ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research, volume 295

1st International Conference on Educational Sciences and Teacher Profession (ICETeP 2018)

Equating Method for Small Sample:

Comparative research on nominal weight mean and linear method

Deni Iriyadi, Wardani Rahayu
Educational Research and Evaluation
Universitas Negeri Jakarta
Jakarta, Indonesia
deni.iriyadi@unj.ac.id, Wardani.rahayu@unj.ac.id

Dali S. Naga
Educational Research and Evaluation
Universitas Tarumanagara
Jakarta, Indonesia
dalinaga@gmail.com

***Abstract*—The purpose of this study was to determine the effectiveness of the Nominal Weight Mean Equating and Linear methods on equating using a small sample as a tool for teachers to equate students' scores in the class. This research is included in comparative research comparing two equating methods. The number of samples each replication in this research is 30. The data in the study used the UN results for mathematics subjects in 2015 DKI Jakarta area with the number of anchor items 20% of the total items (30 items). From this data, replication is 50 times for each sample. By spreading the average RMSE for each sample according to replication, the small average RMSE value shows a stable measurement result. Among these methods of securing the class, the most stable is NWME. Thus, as a suggestion for teachers to use the NWME method to equalize scores to avoid students in different classes. Thus, discrimination against students can be prevented especially in determining the completeness of learning or graduation.**

Keywords—equating; small sample; RMSE; nominal weight mean; linear

I. INTRODUCTION

One of the keys to improving the quality of education nationally is by improving the quality of education at the school level. The increase is related to facilities and infrastructure, teacher competency, learning process, and so on. In doing learning in class, many things are of concern to a teacher. One of them conducted an assessment. This is one component in order to improve the quality of education. Efforts to improve the quality of education can be pursued through improving the quality of learning and the quality of assessment. Assessment is a process carried out in order to monitor the learning process and progress of students as evaluation material for future learning improvement. The results of the assessment are presented in the form of numbers and letters as a sign to determine where the students' mastery of a subject matter.

Assessment is a process carried out in order to monitor the learning process and progress of students as evaluation material for future learning improvement. Aside from being an evaluation material, the results of the assessment are used as a benchmark to see the quality of students' education in an educational unit [1]. Often found in one school there are parallel classes taught by two or more same subject teachers. Each teacher has different teaching characteristics, but in giving tests to students, the teacher is only based on the existing grid. This will produce a different test device.

Test equipment based on the same grid is rare and almost never will produce a truly equivalent test device [2]. Preparing the correct parallel test is not easy. Making the same test device will not be perfectly parallel so that their scores cannot be compared directly [3]. A set of 70 test A device is certainly not the same as 70 on test B. This is because the scales of both devices do not have the same scale [4]. For this reason, a method is needed to follow up on this by comparing the scores obtained by the students from the measurements using different test devices which of course come from the same grid. A process is carried out to eliminate discrimination in the form of equalization. This equalization is considered fair enough. Basically, what is done is only to do a general scaling so that scores from various test devices can be compared. After this is done, the checklist of the test A device can be exchanged with the test B. With the equalization of the results of the acquisition of students, it can be arranged nationally for all packages used in the National Examination through the test package equalization process. Thus, there is no discrimination for students because it has been equalized and also allows for the ability mapping between schools in Indonesia.

The process of equalizing the sector is statistically called Equating. Kilmen et al. states that equating is a statistical method that can be used to convert values from different tests with the same constellation [5]. This process is carried out to determine the relationship between two or more tests [6]. Equating is a method that can be used to carry out the equivalent of the test results using statistical and psychometric methods so that they can be compared to provide a general scale so that the test scores equivalent can be exchanged or compared to one another [7–12].

Various equalization methods based on classical methods have been described by several experts. Nonoh in his study compared the Linear method with Equipercentil [13]. Skaggs compares the Linear, Mean, Unsmoothed, and Log-Linear methods [14]; Asiret and Sunbul which compare the Identity, Mean, Linear, Circle Arc, and Presmooted methods [15]; Livingston and Kim compare the Circle Arc and Linear methods [16]; and Babcock, Albano, and Raymond which compare Nominal Weight Mean, Chained, Linear, Circle Arc, Identity and Synthetic [17]. Based on these methods a new

comparison can be made in the hope of providing the best choice for the use of effective equating methods to find a relatively new method that can be used for small samples. In addition to methods, anchors also play a role in equalizing the sector. Babcock has compared the Nominal Weight Mean method with the Linear method but does not consider small samples and the number of anchors. Based on the condition of education in Indonesia with the number of students in general 30 organizations in the same class with the number of questions 30 being one of the reasons in this study.

This study focuses on the comparison of the method of equivalent equalization based on the classical method. In addition, the use of anchor items is also one of the variables that can affect the results of equalization The NWM method uses anchor items while the linear method is not. In addition, the number of samples used also affects the results of equality. Considering that the target of this study is the classroom 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 the problem of discrimination about scores obtained from 2 different test devices. Liner method was developed to answer this problem. In addition, there is also a Nominal Weight Mean method which is basically developed also for the same reason that is to equalize the small sample. Both methods are considered feasible to be compiled. Both are practical and easy to apply for teachers to avoid discrimination in the assessment process in the classroom.

Test based on the same grid is rare and will almost never produce a truly equivalent test [2]. With the different test devices that are used automatically the resulting scoring of students working on the test equipment will not be on the same scale. Thus, we cannot compare their schemes directly without going through the equalization process so that there is justice for students. When different test devices are tested, the results of the test device cannot be directly compared [18]. Thus, a process is needed so that the scores of the two test devices can be compared. For this purpose, a process called equating is carried out.

Equating is a statistical method to make the score of different test results interchangeable by planning based on the same specifications by converting values from different tests but measuring the same construct [5,19,20]. Furthermore, according to Kolen, to be able to do equating, the test equipment must be prepared based on the same content and the same statistical specification [21]. After equated, the obtained seals can be exchanged.

In general, the large number of anchors will reduce the error of equalization [22–25]. The number of anchor grains also affects the results of equalization which is about 20% of the total number of whole items [24,26,27]. Some previous studies provide recommendations for the number of anchor items that can be used stated that around 16.67% to 33.33% of the anchor points [28]. The use of the right number of anchors is expected to provide a good quality of equality and not cause discrimination.

The Linear method will connect the conversion shell to its origin through linear functions [29]. The method is used to measure the same characteristics in the respondent and the check must come from items that have an even level of difficulty. Linear method [linear method] or straight-line method is the method used for a linear related linear, but the form change [transformation] cannot get out of the straight-line range, the too low, high, or extreme scour is usually cut off. The linear method consists of only 2 statistical concepts namely mean and standard deviation. Converted score. The form of transformation is [30–33]:

$$A^*_Y = a[A_X - c] + d \ \omega\eta\varepsilon\rho\varepsilon \ a = \frac{\sigma_{AY}}{\sigma_{AX}}, c = \mu_{AX}, d = \mu_{AY} \qquad (1)$$

From the formula above, it is known that A * Y is the equalization of respondent's X test results to Y test, a is the ratio of standard deviations of X and Y, c is the average of X, and d is the mean of Y. Nominal Weight Mean is a form of linear equalization method. This method is a simple form of the Tucker Method [17]. NWM makes an equalization assumption about score distribution that allows variance and covariance to be dependent on other values [e.g. test lengths] which can be estimated more accurately by using small samples [34]. This method is one simple method aimed at aligning with small samples [35]. The NWM method replaces the covariant terms and variants with the ratio of the number of items in the total test to the anchor text, making the weight effective [36]. The following is the formula for equalizing scores using the NWM method:

$$\ell_A(b) = b - \mu_2(B) + \mu_1(B) + \left[\frac{N(B)K(A) + N(A)K(B)}{[N(A) + N(B)]K(C)}\right][\mu_2(C_B) - \mu_2(C_A)] \qquad (2)$$

From the formula above, $\mu_1(A)$ it is known that the average package A (package 1), $\mu_2(B)$ is the average package B (package 2), $\mu_1(C)$ is the average anchor in package A (package 1), $\mu_2(C)$ is the average anchor in package B (package 2), *N(A)* is the number of respondents working on package A (package 1), *N(B)* is the number of respondents working on package A (package 2), *K(A)* is the number of items test package A (package 1), *K (B)* is the number of items test package B (package 2), and *K(C)* is the number of anchor items.

The use of equalization methods in small samples can be used at the school level, which in general is the number of students in the small sample category. Thus, the teacher as the implementer of the assessment at school can compare the value of students without causing discrimination. The use of small samples cannot be separated from the purpose of this study. The main target is the teacher. Teachers as executors of assessment in class will always be in touch with all forms of assessment, including assessment involving students. The relatively small number of students will certainly affect the results of the equalization carried out. Skaggs conducted a study comparing several equating methods [14]. The comparison using small samples ranging from 25, 50, 75, 100, 150, and 200 samples shows that for the number of samples 25 to 50 gives good results for the method used when viewed from the SEs value. Livingston and Kim examined using small

samples with 10, 25, 50, and 100 [37]. His research showed that samples with ranges from 25 to 50 gave more accurate results. Parshall, Houghton, and Kromrey used samples of 15, 25, 50, and 100 in the equivalent of linear methods [38].

From some of the research results mentioned above, it shows that the number of small samples moving from 25 to 200 even stated that 10 is still a small sample that can be used to score equalization. This study uses a sample number 30 which is still classified as a small sample. Previously explained the reasons for selecting the number of samples and based on several studies this was justified. This study aims to determine the method of equalizing effective scores for use in small samples that are reviewed from the Root Mean Square Error (RMSE). Several studies have been conducted but in terms of using small samples with NWME and Linear methods have not been done. By that, in the study using both methods and testing which method is good for use in small samples.

II. METHOD

The purpose of this research is to find out the most stable alignment method in equalizing small samples. This study used data from 2 junior high school UN package packages from the Education Assessment Center (PUSPENDIK) for DKI Jakarta in 2015 on mathematics subjects. The second place was chosen based on the characteristics of the UN questions in both of them which have similarities in some items (anchor items) according to the research design that has been determined is equalization on a test device that has an anchor item. The dependent variable in this study is the variance of the RMSE value resulting from the equalization using the predetermined score equalization method. Whereas for the independent variable namely the equalization method (Nominal Weight Mean and Linear).

Selecting samples from each population randomly with a random sampling with replacement with the help of SPSS application. Randomization was conducted 50 times with each randomization taking 30 respondents. From the results of the equalization shovel, the RMSE is then determined as a sealing. Thus, each group will have 50 RMSE values with the following formula [17,39]:

$$RMSE(x) = \sqrt{\frac{\sum_{j=1}^{N}\left(\hat{x}_j - x_j\right)^2}{N}} \quad (3)$$

Where N is the number of respondents, $\hat{x}_j$ result of equalization, and x_j equivalent rate. RMSE is used to determine the accuracy of the equalization methods used [15,40]. The mean of the small RMSE shows high accuracy of an equalization method [41]. Furthermore, according to Kartono small mean values indicate better quality of equalization [42].

III. RESULTS AND DISCUSSION

The following table presents the results of the RMSE calculation for each equalization method with a sample size of 30 with 50 replications, the number of items 30, and anchor item 6 (20% of total items).

TABLE I. DESCRIPTIVE ANALYSIS OF RMSE 50 REPLICATION VALUES

Statistics	Root Mean Square Error (RMSE)	
	NWME	***Linear***
Number of data	50	50
Mean	.8473	1.8975
Std. Error of Mean	.10335	.16321
Median	.6166	1.8007
Modus	.33	.04[a]
Standard Deviation	.73083	1.15404
Variance	.534	1.332
Range	3.07	4.19
Minimum	.03	.04
Maximum	3.10	4.23

From the table above shows the value of drinking for each method 0.03 for NWME and 0.04 for Linear with the highest range owned by the Linear method. This shows that the RMSE value for the Linear method is generally of great value. This is supported by a higher variance value indicated by the Linear method. In addition to this in the box-plot for the RMSE values, both methods also show similar things. Figure 1 shows the RMSE value for the NMWE method having the upper wish line that is longer than the bottom wishker line when compared to the RMSE value for the Linear method. This shows that the RMSE results for equalization using the NWME method provide relatively small results. the small RMSE value shows the accuracy of a method.

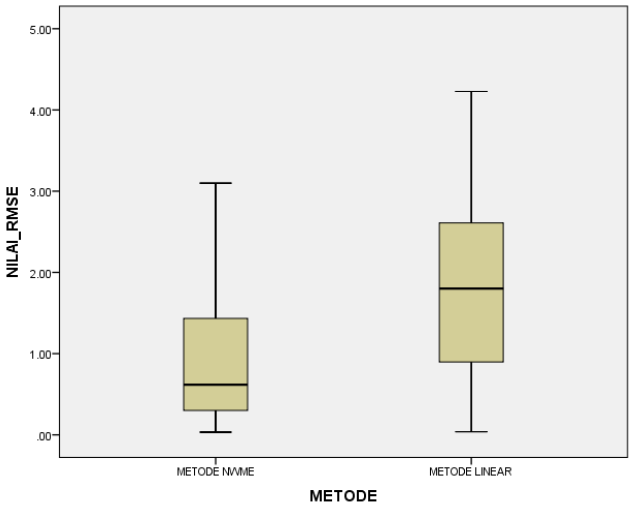

Fig. 1. Boxplot of RMSE values for equalization of NEW and linear methods.

The average RMSE value for equalizing scores using the NWME method is smaller than using the Linear method with a small sample size. RMSE is a method that can be used to see how accurate a method is in making a measurement. RMSE is the difference between the square root value of the predisposing value to the observed value which has meaning

about how much the inequality between actual and predictive values. This RMSE is expected to be of the smallest value.

If the description is descriptive, the mean RMSE value for the NWME method is smaller than the Linear method. This cannot be taken for granted before being tested statistically. Based on the results of the analysis, obtained the value of t arithmetic <0.001 where the value is smaller than the alpha value [0.05]. Thus, both descriptively and inferentially, the average RMSE value for the NWME method is smaller than the Linear method. Livingston states that the mean of the small RMSE shows high accuracy of an equalization method [41]. It was also mentioned by Kartono in his research that small mean values showed better equalization quality [42]. In line with this, it shows that the NWME method is better than the Linear method.

In the Linear method only involves the mean parameters and standard deviations while in the NWME method in addition to these parameters there is also an anchor item parameter. One of the things that cause the NWME method to be better than the Linear method is the presence of anchor items. Anchor items are considered influential in setting scores from 2 value groups. This is in line with the revelation put forward by several researchers namely a large number of anchors will reduce the error of equalization [22–25]. Error from score equalization. The very small number of anchor items with inadequate representation of content can cause equalization problems [24,43]. To get accurate equalization results, the number of anchor items that have enough content in common needs to be considered.

IV. CONCLUSION

Based on the results of the research and discussion above, it can be concluded that the NWM equalization method can be used as an alternative equalization method for the use of small samples. This equalization can be used at the classroom level considering the condition of the number of students belonging to a small sample.

REFERENCES

[1] A.A.P. Antara, and B. Bastari, "Penyetaraan Vertikal Dengan Pendekatan Klasik Dan Item Response Theory Pada Siswa Sekolah Dasar," J Penelit dan Eval Pendidik, vol. 19, no. 1, pp. 13–24, 2015.

[2] R.K. Hambleton, and H. Swaminathan, "Item Response Theory Priciple and Applications," 1985.

[3] N.E. Gronlund, Measurement and Evaluation in Teaching. New York: Macmillan Publishing Company, 1985.

[4] W. Zhu, "Test equating: What, why, how?" Res Q Exerc Sport, vol. 69, no. 1, pp. 11–23, 1998.

[5] S. Kilmen, and N. Demirtasli, "Comparison of Test Equating Methods Based on Item Response Theory According to the Sample Size and Ability Distribution," Procedia-Soc Behav Sci., vol. 46, pp. 1130-134, 2012.

[6] R.K. Hanlbleton, H. Swaminathan, D.J. Rogers, and R.K. Hambleton, "Fundamentals of Item Response Theory Library of Congress Cataloging-in-Publication Data," 1991.

[7] A.P. Herkusumo, "Penyetaraan [Equating] Ujian Akhir Sekolah Berstandar Nasional [UASBN] Dengan Teori Tes Klasik," Junal Pendidik dan Kebud, vol. 17, no. 4, pp. 455–471, 2011.

[8] N.J. Dorans, "Linking scores from multiple health outcome instruments," Qual Life Res., vol. 16, no. SUPPL. 1, pp. 85–94, 2007.

[9] J.S. Kim, and B.A. Hanson, "Test equating under the multiple-choice model," Appl Psychol Meas.. vol. 26, no. 3, pp. 255–270, 2002.

[10] Y. Tong, and M.J. Kolen, "Assessing equating results on different equating criteria," Appl Psychol Meas., vol. 29, no. 6, pp. 418–432, 2005.

[11] D.J. Harris, and M.J. Kolen, "A Comparison of Two Equipercentil Equating Methods for Common Item Equating," Educ Psychol Meas, vol. 50, no. 2, pp. 61–71, 1990.

[12] P. Yin, R.L. Brennan, and M.J. Kolen, "Concordance between ACT and ITED scores from different populations," Appl Psychol Meas, vol. 28, no. 4, pp. 274–289, 2004.

[13] N.S. Aminah, "Karakteristik metode penyetaraan skor tes untuk data dikotomos," J Penelit dan Eval Pendidik, vol. 16, pp. 88–101, 2012.

[14] G. Skaggs, "Accuracy of Random Groups Equating with Very Small Samples," J Educ Meas, vol. 42, no. 4, pp. 309–330, 2005.

[15] S. Aşiret, and S.O. Sünbül, "Investigating test equating methods in small samples through various factors," Kuram ve Uygulamada Egit Bilim., vol. 16, no. 2, pp. 647–668, 2016.

[16] S.A. Livingston, and S. Kim, "Small-Sample Equating by the Circle-Arc Method," ETS Research Report Series, 2008.

[17] B. Babcock, A. Albano, and M. Raymond, "Nominal Weights Mean Equating: A Method for Very Small Samples," Educ Psychol Meas., vol. 72, no. 4, pp. 608–628, 2012.

[18] M. Wu, H.P. Tam, and T.H. Jen, Educational Measurement for Applied Researchers. Singapore: Springer, 2016.

[19] S. Kim, A.A. von Davier, S. Haberman, "an Alternative To Equating With Small Samples in the Non-Equivalent Groups Anchor Test Design," ETS Res Rep Ser, no. 2, pp. 1-40, 2006.

[20] B. Ozdemir, "Equating TIMSS Mathematics Subtests with Nonlinear Equating Methods Using NEAT Design: Circle-Arc Equating Approaches," Int J Progress Educ., vol. 13, no. 2, pp. 116–132, 2017.

[21] M.J. Kolen, "Linking assessments: Concept and history," Appl Psychol Meas., vol. 28, no. 4, pp. 219–226, 2004.

[22] D. Budescu, "Efficiency of Linear Equating As a Function of the Length of the Anchor Test," J Educ Meas, vol. 22, no. 1, pp. 13–20, 1985.

[23] M. Wingersky and L. Cook, "Specifying the Characteristics of Linking Items Used for Item Response Theory Item Calibration1, 2" Educational Testing Service, 1987.

[24] M.J. Kolen, and R.L. Brennan, Test Equating, Scaling, and Linking. 2nd ed. New York: Springer, 2004.

[25] K.L. Ricker, and D.A.A. Von, "The Impact of Anchor Test Length on Equating Results in a Nonequivalent Groups Design," ETS Research Report Series, 2007.

[26] L. Crocker, and J. Algina, "Introduction to Classical and Modern Test Theory," 2006.

[27] W.H. Angoff, Norms Scale and Equivalent Scores. New Jersey: Educational Testing Service, 1984, pp. 508.

[28] B.D. Wright, and M.H. Stone, Best Test Design. Chicago: Mesa Press, 1979.

[29] D.S. Naga, Pengantar Teori Sekor Pada Pengukuran Pendidikan. Jakarta: Gunadarma, 1992.

[30] M.J. Kolen, and R.L. Brennan, "Linear Equating Models for the Common-item Nonequivalent-Populations Design," Appl Psychol Meas, vol. 11, no. 3, pp. 263–277, 1987.

[31] A.A. Mroch, Y. Suh, M.T. Kane, and D.R. Ripkey, "An Evaluation of Five Linear Equating Methods for the NEAT Design," Meas Interdiscip Res Perspect, vol. 7, no. 3–4, pp. 174–193, 2009.

[32] M.T. Kane, A.A. Mroch, Y. Suh, and D.R. Ripkey, "Linear equating for the neat design: A rejoinder and some further comments," Measurement, vol. 8, no. 1, pp. 27–37, 2010.

[33] M. Wiberg, and W.J. van der Linden, "Local linear observed-score equating," J Educ Meas, vol. 48, no. 3, pp. 229–254, 2011.

[34] A.C. Dwyer, "Maintaining Equivalent Cut Scores for Small Sample Test Forms," J Educ Meas, vol. 53, no. 1, pp. 3–22, 2016.

[35] G.T. LaFlair, D. Isbell, L.D.N. May, M.N.G. Arvizu, and J. Jamieson, "Equating in small-scale language testing programs," Lang Test, vol. 34, no. 1, pp. 127–144, 2017.

[36] S. Caglak, "Comparison of Several Small Sample Equating Methods under the NEAT Design," Turkish J Educ, vol. 5, no. 3, pp. 96, 2016.

[37] S. Kim and S.A. Livingston, "Comparisons among small sample equating methods in a common-item design," J Educ Meas., vol. 47, no. 3, pp. 286–298, 2010.

[38] C.G. Parshall, P.D.B. Houghton, abd J.D. Kromrey, "Equating Error and Statistical Bias in Small Sample Linear Equating," J Educ Meas., vol. 32, no, 1, pp. 37–54, 1995.

[39] L.W. Klien, and D. Jarjoura, "The Importance of Content Representation for Common Item Equating With Nonrandom Groups," J Educ Meas., vol. 22, no. 3, pp. 197–206, 1985.

[40] I. Uysal, and S. Kilmen, "Comparison of Item Response Theory Test Equating Methods for Mixed Format Tests," Int Online J Educ Sci, vol. 8, no. 2, pp. 1–11, 2016.

[41] S.A. Livingston, "Small-Sample Equating With Log-Linear Smoothing. J Educ Meas.," vol. 30, no. 1, pp. 23–39, 1993.

[42] Karton, "Equating the Combined Dichotomous and Polytomous Item Test Model in an Achievement Test," J Penelit dan Eval Pendidik, vol. 12, no. 2, pp. 302–320, 2008.

[43] N.S. Petersen, L.L. Cook, and M.L. Stocking, "IRT Versus Conventional Equatig Method: A Comparative Study of Scale Stability," J Educ Stat, vol. 8, no. 2, pp. 137–156, 1983.