JISAE
Journal of Indonesian Student Assessment and Evaluation

https://doi.org/10.21009/JISAE

JISAE (Journal of Indonesian Student Assessment and Evaluation)
ISSN : P-ISSN: 2442-4919 | E-ISSN: 2597-8934
Website : http://journal.unj.ac.id/unj/index.php/jisae

Vol 8 No 1 (2022)

EQUATING METHOD FOR LEARNING OUTCOMES OF ELEMENTARY SCHOOL/MADRASAH STUDENTS

Deni Iriyadi[1],
Universitas Islam Negeri Sultan Maulana Hasanuddin Banten

Address for Correspondence:
1deni.iriyadi@uinbanten.ac.id

ABSTRACT

This study aims to determine the method of equalizing a good score on a small number of items often found at the elementary/madrasah level. This research is a simulation study that compares capabilities in terms of three distributions. The sample size used in this study was 50 responses for the distribution of the ability of each group (average, positive skewness, and negative skewness). Replication was carried out in 50 for each ability group using the help of Wingen3. The root means the squared error is the indicator used to evaluate the equation results. The results showed that groups with the same ability skewness distribution (normal ability distribution, normal ability distribution, positive skewness ability distribution, and negative skewness ability distribution- negative skewness ability distribution - negative skewness ability distribution) would give a lower RMSE score than groups with different ability distributions. A low RMSE value indicates that the error of the measurement results is low.

Keywords: Equating, Nominal Weight Means, Ability Distribution

INTRODUCTION

Assessment is a process that is carried out to monitor students' process and learning progress as evaluation material for future learning improvements. The assessment results are presented in the form of numbers and letters as a sign to determine where the student's mastery of a subject matter is. The assessment carried out by the teacher cannot be separated from the measuring instrument in the form of a test. The tool is packaged in the form of questions made based on the grid. Both the teacher and the government do this. Even though using the same grid is rarely found as a truly equivalent test device. Compiling truly parallel tests is not easy. Making the same test device will not perfectly parallel each other so that their scores cannot be compared directly (Gronlund, 1985). From the existing grid, it becomes the primary reference in arranging each question in different schools and regions. Often found in one school, there are parallel classes taught by two or more teachers of the same subject. 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.

When two values come from 2 different test devices, both values cannot be exchanged. This is because the scores of the two devices do not have the same scale (Zhu, 1998). By equating the scores obtained by students, it can be compared. Thus, there is no discrimination

for students because it has been equated, and it is also possible to carry out mapping of capabilities between schools in Indonesia.

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, 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 grid. Of course, it is 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 appropriate to be used according to the class's characteristics, especially for the number of students. The use of undersized samples is also based on the need that the number of students belongs to a small sample in the classroom.

The process of equating the score is statistically called Equating. Equating is a solution to this problem. Kilmer states that equating is a statistical method used to convert values from different tests with the same construct (Kilmen & Demirtasli, 2012). This process determines the relationship between two or more tests (Hambleton & Swaminathan, 1985). Various kinds of equalization methods have been applied. Some of them are based on the classical method, which is known as more practical to apply.

Many equating methods have been developed. These methods are based on the needs of the world of education. Some are based on classical theories, and some are based on modern theories. Each of them has its advantages. Equating using classical methods is more familiar, rational, and easy to apply (Yin, Brennan, & Kolen, 2004).

Nominal Weight Mean

Some equating methods are developed based on the existing equating method. The method is made as a form of improvement on the weaknesses of the previous method. One of them is the Nominal Weight Mean (NWM) method. This method is a form of linear equating development method from the Tucker Method, intended to equate with small samples (Babcock, Albano, & Raymond, 2012; LaFlair, Isbell, May, Arvizu, & Jamieson, 2017). In this method, the synthetic standard deviation X and Y values are assumed to be the same ($\sigma s(X) = \sigma s(Y)$); thus, this makes the formula for Linear Equating Tucker (Babcock et al., 2012).

(1)
to be $$l_Y(X) = \frac{\sigma_S(X)}{\sigma_S(Y)}[X - \mu_S(Y)] + \mu_S(X)$$

$ly(X) = [X - \mu s(Y)] + \mu s(X)$ (2)

Index ly(X) shows the equalization results for the Tucker and X methods as unit X test unit scores which will be equalized where $\mu s(X)$ and $\mu s(Y)$ is a synthetic average. This simplification process is carried out because synthetic standard deviation cannot be estimated accurately when using small samples. Babcock et al. (2012) explain the synthetic mean in equation (2) as follows:

(3)
and $$\mu_s(X) = \mu(X) + w_1 \frac{Cov\,(X,Z_X)}{Var(Z_x)}[\mu(Z_Y) - \mu(Z_X)]$$
(4)
$$\mu_s(Y) = \mu(Y) + w_2 \frac{Cov\,(Y,Z_Y)}{Var(Z_Y)}[\mu(Z_Y) - \mu(Z_X)]$$

This method assumes that the variance and covariance values are replaced with other values that can be estimated more accurately by using small samples (Dwyer, 2016). Variance and covariation values cannot function properly when the sample used is small (Babcock et al., 2012).

The form of the covariance between the X test device scores and the anchor item score is [cov(X, Zx)/var(Zx)] where X is the total score, and Zx is the anchor score found on the X test device (Babcock et al., 2012), then the covariance between X and Zx is,

$$cov(X, Z_X) = \frac{\sum_{p=1}^{N}\left(X_p - \mu(X)\right)\left(Z_{xp} - \mu(Z_x)\right)}{N-1}$$
$$= \frac{\sum_{p=1}^{N}(X_p Z_{xp}) - \sum_{p=1}^{N}(\mu(X) Z_{xp}) - \sum_{p=1}^{N}\left(X_p \mu(Z_x)\right) + \sum_{p=1}^{N}(\mu(X)\mu(Z_x))}{N-1}$$

N is the number of samples/respondents. Assuming that X and Zx are expressed as deviation scores Xi - μ(X) dan Zi - μ(Zx), then the above equation can be simplified to become,

$$cov(X, Z_X) = \frac{\sum_{p=1}^{N}\left(X_p - \mu(X)\right)\left(Z_{xp} - \mu(Z_x)\right)}{N-1}$$
$$= \frac{\sum_{p=1}^{N}(X_p)\,(Z_{xp})}{N-1}$$
$$= \frac{\sum_{p=1}^{N}\left(\sum_{i=1}^{K(X)} X_{pi}\right)\left(\sum_{j=1}^{K(Z)} Z_{x(pj)}\right)}{N-1}$$

Where Xip is the unit score on X for the p-respondent and the item i with K is the number of items on X so that the Xip value is equal to Xp multiplied by the number of items K.

$$\sum_{i=1}^{K(X)} X_{ip} = \frac{\sum_{i=1}^{K} X_i}{K}[K(X)]$$
$$= [\mu(X_p)][K(X)]$$

So Zip can be written in the form K(X)μ(Xp). The same thing applies to anchor items (Zx(pi)). Thus, the equation above can be written as:

$$= \frac{\sum_{p=1}^{N}\left(K(X)\mu(X_p)K(Z_p)\mu(Z_p)\right)}{N-1} \qquad (5)$$

X and Zx are the respondent's scores of each X test device and anchor items on the X test device, while K is the number of items from each group. K is substituted because the total score for each respondent equals the average score multiplied by the number of items. For the variance of Zx as follows:

$$var(Z_x) = \frac{\sum_{p=1}^{N}\left(Z_{xp} - \mu(Z_x)\right)^2}{N-1}$$

by assuming a deviation score Zi. - μ(Zx), so that the above equation can be written into,

$$var(Z_x) = \frac{\sum_{p=1}^{N}(Z_{xp} - \mu(Z_x))(Z_{xp} - \mu(Z_x))}{N-1}$$
$$= \frac{\sum_{p=1}^{N}(Z_{xp})(Z_{xp})}{N-1}$$
$$= \frac{\sum_{p=1}^{N}\left(\sum_{j=1}^{K(Z)} Z_{x(pj)} \sum_{j=1}^{K(Z)} Z_{x(pj)}\right)}{N-1}$$

Where Zip is the unit score on Z for the p respondent and item j with K is the number of items in Z so that the value of Zpj is equal to the mean of Zp multiplied by the number of items K.

$$\sum_{i=1}^{K(Z)} Z_{ip} = \frac{\sum_{i=1}^{K} Z_i}{K}[K(Z)]$$
$$= [\mu(Z_x)][K(Z)]$$

So the Zip can be written in the form K(Z)μ(Zx). Thus, the equation above can be written as:

$$= \frac{\sum_{p=1}^{N}(K(Z_x)\mu(Z_x))(K(Z_x)\mu(Z_x))}{N-1}$$
$$= \frac{\sum_{p=1}^{N}(K(Z_x)\mu(Z_x))^2}{N-1} \quad (6)$$

substitute equations (8) and (9) into equations below obtained results,

$$\frac{cov(X,Z_x)}{var(Z_x)} = \frac{\frac{\sum_{p=1}^{N}(K(X)\mu(X_p)K(Z_x)\mu(Z_{xp}))}{N-1}}{\frac{\sum_{p=1}^{N}(K(Z_x)\mu(Z_{xp}))^2}{N-1}}$$
$$= \frac{K(X)K(Z_x)\sum_{p=1}^{N}(\mu(X_p)\mu(Z_{xp}))}{K(Z_x)^2\sum_{p=1}^{N}(\mu(Z_{xp}))^2}$$

Assuming the average of the total items X for each respondent is equal to the average of the Zx items, then the equation above will be,

$$= \frac{K(X)K(Z_x)\sum_{p=1}^{N}(\mu(Z_{xp})\mu(Z_{xp}))}{K(Z_x)^2\sum_{p=1}^{N}(\mu(Z_{xp}))^2}$$
$$= \frac{K(X)K(Z_x)\sum_{p=1}^{N}(\mu(Z_{xp})\mu(Z_{xp}))}{K(Z_x)K(Z_x)\sum_{p=1}^{N}\mu(Z_{xp})\mu(Z_{xp})}$$
$$\frac{cov(X,Z_x)}{var(Z_x)} = \frac{K(X)}{K(Z_x)} \quad (7)$$

Using equation (7), the synthetic mean (3) and (4) becomes:

$$\mu_s(X) = \mu(X) + w_1\frac{K(X)}{K(Z_x)}[\mu(Z_Y) - \mu(Z_X)] \quad (8)$$

and

$$\mu_s(Y) = \mu(Y) + w_2\frac{K(Y)}{K(Z_Y)}[\mu(Z_Y) - \mu(Z_X)] \quad (9)$$

Thus the nominal weight replaces the variance and covariance in the Tucker method in the form of total items and anchor items to be the number of respondents (Caglak, 2016), where K indicates the number of items on the test device. While w is related to the number of samples/respondents (N), which is the ratio of the number of samples from X and Y to the total number of samples (Babcock et al., 2012; Caglak, 2016)

$$w_1 = \frac{N(Y)}{N(X)+N(Y)} \quad (10)$$

and

$$w_2 = \frac{N(X)}{N(X)+N(Y)} \quad (11)$$

By substituting equations (8) and (9) to equation (2) is obtained,

$$l_Y(X) = X - \mu_s(Y) + \mu_s(X)$$
$$= X - \left[\mu(Y) + w_2\frac{K(Y)}{K(Z_Y)}[\mu(Z_Y) - \mu(Z_X)]\right] + \left[\mu(X) + w_1\frac{K(X)}{K(Z_x)}[\mu(Z_Y) - \mu(Z_X)]\right]$$
$$= X - \mu(Y) + \mu(X) + \left[w_1\frac{K(X)}{K(Z_x)} + w_2\frac{K(Y)}{K(Z_Y)}\right][\mu(Z_Y) - \mu(Z_X)]$$

Moreover, using the weights of the number of samples in equations (12) and (13), the above equation becomes:

$$l_Y(X) = X - \mu_s(Y) + \mu_s(X)$$
$$= X - \left[\mu(Y) + w_2\frac{K(Y)}{K(Z_Y)}[\mu(Z_Y) - \mu(Z_X)]\right] + \left[\mu(X) + w_1\frac{K(X)}{K(Z_x)}[\mu(Z_Y) - \mu(Z_X)]\right]$$
$$= X - \mu(Y) + \mu(X) + \left[w_1\frac{K(X)}{K(Z_x)} + w_2\frac{K(Y)}{K(Z_Y)}\right][\mu(Z_Y) - \mu(Z_X)]$$

so the equation for the Nominal Weight Mean Equating method will be obtained as follows:

$$Y^*_{NWME} = X - \mu(Y) + \mu(X) + \left[\frac{N(Y)K(X) + N(X)K(Y)}{[N(X) + N(Y)]K(Z)}\right][\mu(Z_Y) - \mu(Z_X)] \quad (12)$$

Y*NWME shows the result of equalizing the score from the Y test device to the X test device. All values contained in the X test device, when substituted in equation (12), the value of the Y test device equalization will be obtained.

Small Sample

Following the explanation from Naiman, Rosenfeld, and Zirkel, which states that the sample size will always affect the calculation results. The number of samples is related to errors that lead to the equalization results. There are two types of errors, namely random errors and systematic errors. Random errors related to the number of samples used, the greater the number of samples, the smaller the random errors generated will be more minor.

METHOD

This research uses simulation data to consider that when making comparisons using several factors determined (the form of distribution of capabilities), using actual data will have several complex problems. When using simulation data, it will be straightforward to condition what has been and will be tested, and the data can represent actual data in the field (Harris & Crouse, 1993; Holland, Davier, Sinharay, & Han, 2006).

Data is generated based on the 3PL model using the WinGen3 program. The program is specifically for generating data on the single model item responses of dichotomous and polytomous, mixed models for several models and several conditions following actual conditions in practice. The WinGen3 Program can generate item response data with values of grain parameters and capabilities for various distributions that correspond to the distribution of actual data (Han & Hambelton, 2014). In this data generation, 50 replication conditions are carried out.

From the results of the equalization score, RMSE was determined to score. Thus, each group will have 30 RMSE values with the following formula (Babcock et al., 2012; Klien & Jarjoura, 1985):

$$RMSE(x) = \sqrt{\frac{\sum_{i=1}^{M}(\hat{x}_i - x_i)^2}{M}}$$

Where M is the number of respondents, the score is equated, score xi score is equalized. RMSE is used to determine the accuracy of the equalization method used (Aşiret & Sünbül, 2016; Uysal & Kilmen, 2016). The mean of a small RMSE shows the high accuracy of an equalization method (Livingston, 1993). Furthermore, according to Karton (2008), the small mean value of RMSE shows a better quality of equalization.

The form of ability distribution is divided into three, namely: (1) normal distribution, (2) favorable skewness distribution, and (3) negative skewness distribution. For the normal distribution using the criteria N (0.1), the favorable skewness distribution uses the criteria (-1.1), and the negative skewness distribution uses the criteria (1,1). Whereas for the number of items using 30 with an anchor item proportion of 20% of the total item.

RESULTS AND DISCUSSION

Table 1. Data Descriptions

Group	Mean	SD
N-N	0,6	0,09
SP-SP	0,7	0,15
SN-SN	0,6	0,15
N-SP	1,7	0,73
N-SN	1,5	0,69
SP-SN	1,3	1,02

The table above shows a description of the sample statistics generated using Wingen. The results of analysts that have been done using the method of nominal weight mean and with a different distribution of abilities in small samples obtained RMSE values of 50 replications carried out using the WinGen3 program. The number of questions used is 30 items with a proportion of 20% anchor (Angoff, 1984; Crocker & Algina, 2008; Hamlbleton, Swaminathan, Rogers, & Hambleton, 1991; Kolen & Brennan, 2004; Wright & Stone, 1979).

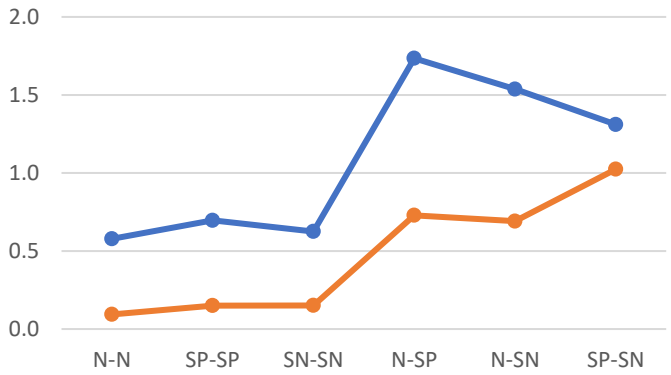

Figure 1. Graph of RMSE Mean Differences Equivalent Couples Group Ability

The graph above shows the average value of the RMSE for each equating that has been carried out for each combination of variations in the ability distribution. It shows that the lowest average value is in the N-N combination group, and the highest average is in the N-SP combination group.

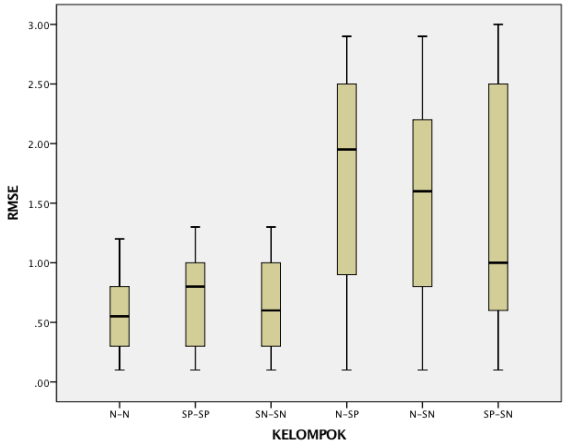

Figure 2. Boxplot of RMSE Value Equivalent Couples Group Ability

Research shows that the average RMSE value for couples with the same distribution is typically distributed with normal distribution, skewness is positively distributed with fair skewness distribution, and negative skewness distribution with negative skewness distribution is lower than ability pairs who have different ability distributions. The opinion of several experts supports this, one of which states that the similarity in the form of the initial distribution of the two test devices compared provides accurate equalization results as explained by Zhu (1998) that the form of the score distribution of the two test devices must be the same Likewise from the variance values also show different results when the pairs of groups have the same characteristics. Minor variants will be obtained for couples of abilities that are equally distributed while those with different distributions have more significant variance. This shows that with a smaller average and a smaller variance value, the consistency of the RMSE value generated from some replications is perfect. Thus, for all replications, it produces a small RMSE value. This is different from the ability pairs with different distributions and a considerable average value. Visually, this can be seen in Figures 1 and 2 above. The same thing is shown by Kilmen & Demirtasli (2012), which shows that the results are accurate when the distribution of the two groups is the same. Besides that, in other studies, Uysal & Kilmen (2016) conducted research on the distribution of capabilities divided into three: normal distribution, favorable skewness distribution, and negative skewness distribution. The results showed that groups that had the same ability distribution (a normal distribution with normal distribution, favorable skewness distribution with positive skewness distribution, and negative skewness distribution with negative skewness distribution) produced a low Equating Error when compared to groups that had a distribution of abilities that different from each other. In addition, the similarities in the form of the initial distribution of the two test devices compared provide accurate equalization results.

In Figure 2 above, ability group pairs with the same distribution appear to have the upper whisker line, which is longer than the lower whisker line. In other words, the results of the RMSE for equalization using the NWM method on the same initial ability distribution are those that are generally distributed with normal distribution, skewness is positively distributed with a favorable distribution of skewness, and distribution of negative skewness with an unfavorable distribution of skewness has the same characteristics. All three have a whisker line, the upper part of which is longer than the bottom. In addition to the value of drinking, all three have similarities. Based on this, it can be concluded that the RMSE value for all three generally yields relatively tiny results. Unlike the ability group partners who have unequal distribution of abilities, the resulting range is quite extensive, and the value of drinking is relatively more excellent.

Zhu (1998) explained that the distribution of scores from the two test kits must be the same. This can support the equalization results rather than the forms of distribution of the two different factors. Besides that, Muraki, Hombo, & Lee (2000) stated that the equate of scores using the classical method, one of the things that can support the accuracy of the equating results, namely the distribution of scores in both groups of abilities, must be the same even though the mean and standard deviation is different. The same is done by Toni, which shows that the equalization results will be good when the raw score distribution is the same. Flanagan said that scores of two or more test devices could be compared or matched when they have identical distributions (Kolen, 2004). (Masse, Allen, Wilson, & Williams, 2006) say that some assumptions for getting a good equalization result are the distribution of scores between two scales or test kits. The accuracy value is measured from a small RMSE value.

The most commonly used model is the linear equation model in equalizing scores. However, this method assumes that in the target population, the distribution of scores on the X test device and on the Y test device only differs in mean and standard deviation (does not take into account the ability distribution) (Davier, 2007). Reflecting on this, the assumption

is difficult to accept, given that the preparation of the test kit is only guided by one standard grid. When the test forms differ in difficulty, the equalization relationships between them are usually not linear. Nonlinear methods are used when assuming the difficulty level between X test devices and Y test devices is different (Albano, 2015).

CONCLUSION

Based on the research results, it can be concluded that the method of equalizing nominal weight means can be used as an alternative method of equalization for the use of small samples. This type of equalization with this method is relatively easy to use, considering that the method is part of the classical method. In addition, a well-implemented application in small samples also provides added value for use. This equalization can be used at the class level considering the condition of the number of students belonging to the small sample and seeing the similarities in the distribution of students' ability values. Thus, teachers no longer make students "victims" of the same inequality as some developed test kits. No more discrimination against students to determine their graduation.

REFERENCE

Albano, A. D. 2015. A General Linear Method for EquatingWith Small Samples. *Journal of Educational Measurement*, *52*(1), 55–69.

Angoff, W. H. 1984. *Scales, Norms, and Equivalent Scores*. New Jersey: Educational Testing Service. https://doi.org/10.1063/1.3131262

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. https://doi.org/10.19128/turje.16916

Crocker, L., & Alglna, J. 2008. *Introduction to Classical and Modern Test Theory*. (M. Stranz, Ed.), *Harcourt Brace Jovanovich College Publishers*. USA: Cengage Learning.

Davier, A. A. von. 2007. New Results on the Linear Equating Methods for the Non-Equivalent-Groups Design. *Journal of Educational and Behavioral Statistics*, *33*(2), 186–203. https://doi.org/10.3102/1076998607302633

Dorans, N. J. 2007. Linking scores from multiple health outcome instruments. *Quality of Life Research*, *16*(SUPPL. 1), 85–94. https://doi.org/10.1007/s11136-006-9155-3

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

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

Hambleton, R. K., & Swaminathan, H. 1985. *Item Responde Theory Principle and Application*. New York: Springer.

Han, K. T., & Hambelton, R. K. 2014. *User 's Manual : WInGen3*.

Harris, D. J., & Crouse, J. D. 1993. A Study of Criteria Used in Equating. *Applied Measurement in Education*, *6*(3), 195–240. https://doi.org/10.1207/s15324818ame0603_3

Holland, P. W., Davier, A. A. Von, Sinharay, S., & Han, N. 2006. *Testing the Untestable Assumptions of the Chain and Poststratification Equating Methods for the NEAT Design*. https://doi.org/10.1002/j.2333-8504.2006.tb02023.x

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

Kilmen, S., & Demirtasli, N. 2012. Comparison of Test Equating Methods Based on Item Response Theory According to the Sample Size and Ability Distribution. *Procedia - Social and Behavioral Sciences*, *46*(1980), 130–134. https://doi.org/10.1016/j.sbspro.2012.05.081

Kim, J. S., & Hanson, B. A. 2002. Test equating under the multiple-choice model. *Applied Psychological Measurement*, *26*(3), 255–270. https://doi.org/10.1177/0146621602026003002

Klien, L. W., & Jarjoura, D. 1985. The Importance of Content Representation for Common Item Equating With Nonrandom Groups. *Journal of Educational Measurement*, *22*(3), 197–206. https://doi.org/10.1111/j.1745-3984.1985.tb01058.x

Kolen, M. J. 2004. Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*(4), 219–226. https://doi.org/10.1177/0146621604265030

Kolen, M. J., & Brennan, R. L. 2004. *Test Equating, Scaling, and Linking* (2nd ed.). New York: Springer.

Kolen, M. J., & Brennan, R. L. 2014. *Test Equating, Scaling, and Linking* (Third Edit). New York: Spinger.

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

Livingston, S. A. 1993. Small-Sample Equating With Log-Linear Smoothing. *Journal of Educational Measurement*, *30*(1), 23–39. https://doi.org/10.1111/j.1745-3984.1993.tb00420.x

Masse, L. C., Allen, D., Wilson, M., & Williams, G. 2006. Introducing equating methodologies to compare test scores from two different self-regulation scales. *Health Education Research*, *21*(August), 110–120. https://doi.org/10.1093/her/cyl088

Muraki, E., Hombo, C. M., & Lee, Y. W. 2000. Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*(4), 325–337. https://doi.org/10.1177/01466210022031787

Tong, Y., & Kolen, M. J. 2005. Assessing equating results on different equating criteria. *Applied Psychological Measurement*, *29*(6), 418–432. https://doi.org/10.1177/0146621606280071

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

Wright, B. D., & Stone, M. H. 1979. *Best Test Design*. Chicago: Mesa Press. https://doi.org/10.1016/B978-0-12-238180-5.50013-6

Yin, P., Brennan, R. L., & Kolen, M. J. 2004. Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement*, *28*(4), 274–289. https://doi.org/10.1177/0146621604265034

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