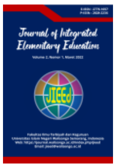

Journal of Integrated Elementary Education
ISSN 2776-1657 (online) 2828-223X (printed)
Volume 2, Number 2, September 2022, Page 117-129
Website: https://journal.walisongo.ac.id/index.php/jieed

Linear and Separate Calibration Methods of Equating Continuous Assessment Scores of Public and Private Elementary Schools

Nofisat Adeola Olanigan,* Alaba Adeyemi Adediwura, Olawale Ayoola Ogunsanmi
Obafemi Awolowo University, Ile-Ife
*Correspondence author: olaniadeola2@gmail.com

DOI : https://doi.org/10.21580/jieed.v2i2.11677
Received: 2022-05-17, Revised: 2022-09-14,
Accepted: 2022-09-14, Published: 2022-09-15

Abstract

This study determined the mean ability estimates of students in public and private secondary schools when their scores are equated through linear equating, and it examined which of the equating methods is more efficient. This study adopted a descriptive research design. The population for the study comprised 24,874 candidates that registered and sat for the 2016 June/July National Examinations Council (NECO) mathematics examination in Ogun State. A sample of 1139 candidates were selected from both public and private schools using the multi-stage sampling procedure. The research instruments used for the study were secondary data sources. The data was analyzed using Multidimensional Item Response Theory (MIRT), Separate and Linear Equation. The results showed that private school candidates' ability score ($\bar{x}$ = -0.001, SD = 0.961), ($\bar{x}$= -0.001, SD = 0.961) was higher than public school candidates' ability score ($\bar{x}$= -0.865, SD = 1.058), ($\bar{x}$= -0.626, SD = 0.970) when equated using separate calibration and linear equating methods respectively and that the difference observed in the ability estimate of examinees from public school and private school were significant (t 1138 = - 14.431, p < 0.05) and (t 1138 = -10.876, p < 0.05) when their scores were equated with each of the equating methods. However, the results showed that the linear equating method was more efficient.

Keywords: Continuous Assessment, Linear Equating, Separate Calibration, Test Equating Methods

INTRODUCTION

Continuous assessment practice in Nigeria requires that teachers generate students' continuous assessment scores (CAS). The non-uniformity of scores allotted to CA attest to the fact that there is inherent problem of comparability of standard. The differences may be in the quality of teachers across schools or in the quality of instructions and assessment which could result in non-uniformity of students' continuous assessment scores that vary from teacher to teacher and from school to school. The assessment result, in which test scores are paramount, has a consequential effect on students' future. Therefore, the need for the standardization and placing of test scores on common scale to enhance fairness and comparablity is as important as given the test itself. Different approaches can be used in ensuring that test scores are standardized and placed on a common scale, and one of such approach which has not been properly explored in Nigeria is test score equating.

Research has shown that the differences in the quality of tests and other assessment instruments used in different schools as well as differences in the procedures of scoring and grading the various assessments in the various schools could pose problem of comparability of standard, as some teachers may set apparently difficult test items, which students may see as a threat to the class while some teachers may set easy items or unduly inflate continuous assessment scores of the students to favour their schools. This inherent problem has bedeviled continuous assessment practices in Nigeria. There are so many problems that have been identified as bottlenecks in the implementation of continuous assessment practice in Nigerian schools. They include: Comparability of standard, problem of record keeping, continuity of records, problem of cheating, and misconception of the concept of continuous assessment (Jones et. al, 2022; Schalet et. al, 2021; Emaikwu, 2004; Onjewu, 2007; Akin-Arikan & Gelbal, 2021).

Comparability research conducted by Makiney, Rosen and Davis, (2003); Pinsoneault, (1996); focused on the differences in means and standard deviations of test scores. The above authors placed little emphasis on underlying measurement issues like item parameters. Raju, Laffitte, and Byrne (2002) stated that "without measurement equivalence, it is difficult to interpret observed mean score differences meaningfully." It is therefore imperative not to only use mean and standard deviation in the comparability of student CA in Nigeria but also include other measurement issues such as item discrimination, difficulty and guessing parameters. As a result, enhancement strategies such as moderation, self-assessment and test scores equating have been suggested as educational standards control mechanisms in Nigeria (Afemikhe, 2007). Moderation and to some extent self-assessment has been practiced especially at the tertiary level and not at the secondary level of education in Nigeria, whereas test score equating is not practiced at all. It is therefore pertinent to carry out a study on test score equating in Nigeria using CA scores from teacher-made-tests and researcher's mathematics achievement test score.

Comparability of standard of continuous assessment had been one of the major problems being identified by scholars, stakeholders, since inception of 9-3-3-4 system of education. This problem seems to arise as a result of difference in personnel and the practice of continuous assessment. Other problems associated with this may include examination malpractices, do-or-die syndrome in examinations, too much of paper qualification, lack of funding, attitude of students, parents and teachers towards continuous assessment, teachers' integrity, lack of commitment, quality of assessment instruments, inconsistency in instrument administration, categories of schools, differences in procedures of scoring and grading and collation of continuous assessment grade. Some conceptual solution had been suggested regarding the issues of comparability of standard in CA, but these seem not to be working. This is probably due to the fact they are not based on solid statistical background. There is therefore the need to examine how scores from multiple assessment standards can be made comparable using a statistical approach (test score equating methods).

The statistical process of making test scores comparable is called test equating (Kolen & Brennan, 2004). Test equating is a process used in comparing the test scores of more than one test form administered to examinees or group of examinees. Test equating also refers to the statistical process of determining comparable scores on different forms of an exam. It can be accomplished using either Classical Test Theory (CTT) or Item Response Theory (IRT). Test equating was seen as a measurement process that involves test development, administration, analysis, scoring, reporting, and evaluation (Hattie, Jaeger & Bond, 1999).

In the conduct of test scores equating, numerous methods were available to the researcher which has also been used by scholars in equating continuous assessment but these methods may not be of equal efficacy. Some of the methods of equating continuous assessment used by these scholars' includes; Mean Equating, Linear equating, Levine equally reliable linear equating, Tucker linear equating, Chained linear equating, Equi-percentile equating , Frequency estimation equi-percentile equating, Chained equi-percentile equating, One parameter logistic (Rasch) model equating Concurrent calibration, Fixed based procedure, Equating constant procedure. However, this study is centered on Linear and Separate Calibration Equating Test methods.

Linear equating assumes that, apart from differences in means and standard deviations, score distributions on two forms of a test are the same, this follows difficulty differences to vary along the score scale of the two assessments. Given this assumption, scores on the two forms can be matched using their z scores. The linear conversion is defined in terms of the mean and Standard deviation of the two scores (X and Y).

Linear equating is sometimes called linear conversion. In this case, A is referred to as the slope of the linear conversion and B the intercept. It allows the relative difficulty of two or more forms of tests to vary along the score scale. When the standard deviation are equal, linear equating becomes the same as mean equating described by Kolen and Brennan (1995). Linear equating is appropriate if the score distributions differ only in mean and standard deviation. If there are differences beyond these first two moments, or if the shape of the distributions differ, then linear equating is inappropriate. If multiple scales are aggregated to form a composite or total score, the score range might be large enough to permit linear equating. However, caution is needed when assuming that score distributions will be reasonably stable over multiple assessment populations. Large score distribution differences can invalidate a linear equating because the equating transformation would differ for another dataset (Eiji, Catherine &Yong Won, 2000).

Separate calibration methods can be examined under two headings: moment methods (mean-mean, mean-sigma and robust mean-sigma) and characteristic curve methods. The first two are moment's methods, which are attractive because of their statistical simplicity. The latter two are characteristics curves methods, which use more available information from item parameters and thus are expected to produce more adequate scaling results (González, J., & Gempp, R., 2021; Hanson & Beguin, 2002). There is therefore the need to determine which of the test score equating methods (linear equating and separate calibration) would result in the most efficient or accurate technique for the comparison of students' continuous assessment measure.

The main objective of the study is to determine the efficient technique for the comparison of student's continuous assessment measure.

The specific objectives of this study are to:

a. assess the item parameter estimates of the Mathematic Achievement Test (MAT) used for Test Score equating of public and private schools;
b. determine the mean ability estimates of students in public and private secondary schools when their scores are equated through separate calibration;
c. determine the mean ability estimates of students in public and private secondary schools when their scores are equated through linear equating; and
d. examine which of the equating methods is more efficient.

From the objectives of the study, the following research questions were raised.

a. What are the mean ability estimates of students in public and private secondary schools when their scores are equated through separate calibration?
b. What are the mean ability estimates of students in public and private secondary schools when their scores are equated through linear equating?
c. Which of the equating methods is more efficient?

From the study objectives, the following research hypotheses were tested.

a. There is no significant difference in the ability estimate of students in public and private secondary schools for scores equated through separate calibration.

b. There is no significant difference in the ability estimate of students in public and private secondary schools for scores equated through linear equating.

METHODOLOGY

This study adopted an ex-post facto research design. This design was adopted for the study because it allowed analysis to be performed on existing data. The Population for the study comprised the 24,874 candidates that registered and sat for the 2016/2017 June/July NECO Mathematics examination in Ogun State. The 24,874 candidates were made of 11,671 males and 13,203 females (Ogun State Ministry of Education, Science and Technology 2016). A sample of 1139 candidates were selected using multistage sampling technique. A total of two LGAs were selected from each of the three senatorial districts of the State and from each of the selected LGAs, four schools were selected using non-proportional stratified random sampling technique with school ownership (Public and Private) serving as basis for stratification to make a total of 24 schools. All candidates in the selected 24 schools who sat for 2016 NECO Senior School Certificate Mathematics examination and whose result and Continuous Assessment result were made available were selected as sample for the study.

The study made use of two research instruments. They include softcopy of NECO mathematics Marksheet Record (MMR) that was made available by National Headquarters of the examination body. The second instrument was mathematics continuous assessment records of the selected schools for candidate that sat for the June/July SSCE. The records were made available by Ogun State Ministry of Education, Science and Technology.

The data were subjected to Separate equating to answer question one and Linear equating was used to answer question two. All hypotheses were tested using t-test at 0.05 level of significance.

RESULTS

Research question 1: What are the mean ability estimates of students in public and private secondary schools when their scores are equated through separate calibration?

To answer this research question, the ability estimates of the candidates from public and private schools were first estimated and then the ability estimates of candidates in public school were placed on the same scale ability estimates of candidates in private school through Lord-Stocking Test Characteristics Curve method of separate calibration equating method for the ability score of candidates in public school transformed to the scale of candidates in private school). To achieve the equating, linear equating formula given by Kolen and Brennan (2014) was used. The formula is given by

$\theta_{yi} = A\theta_{xi} + B$ ---------------------
4.1.3

Where A = Slope

B = Intercept

θ_{yi} = Ability of candidates from public school on the scale of ability of candidates from private school

$A\theta_{xi}$ = Ability of candidates from public school

To obtain the value of the slope and the intercept, the item parameter estimates of the MAT in public school and private school were respectively estimated using MIRT package and then the item parameter estimates obtained in public school were placed on the scale of those obtained from private school using equate IRT package of R Language and environment for statistical computing. The result is presented in Table 1.

Table 1

Equating constants of public and private schools' ability estimate under Separate Calibration

Slope (A)	Intercept (B)
1.10795	-0.85945

After the equating, the mean of the ability estimates of candidates from public school that have been placed on the scale of ability estimates from private school and the ability estimates of candidates of private school were obtained. The result is presented in Table 2.

Table 2

Mean of ability estimate of candidates of public school and private school after equating under separate calibration

	Public school	Private school
$\bar{\ddot{x}}$	-0.865	-0.001
SD	1.058	0.961

Table 2 showed that private school candidates' ability score (x = -0.001, SD = 0.961) was higher than public candidates' ability (x = -0.865, SD = 1.058). This result showed that private school candidates performed better than their counterpart in public school.

Research question 2:What are the mean ability estimates of students in public and private secondary schools when their scores are equated through linear equating?

To answer this research question, the ability estimates of the candidates from public school and private school were first estimated (see Appendix I) and then the ability estimates of candidates in public school were placed on the same scale ability estimates of candidates in private school through mean-sigma(linear equating) method of separate calibration equating method (see Appendix II for the ability score of candidates in public school transformed to the scale of candidates in private school). To achieve the equating, linear equating formula given by Kolen and Brennan (2014), was used. The formula is given by

$\theta_{yi} = A\theta_{xi} + B$---------------------- 4.1.4

Where A = Slope

B = Intercept

θ_{yi} = Ability of candidates from school type one on the scale of ability of candidates from school type two

$A\theta_{xi}$ = Ability of candidates from public school

To obtain the value of the slope and the intercept, the item parameter estimates of the MAT in public school and private school were respectively estimated using MIRT package and then the item parameter estimates obtained in school type one were placed on the scale of those obtained from school type two using equate IRT package of R Language and environment for statistical computing. The result is presented in Table 3.

Table 3

Equating constants of public and private schools ability estimate under linear

Slope (A)	Intercept (B)
1.01590	-0.62077

After the equating, the mean of the ability estimates of candidates from public school that have been placed on the scale of ability estimates from private school and the ability estimates of candidates of private school were obtained. The result is presented in Table 4.

Table 4

Mean of ability estimate of candidates of public school and private school after equating under Linear

	Public school	Private school
$\bar{x}$	-0.626	-0.001
SD	0.970	0.961

Table 4 showed that private school candidates' ability score ($\bar{x}$ = -0.001, SD = 0.961) was higher than public school candidates' ability (x= -0.626, SD = 0.970). This result showed that private school candidates performed better than their counterpart in public school.

Research question 3: Which of the equating methods is more efficient?

To identify which of the two methods of equating used in this study was the most effective, the standard error of the equating methods was obtained and compared. The most effective method is the equating method that produced the smallest standard error of equating (Kolen and Brennan, 2014). According to Kolen and Brennan (2014), the standard error of equating is obtained by taking the standard deviation of the equated scores. The standard errors of equating of the separate and linear equating methods are presented in Table 5.

Table 5

Standard error of equating of separate and linear equating methods

	Separate calibration	Linear equating
SD error of equating	1.058	0.970

Table 5 showed the standard error of equating of the equated scores obtained using separate calibration equating method and linear equating method. The table showed that the standard error of equating of the linear equating method was lesser (0.970) than that of the separate calibration equating method (1.058). This result showed that the linear equating method was more efficient than the separate calibration method of test score equating.

*Hypothesis 1:*There is no significant difference in the ability estimate of students in public and private secondary schools for scores equated through separate calibration.

To test this hypothesis, the difference in the estimated public and private school students' ability using separate calibration was determined using t-test statistic and the result is presented in Table 6.

Table 6

Difference in public and private school ability estimate of candidates after equating under separate calibration

School type	$\bar{x}$	SD	Mean Difference	t	df	Sig.(2-tailed
public	-0.865	1.058	-0.86423	-14.431	1138	0.000
private	-0.001	0.961				P < 0.05

Table 6 showed that the difference observed between the ability estimate of examinees from public school and private school was significant t_{1138} = - 14.431, p < 0.005). This result showed that examinees in public school significantly performed better than their counterpart in public school.

Hypothesis 2:There is no significant difference in the ability estimate of students in public and private secondary schools for scores equated through linear equating.

To test this hypothesis, the difference in the estimated public and private school students' ability using linear equating was determined using t-test statistic and the result is presented in Table 7.

Table 7

Difference in public and private school ability estimate of candidates after equating under linear equating

School type	$\bar{x}$	SD	Mean Difference	t	df	Sig. (2-tailed)
public	-0.626	0.970	-0.62511	-10.876	1138	0.000
private	-0.001	0.961				P < 0.05

Table 7 showed that the difference observed between the ability estimate of examinees from public school and private school was significant $t_{1138} = -10.876$, p < 0.005). This result showed that examinees in private school significantly performed better than their counterpart in public school.

DISCUSSION

Research question one revealed the mean ability estimates of students in public and private secondary schools when their scores are equated through separate calibration. After the equating, the mean of the ability estimates of candidates from public schools that have been placed on the scale of ability estimates from private schools and the ability estimates of candidates of private school obtained shows that private school candidates' ability score was higher than the public-school candidates' ability. This result implied that private schools' candidates performed better than their counterpart in public schools. These findings were not in agreement with the Agah (2013) who revealed in his study that the mean ability estimates of students in public school and private did not show any significant difference and also observed no difference in the mean ability estimates and other moments (standard deviation and variance). This implied that students in public school did not perform better than students in private school.

Research question two also showed the mean ability estimates of students in public and private secondary schools when their scores are equated through linear equating. After the equating, the mean of the ability estimates of candidates from public schools that have been placed on the scale of ability estimates from private schools and the ability estimates of candidates of private school obtained, shows that the private school candidates' ability score was higher than public school candidates' ability. Therefore, private schools' candidates performed better than their counterpart in public schools. These findings were not in agreement with the Agah (2013) who investigated the item parameter consistency values (Item Functioning) for public school and private school revealed in his study that the mean ability estimates of students in public school and private school did not show any significant difference and also observed no difference in the mean ability estimates and other moments (standard deviation and variance). This implied that students in public school did not perform better than students in private school.

Research question three also showed the most effective equating methods between the linear and separate calibration methods. After the equating, the result showed that the linear equating method was more effective than the separate calibration method to test score equating because the standard error of equating of the linear equating method was lesser. These findings were not in consonance with Yang (1997) who found linear equating to produce the largest amount of error. Yang's (1997) result shows that IRT equating methods were better than the linear (Tucker).

Research Hypothesis one provides the result of difference in public and private school estimate of candidates equating under separate calibration. The findings from the result showed that the difference observed between the ability estimate of examinees from public school and private school was significant. Hence, examinees in private schools significantly performed better than their counterpart in public schools. Consequently, after equating under linear equating from the hypothesis two raised, the findings showed that the difference observed between the ability estimate of examinees from public school and private school was also significant. Hence, examinees in private schools significantly performed better than their counterpart in public schools. These findings was also against Agah (2013) who in his study also revealed that the mean ability estimates of students in public school and private school did not show any significant difference when equated through separate calibration and linear equating. These findings are not in consonant with Hanson and Beguin (1999) who investigated the performance of separate versus concurrent estimation in putting item parameter estimates for two forms of a test administered in a common item equating design on the same scale. Their results among others showed that, the differences among the item parameter scaling methods used in separate estimation were much larger than the differences between concurrent estimation and the better performing scaling methods in separate estimation. This finding seems to negate that of Morrison and Fitzpatrick (1992) who found that concurrent calibration resulted in the least amount of equating error among four equating methods considered in their study. The finding of this study also differs from that of Yang (1997) who found linear equating to produce the largest amount of error. Yang's (1997) result shows that IRT equating methods were better than the linear (Tucker), Other research that compare separate and concurrent calibration have concluded that concurrent estimation performed somewhat better than separate estimation (Zhu et. al, 2022; Petersen, Cook, & Stocking, 1983; Wingersky, Cook, &Eignor, 1987), while Kim and Cohen (1998), concluded that the performance of separate estimation was equal to or better than concurrent estimation. The findings of this study have shown that linear equating method performed better than the separate calibration in the estimation of students' ability.

CONCLUSION

The study concluded that linear equating method is more efficient than separate calibration method in equating public and private secondary schools' continuous assessment. It is therefore recommended that CA of both public and private should be made equivalent so that the students can be on the same ability level and that test scores equating should be used to standardize students' continuous assessment scores.

REFERENCES

Afemikhe, O. A., (2007). *Assessment and Educational Standard Improvement: Reflections from Nigeria*. A paper presented at the 33rd Annual conference of the International Association for Educational Assessment held at Baku, Azerbaijan. September 16th 21st 2007.

Agah, J.J, (2013). Relative Efficiency of Test Scores Equating Methods in Comparison of Students Continuous Assessment Measures. Ph.D Thesis submitted to the Department of Science Education, University of Nigeria, Nsukka. https://www.unn.edu.ng/publications/files/AGAH%20JOHN%20JOSEPH.pdf

Akin-Arikan, Ç., & Gelbal, S. (2021). A Comparison of Kernel Equating and Item Response Theory Equating Methods. *Eurasian Journal of Educational Research*, *93*, 179-198. https://eric.ed.gov/?id=EJ1299641

Eiji, M., Catherine, M. H. & Yong-Won, L., (2008). Equating and Linking of Performance Assessment. *Applied Psychological Measurement. 24*(4): 325-337. Available at: http://upm.sagepub.com/cgi/content/abstract/24/4/325.

Emaikwu, S. O., (2004). *Relative Efficiency of Four Multiple Matrix Sample Models in Estimating Aggregate Performance from Partial Knowledge of Examinees Ability Levels*. An Unpublished PhD Thesis. University of Nigeria, Nsukka. https://globalacademicgroup.com/journals/the%20nigerian%20academic%20forum/Sunday11.pdf

González, J., & Gempp, R. (2021). Test comparability and measurement validity in educational assessment. In *Validity of Educational Assessments in Chile and Latin America* (pp. 173-204). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-78390-7_8

Hanson, B. A. & Beguin, A. A., (2002). Obtaining a Common Scale for IRT Item Parameters using Separate Versus Concurrent Estimation in Common Item Nonequivalent Groups Equating Design. *Applied Psychological Measurement. 26*(1): 3-24. https://doi.org/10.1177/014662160202600100

Hanson, B. A. & Beguin, A. A., (2002). Obtaining a Common Scale for IRT Item Parameters using Separate Versus Concurrent Estimation in Common Item Nonequivalent Groups Equating Design. *Applied Psychological Measurement. 26*(1): 3-24. https://doi.org/10.1177/014662160202600100

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education*, 24 (1), 393-446. https://doi.org/10.3102/0091732X024001393

Iris Eekhout, Ann M. Weber, Stef Buuren et al. Equate groups: An innovative method to link multi-item instruments across studies, 29 March 2022, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-1439153/v1]

Jones, P., Tong, Y., Liu, J., Borglum, J., & Primoli, V. (2022). Score comparability between online proctored and in-person credentialing exams. *Journal of Educational Measurement.* https://doi.org/10.1111/jedm.12320

Kasali, J., & Adeyemi, A. (2022). Estimation of Item Parameter Indices of NECO Mathematics Multiple Choice Test Items among Nigerian Students. *Journal of Integrated Elementary Education, 2*(1), 43-54. doi: https://doi.org/10.21580/jieed.v2i1.10187

Kim, S. H. & Cohen, A. S. (1998). A comparison of Linking and Concurrent Calibration under Item Response Theory. *Applied Psychological Measurement*, 22 (2), 131-143. https://doi.org/10.1177/01466216980222003

Kolen, M. J. & Brennan, R. L., (2004). *Test equating, scaling and linking* (Second Edition). USA: Springer. https://link.springer.com/book/10.1007/978-1-4939-0317-7

Kolen, M. J., & Brennan, R. L. (2013). Test equating: Methods and practices. Springer Science & Business Media. https://link.springer.com/book/10.1007/978-1-4757-2412-7

Kolen, M. J., & Brennan, R. L. (2014). Item response theory methods. In *Test Equating, Scaling, and Linking* (pp. 171-245). Springer, New York, NY.

Makiney, J. D., Rosen, C., Davis, B.W., Tinios, K. & Young, P. (2003). *Examining the Measurement Equivalence of Paper and Computerized Job Analyses Scales.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FLORIDA

Morrison, C. A. & Fitzpatrick, S.J. (1992). Direct and Indirect Equating: *A Comparison of Four Methods using the Rasch Model*. Available at: https://eric.ed.gov/ERICdoes/data/pdf

Onjewu, M. A., (2007). *Assuring Fairness in the Continuous Assessment Component of School Based Assessment Practice in Nigeria*. Paper Presented at the 33rd Annual Conference of the International Association for Educational Assessment. Baku, Azerbaijan.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, 8, 137-156. https://doi.org/10.3102/10769986008002137

Pinsoneault, T. B, (1996). Equivalency of Computer-Assisted Paper-and-Pencil administered version of the Minnesota Multiphasic Personality Inventory-2.*Computers in Human Behavior, 12*, 291-300. https://doi.org/10.1016/0747-5632(96)00008-8

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517–529. https://doi.org/10.1037/0021-9010.87.3.517

Schalet, B. D., Lim, S., Cella, D., & Choi, S. W. (2021). Linking scores with patient-reported health outcome instruments: A validation study and comparison of three linking methods. *Psychometrika*, *86*(3), 717-746. https://link.springer.com/article/10.1007/s11336-021-09776-z

Wingersky, M. S., Cook, L. L., &Eignor, D. R. (1986). *Specifying the characteristics of linking items used for item response theory item calibration*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. https://doi.org/10.1002/j.2330-8516.1987.tb00228.x

Yang, W. (1997). The Effect of Content Mix and Equating Method on the Accuracy of Test Equating using Anchor. *Item Design*. Available at: http://eric.ed.gov

Zhu, W., Konishi, D., Welk, G., Mahar, M., Laurson, K., Janz, K., & Baptista, F. (2022). Linking Vertical Jump and Standing Broad Jump Tests: A Testing Equating Application. *Measurement in Physical Education and Exercise Science*, 1-9. https://doi.org/10.1080/1091367X.2022.2112683