness and negative skewness data distribution, the average RMSE with the Simplified Circle Arc Method is smaller than average RMSE resulting in equalization of the score with Nominal Weight Mean method. A small RMSE value indicates a fairly accurate result of equalization.

Keyword*: equating; simpified circle arc; nominal weight mean; data distribution.*

INTRODUCTION

Assessment is a process carried out in the learning process by increasing learning from students as an evaluation material for improving learning to renewal. Apart from being an evaluation material, the results of the assessment are made as a benchmark to see the quality of students in an Education (Antara & Bastari, 2015). The commonly used form of assessment is a multiple choice test because this form of test can be easily used to measure several aspects at once (Ebel & Frisbie, 1991). The preparation of these items is always based on the grid. Both the teacher and the government do this. In Indonesia, the government issues a grid for each subject tested in the National Examination so that students can focus more on learning. The government in compiling the items to be tested on students certainly must consider many things, one of which is regarding advice and infrastructure. Equitable distribution of facilities and infrastructure in Indonesia has not been fulfilled. Thus, the government made several test kits that were adjusted to the balance. Since 2007 the number of test kits in the UN has changed. Which initially only had one test device to become 20 packages in 2013 to 160 packages in 2014 throughout Indonesia. The package is divided into 8 regions which will print different test devices (Pos, 2018). In the

National Exam Standard Operational Procedure issued by the National Education Standards Agency, there is no clearly stated number of packages throughout Indonesia, but certainly in one test room, there ismore, than one test. As a result of this, the questions made will have differences even though the content contains the same thing.

In practice, the assessment process for the smallest unit occurs at the school level. At the primary to the high school level, the Indonesian government has standards regarding the number of students in one class. The number is not more than 50 students even in a number of excellent schools the number of students in the class is not more than 36 people. The amount is based on the Republic of Indonesia Minister of Education and Culture Regulation Number 17 of 2017 concerning New Student Admission at Article 24, which is the maximum number of students in one class with a maximum of 36 people (Kemendikbud, 2017).

The fewer the number of students, the process of learning and teaching can be more effective for teachers. In this regard, teachers as implementers of learning in the classroom certainly need a form of assessment to see the achievements of their students. As explained earlier that in one class level, sometimes there is more than one teacher who teaches the same subject. In compiling test kits, they are only based on the agreed upon the grid. Of course, it's unfair when the grades of class A are compared to class B taught by different teachers. Therefore, it is necessary to use an equalization method that is considered appropriate for use in accordance with the characteristics in the class, especially for the number of students. Several studies have been conducted for that with a limited number of samples (sampel kecil).

The results of the assessment of different test kits are treated equally without regard to several aspects such as the level of difficulty. This can be beneficial or detrimental to some students. Being an unfair thing for students. The main problem is how to interpret the results of the acquisition of students who have worked on different test kits to prevent discrimination. When student exam results are used as a benchmark for graduation obtained from different test kits, of course this is not appropriate. The secret comes from a different test device even though the same grid cannot be directly compared. When two values come from two different test devices compared to the value of the two they cannot be exchanged. 70 of the X test kits are certainly not the same as 70 in the Y test device. This is because the scales from both devices do not have the same scale (Zhu, 1998). For this reason, a process is carried out to eliminate discrimination in the form of equal equivalence. This equalization is considered fair enough. Basically what is done is only to do general scaling so that the scores of various test devices can be compared (Zhu, 1998). Once this has been done, it will be scorched from the X test device and set the Y test device on the same scale.

Various methods of equalization based on classical methods have been presented by several experts. Aminah (2012) in his research comparing the Linear (Tucker and Levine) method with Equipercentil (Braund-Holland and Chained), Skaggs (2005) which compares the Linear, Mean, Unsmootied, and Log-Linear methods, Ozdemir (2017) compare the Equipercentil method with Circle Arc, Aşiret & Sünbül (2016) which compares the methods of Idenetity, Mean, Linear, Circle Arc and Presmooted, Livingston & Kim (2008) which compares the Circle Arc and Linear methods, serta Babcock, Albano, & Raymond (2012) which compares the Mean Weight Nominal, Chained, Linear, Circle Arc, Identity and Synthetic. Based on these methods a new comparison can be made in the hope of providing the best choice for the use of an effective equating method. Livingston & Kim (2010b) conducted research by comparing the Simetryc and Simplified Circle Arc methods with several other methods but did not compare the accuracy between the two Circle Arc methods.

Ozdemir (2017) states that the Circle Arc method has superior results compared to the equipercentil method where both methods are classified as nonlinear methods based on the classical method. This is based on the generated RMSE. Livingston and Kim made modifications to the Circle Arc method that already existed before and divided the method into two forms, one based linearly while the other contained nonlinear elements even though there were still linear elements (Livingston & Kim, 2008). Furthermore in another study, Livingston used this method to do a number of different conditions including the number of samples and showed that this method provides accurate results based on RMSD values and bias (Livingston & Kim, 2009, 2010). Research

conducted by Aşiret & Sünbül (2016) states that the Circle Arc method produces a lower equalization error than the other methods for using small samples.

The form of distribution also has a role in the equalization process. In line with this, research Uysal & Kilmen (2016) suggests that the distribution of abilities also influences the results of equalization. The study uses a modern theoretical approach so as to estimate the ability of respondents. Furthermore, Uysal and Kilmen divide the 3 distributions namely Normal, Positive Skewness, and Negative Skewness. In line with this, S. Kim, von Davier, & Haberman (2008) states that choosing the form of distribution can be influenced by the form of group distribution which will be equalized. The results of previous studies also stated that the distribution of abilities also had an effect on equating results (Uysal & Kilmen, 2016). The difference is about distribution used which previously saw the distribution of capabilities, this study uses data distribution.

The classic method is identical to the raw score or total score obtained by the respondent from the results answering a number of questions. As previously explained, the distribution used in several studies is the distribution of abilities. This is interesting to be one aspect studied. The question arises about this when applied to the distribution using the classical method which basically uses the raw score, of course the distribution used is the distribution of the raw score itself. The existing data distribution is not only normal, but when viewed from the distribution skewness, the data distribution is divided into two, namely positive skewness with a longer tail to the right and negative skewness with a longer tail to the left. Both types of distribution fall into the category of abnormal distribution. Thus, there are three types of distribution that will be the focus of this study namely (1) normal distribution, (2) positive skewness distribution, and (3) negative skewness.

With the explanation described above, it is deemed necessary to conduct research on effective equalization methods to be used at the school level with a limited number of samples. It aims to prevent discrimination against student grades. Considering the target of this research is the class teacher, a small sample is used as a representation of the number of students in the class belonging to the small sample. Several equalization methods were developed to be able to overcome discrimination issues regarding scores obtained from two different test devices. Liner method was developed to answer the problem. In addition, there is also a method of Nominal Weight Mean Equating which is basically developed also for the same reason, namely for equalization of the scale in small samples. Both methods are considered appropriate to be compiled. Both are practical and easy to implement for teachers to avoid discrimination in the assessment process in class.

METHOD

This study used two sample groups. The two groups were given different types of test kits but came from the same grid. This study uses data from 2 SMP National Exam packages from the Education Assessment Center (PUSPENDIK) for the DKI Jakarta and Tengerang areas in 2015 on mathematics subjects. The selection of the two places was based on the characteristics of the National Examination on both of them who had similarities on several items (anchor items) in accordance with the predetermined research design, namely equalization on test devices that have anchor items. Analysis of the Rasch model is used to measure the level of suitability of the respondents (person fit) with the model with the acceptable criteria for Outfit Mean Square (MNSQ) value of $0.5 < MNSQ < 1.5$. For the X test equipment 2135 responses were received while the Y test was obtained by 2271 students. After analyzing the Rasch model, there were 2 respondents who were not fit for the X test equipment and 233 respondents who were not fit for the Y test device. Thus, in the X test device there were 2133 responses of students and the Y test kit had 2048 responses from students.

This research uses RMSE replication results as a tool to evaluate the results of the score equalization. Selecting samples from each population at random with random sampling with replacement with the help of the SPSS application. Randomization was carried out 50 times as shown in the following table 1 which explains the RMSE of each replication carried out for the form of data distribution and the equalization method. Each equalization method consists of 50 replications which means there are 50 RMSE from each data distribution.

Tabel 1. Replication of Forms of Data Distribution

Data Distribution (B)	Equating Method (A)	
	SCA (A_1)	NWME (A_2)
Normal Distribution (B_1)	$RMSE_1$	$RMSE_1$
	$RMSE_2$	$RMSE_2$
	…….	…….
	$RMSE_{50}$	$RMSE_{50}$
Skewness Positive Distribution (B_2)	$RMSE_1$	$RMSE_1$
	$RMSE_2$	$RMSE_2$
	…….	…….
	$RMSE_{50}$	$RMSE_{50}$
Skewness Positive Distribution (B_3)	$RMSE_1$	$RMSE_1$
	$RMSE_2$	$RMSE_2$
	…….	…….
	$RMSE_{50}$	$RMSE_{50}$

This study uses the Simplified Circle Arc and Nomial Weight Mean Equating method. In short, the equation for equalization can be written using the Nominal Weight Mean Equating method according to equation (1) as follows:

$$Y^*_{NWME} = X - \mu(Y) + \mu(X) + \left[\frac{N(Y)K(X) + N(X)K(Y)}{[N(X) + N(Y)]K(Z)}\right][\mu(Z_Y) - \mu(Z_X)] \quad (1)$$

From the results of the equalization, the equalization value of Root Mean Square Error (RMSE) is determined. RMSE is used to evaluate the results of a research study (Chai & Draxler, 2014). Thus, this can also be used to determine the accuracy of several equalization methods in conducting equal equivalence. Each RMSE value is determined using equation (2) as follows (Babcock et al., 2012; Joo, Lee, & Stark, 2016; ; Shin, 2015):

$$RMSE_{NWME} = \sqrt{\frac{\sum_{i=1}^{N}\left(Y^*_{NWME_i} - X_i\right)^2}{N}} \quad (2)$$

Where N is the number of respondents, is the same as the equalization result, and X_i is equal. RMSE is used to determine the accuracy of the equalization method used (Aşiret & Sünbül, 2016; Uysal & Kilmen, 2016). According to Kartono (Karton, 2008) that a small mean value indicates a better quality of equalization. As explained earlier that the RMSE value is a value that indicates the good or not the results of a measurement. This value is obtained from each result of replication carried out. The number of RMSE values depends on the amount of replication performed (M). To assess the accuracy of the RMSE results given, then a mean test was made of these values.

$$\mu_{RMSE_{NWME}} = \frac{\sum_{i=1}^{M} RMSE_{NWME_i}}{M} \quad (3)$$

The equation for equalization using the Simplified Circle Arc method for linear components and the component curve are as follows:

$$Y^*_{SCA_1} = Y_{min} + Y_{c(SCA)} + \frac{Y_{max} - Y_{min}}{X_{max} - X_{min}}(X - X_{min}) + \sqrt{r^2_{SCA} - (X - X_{c(SCA)})^2} \quad (4)$$

Or

$$Y^*_{SCA_2} = Y_{min} + Y_{c(SCA)} + \frac{Y_{max} - Y_{min}}{X_{max} - X_{min}}(X - X_{min}) - \sqrt{r^2_{SCA} - (X - X_{c(SCA)})^2} \quad (5)$$

Similar to the Nominal Weight Mean Equating method, the RMSE value in the Simplified Circle Arc method is also calculated using equation (4) dan (5). A small RMSE value shows the results of good equivalence.

$$RMSE_{SCA} = \sqrt{\frac{\sum_{i=1}^{N}\left(Y^*_{SCA_i} - X_i\right)^2}{N}} \quad (6)$$

From the several RMSE values obtained then the mean value is calculated. In relation to the small value of RMSE, it is expected that the average of the RMSE group is of little value, so the value of the parameters we obtain is quite sharp or sufficiently accurate.

$$\mu_{RMSE_{SCA}} = \frac{\sum_{i=1}^{M} RMSE_{SCA_i}}{M} \quad (7)$$

A small RMSE value shows a better quality of equalization (Karton, 2008). In addition, the RMSE value can also be used to determine the accuracy of the equalization

method used (Aşiret & Sünbül, 2016; Uysal & Kilmen, 2016).

RESULT AND DISCUSSION

Result

The data in this research used the National Examination data for junior high school students from two different places which consisted of 40 items of the exam. The Rasch model analysis was applied to obtain items that were fit to be used in the subsequent analysis with acceptable criteria for Mean Square (MNSQ) value in which 0.5 < MNSQ <1.5 and the value of Z-standard was -2 <ZSTD Outfit <2 (Anshel, Weatherby, Kang, & Watson, 2009; Linacre JM, 2006; Neumann, Neumann, & Nehm, 2011; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). However, some experts did not recommend the ZSTD criteria when the sample size was more than 500. From the fit item analysis, in X test kit ,1 item that was not fit was item 30 as it had MNSQ of 1.70 while for in Y test 4 items were found not fit, which were point 3, 10, 37, 26 with the MNSQ value of 2.73; 1.76; 1.58; 1.51 respectively. Based on the calculation of the Rasch model and the design of the previous study, 30 items were selected in which there were 6 anchor items (20% of the total items) that were used as research instruments. In X test kit the anchor items were item 2, 6, 17, 22, 34, and 38 (6 items in total), while the non-anchor items were item1, 3, 4, 5, 7, 8, 11, 12, 13, 15, 19, 20, 21, 23, 25, 28, 29, 31, 32, 33, 35, 36, 37, and 40 (24 items in total). In Y test kit, (POC5530 question code) the anchor items were item 4, 8, 17, 29, 32, and 39 (6 items in total), while the non-anchor items were item 1, 2, 5, 6, 7, 9, 11, 13, 14, 15, 16, 19, 22, 23, 25, 28, 30, 31, 33, 34, 35, 36, 38, and 40 (24 items in total). Thus, for the two test kits, there were 30 questions with each of the tests having 24 non-anchor items and 6 anchor items.

Out of 40 items from the two test packages, 30 items were selected having 20% (6 items) of anchor items. The item was then used to calculate the equalization score using two different methods (simplified circle arc and moninal weight mean). In addition, the conditions for the dimensions were also tested against these items to ensure that the instruments used only measured 1 dimension. For the X test kit, based on the results of the Rasch model analysis, the raw variance value was 30.4% while for the Y test kit the raw variance value was 33.1%. Both of these values were above the minimum value of the minimum requirement, which was 20% (Hsiao, Shih, Yu, Hsieh, & Hsieh, 2015; Sinnema, Ludlow, & Obinson, 2016).

From the results of the analysis of the Rasch model, replication of the number of respondents is available to assess the RMSE produced. Replication is done 50 times for each group of data then the RMSE value is calculated. The following are the RMSE results from the Simplified Circle Arc and Nomial Weight Mean methods.

Tabel 2. Description of RMSE Value of Equivalent Score Results

Statistic	*Simplified Circle Arc*			*Nomial Weight Mean*		
	N-N	SP-SP	SN-SN	N-N	SP-SP	SN-SN
N	50	50	50	50	50	50
Average	0.34	0.44	0.29	0.61	0,75	0.44
Variance	0.08	0.10	0.04	0.22	0,25	0.08
Median	0.28	0.37	0.23	0.50	0,70	0.35
Maximum	1.10	1.37	0.70	1.70	1,90	1.10
Minimum	0.05	0.05	0.06	0.10	0,10	0.10

Note. N – N = normal distribution with normal distribution
SP – SP = positive skewness distribution with positive skewness distribution
SN – SN = negative skewness distribution with negative skewness distribution

Discussion

Table 2 shows the distribution using the simplified circle arc method more effectively to use based on the RMSE value generated. RMSE is used to evaluate the results of a research study (Chai & Draxler, 2014). Thus, this can also be used to determine the accuracy of several equalization methods in conducting equal equivalence. Of the three forms of data

distribution pairs, all of them produce low values rather than RMSE from the nominal weight mean method. As previously explained, a low RMSE value indicates the results / quality of an equal score is good and can be used (Aşiret & Sünbül, 2016; Uysal & Kilmen, 2016). Figure 5 shows a mean comparison of RMSE between the simplified circle arc method and the nominal weight mean. It is clear that the difference between the two is that from all conditions the data distribution pairs all produce a mean RMSE from the simplified circle arc method that is low.

If viewed from the variance value, there are differences between the SCA and NWME methods. Variance shows the diversity of data from a group. The diversity can be seen from the difference between the unit score and the average value. The greater the difference between the unit score against the average value will result in a large variance which means that the data in the group is diverse or in other words inconsistent. This consistency is closely related to the precision or accuracy of a measurement if done repeatedly. Precision of a measurement system means the extent to which repetition measurements in unchanged conditions get the same results (Taylor, 1997). Precision can be observed from the amount of variance that is ownedi. .. Large measurement errors can cause accuracy of the measurement process to be doubted. This makes measurement errors a matter to be taken into account. The amount of measurement error made can be verified through its variance. This is done by repeating measurements then the variance of the measurement results is small indicating the precision of the measurements made. The similarity is measured through the variance of the measurement results

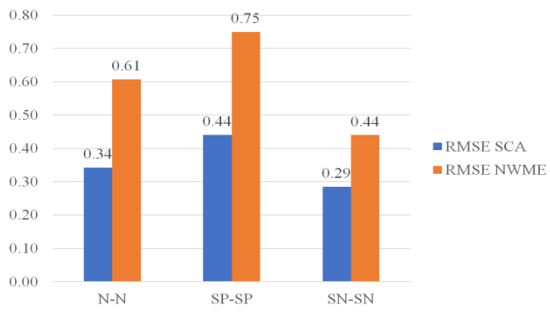

Figure 1. Average RMSE from the simplified circle arc and nomial weight mean methods

Figure 1 shows the position of the average RMSE simplified circle arc method which is lower than the average RMSE nominal weight mean method. While Figure 2 shows the variance value of the average RMSE. The two images show a simplified circle arc method that is more accurate than the nominal weight mean equating method. Besides being reviewed from the average RMSE also from the variance value which is always smaller. A small variant will give good measurement results (Suero et al., 2017; Verde et al., 2006; Walther et al., 2005).

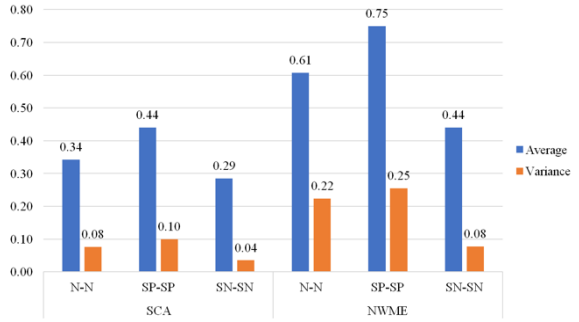

Figure 2. RMSE Review and Score Equivalence Variance

The simplified circle arc method will always provide accurate equalization results for all forms of data distribution pairs. The method is divided into 2 parts, namely linear components and curve components. The curve component in the simplified circle arc method has a role on the accuracy of the equalization results produced. This component is divided into two parts depending on the position of the middle value of the transformation results. When the middle value of transformation results is positive, then use equation (5) while when the middle value is negative, then use equation (4) In the form of normal data distribution, it is not a problem in both equations. Both equations (4) and (5) will all result in the equal score scoring than the nominal weight mean method.

In equating with positive skewness data distribution, the value of equalizing the Simplified Circle Arc method will be careful when on the curve component to use the equation (5).This happens because the positive skewness distribution of data is likely to be at a low value. The average data will be of little value. When the data is entered in equation, the mean value is getting smaller. The small mean means that the data in the group are generally small (such as positive skewness conditions). In relation to variance which is the distance

between the data and the average value, it will produce a small variance value as a result of the density of the data with the mean. If seen from the formula for variance it appears that the difference between the value of the equalization and the average value is reduced.

In equating with the distribution of negative skewness data, the value of the equalization method of the Simplified Circle Arc method will be reflected when on the component curve for the Simplified Circle Arc method using equation (4). This happens because the negative skewness distribution of data is likely to be at a high value. The equalization curve will curve open down (positive). If it is seen from the equation for variance, it appears that the difference between the value of the equalization and the average value of the equalization results. Of course the Simplified Circle Arc method will be small. Similar to the form of positive skewness data, in the form of negative skewness data groups of data generally gather at a high value summed with a value in the form of a curve so that the value group has a smaller range so that the variance will also be smaller. Equation (5) is what makes the result of equal equations on the Simplified Circle Arc method for the initial data form with negative skewness distribution will result in small variations so that specifically in the form of negative skewness data the meticulous method of equality is used namely Simplified Circel Arc with curve equation (4). In contrast, when the value group with negative skewness distribution is reduced by a group of data in the form of a curve, the range of values produced will be greater as well as the resulting variation.

ACKNOWLEDGMENTS

Many thanks are given to universities for providing complete facilities and also to the Education Assessment Center which provides the results of National Exams for mathematics subjects in DKI Jakarta and Tangerang in 2015

CONCLUSION AND SUGGESTION

As already explained, the small average value of RMSE shows the results of a good measurement of a method. The mean RMSE is obtained from repeated measurements through replication. The variance of the RMSE values in the simplified circle arc method is small. This shows that the method is precise in making measurements. The mean and variance of RMSE shows that the simplified circle arc method is better than the nominal weight mean equating method. Thus, the teacher as the executor of the assessment in the class can use the method to equalize the values of students from different classes. This was done to eliminate discrimination against students.

Students are taught by different teachers at one level of education, so all forms of teaching and assessment will be different. Of course, the test device that is made will also be different even though it uses the same grid guide. The results of the assessment of different test kits are treated equally without regard to several aspects such as the level of difficulty. This can be beneficial or detrimental to some students. Being an unfair thing for students. The main problem is how to interpret the results of the acquisition of students who have worked on different test kits to prevent discrimination. When student exam results are used as a benchmark for graduation obtained from different test kits, of course this is not appropriate. The secret comes from a different test device even though the same grid cannot be directly compared. Through this score distribution, this can be overcome, especially the equalization of scores by using the simplified circle arc method. By equalizing these values, a benchmark of common values will emerge that can be used to determine whether or not a student pass.

REFERENCES

Aminah, N. S. (2012). Karakteristik metode penyetaraan skor tes untuk data dikotomos. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *16*(Special Issue for UNY's 48th Dies-Natalis), 88–101. https://doi.org/10.21831/pep.v16i0.1107

Anshel, M. H., Weatherby, N. L., Kang, M., & Watson, T. (2009). Rasch calibration of a unidimensional perfectionism inventory for sport. *Psychology of Sport and Exercise*, *10*(1), 210–216. https://doi.org/10.1016/j.psychsport.2008.07.006

Antara, A. A. P., & Bastari, B. (2015). Penyetaraan Vertikal Dengan Pendekatan Klasik Dan Item Response Theory Pada Peserta didik Sekolah Dasar. *Jurnal Penelitian Dan Evaluasi Pendidikan*,

19(1), 13–24. https://doi.org/10.21831/pep.v19i1.4551

Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Kuram ve Uygulamada Egitim Bilimleri*, *16*(2), 647–668. https://doi.org/10.12738/estp.2016.2.2762

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal Weights Mean Equating: A Method for Very Small Samples. *Educational and Psychological Measurement*, *72*(4), 608–628. https://doi.org/10.1177/0013164411428609

Caglak, S. (2016). Comparison of Several Small Sample Equating Methods under the NEAT Design. *Turkish Journal of Education*, *5*(3), 96–118. https://doi.org/10.19128/turje.16916

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev*, *7*, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Dwyer, A. C. (2016). Maintaining Equivalent Cut Scores for Small Sample Test Forms. *Journal of Educational Measurement*, *53*(1), 3–22. https://doi.org/10.1111/jedm.12098

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement*. *Educational Researcher* (Fifth Edit). New Delhi: Rajkamal Electirc Press. https://doi.org/10.2307/1175572

Hippel, P. Von. (2010). Skewness. *International Encyclopedia of Statistical Science*, *100*, 1–4. https://doi.org/http://dx.doi.org/10.4135/9781412952644

Hsiao, Y. Y., Shih, C. L., Yu, W. H., Hsieh, C. H., & Hsieh, C. L. (2015). Examining unidimensionality and improving reliability for the eight subscales of the SF-36 in opioid-dependent patients using Rasch analysis. *Quality of Life Research*, *24*(2), 279–285. https://doi.org/10.1007/s11136-014-0771-z

Joo, S.-H., Lee, P., & Stark, S. (2016). Evaluating Anchor-Item Designs for Concurrent Calibration With the GGUM. *Applied Measurement in Education*, 1–14. https://doi.org/10.1177/0146621616673997

Kartono. (2008). Equating the Combined Dichotomous and Polytomous Item Test Model in an Achievement Test. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *12*(2), 302–320.

Kim, S., Davier, A. A. von, & Haberman, S. (2008). Small-Sample Equating Using a Synthetic Linking Function. *Journal of Educational Measurement*, *45*(4), 325–342. https://doi.org/10.1111/j.1745-3984.2008.00068.x

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement*, *47*(3), 286–298. https://doi.org/10.1111/j.1745-3984.2010.00114.x

LaFlair, G. T., Isbell, D., May, L. D. N., Gutierrez Arvizu, M. N., & Jamieson, J. (2015). Equating in small-scale language testing programs. *Language Testing*, *34*(1), 1–18. https://doi.org/10.1177/0265532215620825

Linacre JM, W. B. (2006). *A user's guide to Bigsteps, Winsteps*. *MESA Press*.

Livingston, S. A., & Kim, S. (2008). *Small-Sample Equating by the Circle-Arc Method*. *ETS Research Report Series*. https://doi.org/10.1002/j.2333-8504.2008.tb02125.x

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, *46*(3), 330–343. https://doi.org/10.1111/j.1745-3984.2009.00084.x

Livingston, S. A., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement*, *47*(2), 175–185. https://doi.org/10.1111/j.1745-3984.2010.00107.x

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education*, *33*(10), 1373–1405. https://doi.org/10.1080/09500693.2010.511297

Ozdemir, B. (2017). Equating TIMSS Mathematics Subtests with Nonlinear Equating Methods Using NEAT Design:

Circle-Arc Equating Approaches. *International Journal of Progressive Eduaction*, *13*(2), 116–132.

Pos, L. (2018). UN 2014, Jumlah Paket Soal Ditambah. Retrieved from Acceshttps://www.linggapos.com/14812_un-2014-jumlah-paket-soal-ditambah.html

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33. https://doi.org/doi:10.1515/bile-2015-0008

Sainani, K. L. (2012). Dealing With Non-normal Data. *PM and R*, *4*(12), 1001–1005. https://doi.org/10.1016/j.pmrj.2012.10.013

Shin, M. (2015). *An Investigation of Subtest Score Equating Methods under Classical Test Theory and Item Response Theory Frameworks*. University of Massachusetts.

Sinnema, C., Ludlow, L., & Obinson, V. (2016). Journal of Educational Administration and History. *Journal of Educational Administration and History*, *35*(2). https://doi.org/10.1080/713676155

Skaggs, G. (2005). Accuracy of Random Groups Equating with VerySmall Samples. *Journal of Educational Measurement*, *42*(4), 309–330. https://doi.org/10.1111/j.1745-3984.2005.00018.x

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *33*(8), 1–11. https://doi.org/10.1186/1471-2288-8-33

Tabor, J. (2010). Investigating the Investigative Task: Testing for Skewness An Investigation of Different Test Statistics and their Power to Detect Skewness. *Journal of Statistics Education*, *18*(2), 1–13. https://doi.org/10.1002/jmri.20253

Taylor, J. R. (1997). *An Introduction to Error Analysis: The Studi of Uncertainties in Physical Measurements*. Sauslito: Unversity Science Books.

Treptow, R. S. (1998). Precision and Accuracy in Measurements A Tale of Four Graduated Cylinders. *Journal of Chemical Education*, *75*(8), 1–4. https://doi.org/10.1021/ed075p992

Uysal, İ., & Kilmen, S. (2016). Comparison of Item Response Theory Test Equating Methods for Mixed Format Tests. *International Online Journal of Educational Sciences*, *8*(2), 1–11. https://doi.org/10.15345/iojes.2016.02.001

Zhu, W. (1998). Test equating: What, why, how? *Research Quarterly for Exercise and Sport*, *69*(1), 11–23. https://doi.org/10.1080/02701367.1998.10607662