COMPARABILITY OF ITEM PARAMETERS OF EQUIVALENT EXAMINATIONS USING LINEAR TEST SCORE EQUATING METHOD

By

ADEYEMO Emily Oluseyi
Department of Educational Foundations and Counselling
Faculty of Education
Obafemi Awolowo University, Ile-Ife. Nigeria
e-mail :-seyiadeyemo2007@yahoo.com
+2348033907463

COMPARABILITY OF ITEM PARAMETERS OF EQUIVALENT EXAMINATIONS USING LINEAR TEST SCORE EQUATING METHOD

Abstract

This study examined how items from two examining bodies could be compared to confirmed the respective psychometric properties. It determined the comparability of the items in terms of examinee's scores and item parameters using test score equating method. The study adopted a survey research design The population comprised all Senior Secondary School three (SSS III) in Ekiti. The sample size consisted 360 students selected from the three senatorial districts using simple random sampling techniques . Two instruments were used for the study, these were the Mathematics Achievement Test A (MAT A) and the Mathematics Achievement Test B adopted from 2015 Ekiti State Unified Examinations and 2015 WAEC respectively. These instruments each contained 50 Multiple-choice items . The instruments were administered on 30 SSS III students from each of the 12 schools selected . The data collected from the administered MAT A and B were analyzed and the hypotheses tested using t – test statistics and the application of test score equating method. The difference in the difficulty indices of both examinations was significant ($t = 8.682$ $p < 0.05$). there was a significant difference in the discrimination power of the Ekiti State Unified and WAEC mathematics examinations ($t = 16.664$, $p < 0.05$). the relationship in students' performance in the two examinations was significant($t = 4.664$, $p < 0.05$) and ($r = 0.173$, $p = 0.001$). And finally ($t= 3.330$, $p = 0.001$) at 0.05 level of significant students' performance in the State Unified mathematics examination made a statistically significant contribution to students' success in the WAEC mathematics examination The study concluded that the comparability of the two mathematics examination items in terms of examinee's scores and item parameters using linear equating method administered were equivalent and the scores yielded from both examinations were symmetry.

Key words :- Comparability, Test score equating, Items parameter, Unified test , Equivalent scores

COMPARABILITY OF ITEM PARAMETERS OF EQUIVALENT EXAMINATIONS USING LINEAR TEST SCORE EQUATING METHOD

Comparability among different test items are widely conducted by test developers , educational researchers and psychometricians . Equating, calibration and moderations are different terms used by educational researchers to describe the comparisons .Equating according to Doran (2004) is a process of deriving a function mapping score on an alternate form of a test on the scale of an Anchor form such that after equating, any given scale score has the same meaning regardless of which test form was administered. Equating is a technical procedure or process, conducted to establish comparable score on different version of a test, allowing them to be used interchangeably. It is used to adjust person location estimate from different metrics and estimate on a common metric for the purpose to facilitate comparing individuals scores. Equating method can be used to adjust for differences in difficulty across alternate forms , resulting in comparable score scales and more accurate estimates of ability, establishing validity across forms and years, test fairness and test anxiety etc.

When two tests forms have been successfully equated, educators can validly interpret performance on one just form as having the same substantive meaning compared with the equated score of the other test form. The use of different forms of equivalent test or different test is to measure the same constructs and raise the issue of comparability of the test score. After the two test are equated, pairs of equivalent test score becomes identical. The outcomes after equating are the reference form test scores remaining as they are and the test scores on each of the other forms are equated to the referenced test score by a method selected for equating in test score in all forms measured to the same scale.

The process of equating tests begins by understanding how to link two test, then several tests and finally connecting all possible tests. The term *linking* is used to describe the transformation from a score on one test to a score on another test. The main aim of connecting tests intended to measure the same variable is to ensure that the separate measures each test implies are expressed together on a single common scale. It also involves the practice of pairing or matching scores on two test forms with no strong claim that the paired scores have the same substantive meaning.

When two forms of test are said to be equated, they measure the same contents cognitive process and support the same inferences about what students know and can do with the aim of producing interchangeable scores. To make achievement test meaningful and useful according to Koretz (1999) ,they are subjected to statistical treatment like linking, scaling or equating. Equating methods can be used to adjust for differences in difficulty across alternate forms, resulting in comparable score scales and more accurate estimates of ability. Test equating is a process used to make test scores across different forms of the same test interchangeable. When two test forms have been successfully equated, educators can validly interpret performance on one test form as having the same substantive meaning compared to the equated score of the other test form. The use of different forms of the same test (or different test aiming to measure the same constructs from year to year, or school to school raises the issue of the comparability of test scores.

After two test are equated, pairs of equivalent scores become available. For example, such a pair of equivalent scores could be (17, 19) which indicates that a total scores of 17 on the first paper is equivalent to a total score of 19 on the second paper. The outcomes after equating are the reference form test scores remaining as they are and the test scores on each of the other forms are equated to the reference test scores by a method selected for equating in test scores. Livingstone (2004) provided a simple general definition of equating as a score on a new form and a score on

another form are equivalent in a group of individuals that have taken the form if they represent the same relative position in the group.

Linking is a concept different from equating and does not support the same interpretations supported by equating. Some of the confusion in the use of these terms is likely based on the fact that the same procedures are used in both linking and equating (CCSSO 2011). In both linking and equating the scores on one test form are matched to or paired with scores on another test form. For example, students' scores on a statewide standards based assessment can be paired or linked to scores on a standardized norm referenced test. Such a linking would result in a table with two columns; each row would link a score on a state test form to a particular score on a nationally norm-referenced test form (and vice versa), thereby the two are illustrated in Table one

Table I -Standards- Based Assessment Scores Vs. Norm- Referenced Test Score

Standards-Based Assessment(SBA) Scores	Norm-Referenced Test(NRT)Score
325	422
336	439
351	447

CCSSO 2011

The proper interpretation of this linking is described with statements like "Students who make a 325 on SBA will most likely earn a score of 422 on NRT" "Those who get a 447 on the NRT would be expected to score a 351 on the SBA" Equating two test forms supports a much stronger claim. If the SBA and NRT were successfully equated, a valid interpretation would be"Student with a score of 325 and students with NRT score of 422 have a very similar level of knowledge and skill with respect to what is being measured."

To say two forms are equated is to say that they measure the same content and cognitive processes and support the same inferences about what students know and can do. This is a very strong claim. Linking is a much weaker claim that merely asserts an association between scores on different assessments. Equating also asserts an association between scores, but equating has the additional connotation that these paired scores have the same substantive meaning. The connecting of different tests form can be done in several ways. Wright and Stone (1999) have suggested that when an easy test is linked or connected with a hard test, a set of common items are included in both tests, so that the common items become hard in the easy test and become easy in the hard test.

When equating is conducted, equating error is bound to happen. This is because examiners who take tests are considered to sample from a population, therefore, equating errors are present in systematic error and random equating error. Systematic errors are usually caused by assumption failures in the equating method, bias in the sample statistics etc. Random equating errors on the other hand are caused by sampling errors. Both systematic and random equating error influence the interpretation of results therefore caution needed to be taken when sampling tests for equating.

The following five requirements of equated scores are important as highlighted by Dorans (2000):

- instruments to be equated must measure the same construct equally well.
- : instruments that measure the same construct but which differ in reliability should not be equated.
- the linking function for equating outcome scores of instrument *Y* to those of instrument *X* should be *inverse* of the linking function for equating the outcome scores of *X* to those of Y.

- : the respondent ought to be indifferent about being tested by either one of two instruments that have been equated. This is in two parts: (i) the average or expected test performance should be equal so it requires equating methods to track and take account of differential test difficulty. (ii) The examinee ought to have the same expected distribution of performance on either one of two equated tests. " equating is either impossible or unnecessary"
- : the choice of the (sub) population used to compute the equating function between the scores of instruments X and Y should not matter i.e., the equating function used to link the outcomes of X and Y should be population invariant.

. There are conditions that needed to be met for equated scores i.e. the scores obtained after applying equating methods. Many procedures for equating tests have developed over the years, Holland & Dorans (2006) considered three factors when attempting to develop a taxonomy of methods: these are common population versus common-item data collection designs, observed-score versus true-score procedures, and linear versus nonlinear methods. Because equating is an empirical procedure, it requires a data collection design and a procedure for transforming scores on one test form to scores on another. Linear methods produce a linear function for mapping the scores from **X** to **Y**, while nonlinear methods allow the transformation to be curved. Observed-score procedures directly transform (or equate) the observed scores on **X** to those on **Y**. True-score methods are designed to transform the true scores on **X** to Z the true scores of **Y**. True score methods employ a statistical model with an examinee's true score defined as his or her expected observed test score based on the chosen statistical model. The psychometric models used are those of classical test theory and item response theory. Holland and Hoskens (2003) have shown how these two psychometric models may be viewed as aspects of the same model.

There are assumptions of equating are criteria that test equating must meet before being successfully employed. These criteria according to Doran (2000) are

:i The two tests must measure the same construct (or same ability).

ii The equating must achieve equity (i.e., for individuals of a given proficiency or ability level, the conditional distributions of scores on each test must be equal)

.iii The equating transformation should be symmetric (i.e., the equating of Y to should be the inverse of the equating of X to Y)

iv The equating transformation should be invariant across sub-populations of the population on which it is derived.

According to Kolen and Brennan (2004), these assumptions have been used as the principal basis for developing equating procedures. Dorans and Holland (2000) stated that there is not much of a theoretical underpinning for test equating. They added that each of these assumptions can be criticized as being vague, irrelevant, impractical, trivial,or hopelessly stringent and, even if they are violated, equating can still be carried out, and there is rarely any indication in the data to alert that something inappropriate has been done. However, Kolen and Brennan (2004) have also stated that in equating practice, every effort should be made to assure that the assumptions are satisfied to the greatest extent possible.

Statement of the Problem

Results from public examination bodies showed that students are not performing well in external examination compared with their performance in teachers' made test, especially in the Unified which was conducted to qualify them for the final SSSE . Several reasons have been given among which were inadequate infrastructures, poor teaching methods etc. An important area which seemed to have been overlooked is the comparability of the test item therefore the need to equate the state unified and the WASC Mathematics for 2015 test items to see how comparable the items are

The purpose of this study was to examine how items from the two examining bodies the Ekiti State Unified Examination & WASSCE can be made comparable and to confirm the respective psychometric properties of the test items in the Mathematics examination.
The specific objectives of the study were to:

i. determine the item parameters of the Ekiti State Unified Mathematics Examination items;
ii. determine the item parameters of the WASCE Mathematics items and;
iii. determine the comparability of the two mathematics examination items in terms of examinee's scores and item parameters using test score equating method.

Method:- The study adopted a survey research design which entails gathering information from a representative sample of a population. The study population comprised all Senior Secondary School three (SSS III) in Ekiti. The sample size consisted 360 students selected from the three senatorial districts using simple random sampling techniques . Two instruments were used for the study, these were the Mathematics Achievement Test A (MATA) and the Mathematics Achievement Test B adopted from 2015 Ekiti state Unified Examinations and 2015 WAEC respectively. These instruments each contained 50 Multiple-choice items . The instruments were administered on 30 SSS III students from each of the 12 schools selected . The data collected from the administered MAT A and B were analyzed and the hypotheses tested using t – test statistics and the application of test score equating method.

Results.
Research Question:- What is the relationship between the two mathematics examination items in terms of examinee score and item parameters?
In order to establish the equivalence of WAEC and UNIFIED, examinees' test scores on the two forms of test were compared. Table 2 presents the descriptive statistics of the test scores.

Table 2: Descriptive statistics of 50-item UNIFIED and 2014 WAEC

	Mean	Sd	Min	Max
WAEC(x)	22.57	18.55	0	50
UNIFIED(y)	15.62	6.96	13	30

Table 2 shows the descriptive statistics of the test scores of the examinees on WAEC and UNIFIED tests. Table 2, revealed that WAEC items were more difficult (M = 22.57; SD = 18.55) than the UNIFIED items. (M = 15.62; SD = 6.96). Thus, to assess the equivalent scores of the tests, the test scores emanating from the administration of the items were linked. To achieve this, WAEC test scores were transformed to the scale of UNIFIED test using linear equating. According to Kolen and Brennan (2014), linear equating is represented by

$m_y(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + [\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)]$ -------- Equation 1

Where $\frac{\sigma(Y)}{\sigma(X)}$ = Slope usually represented with A

$\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)$ = Intercept, usually represented with B

On substitution, equation 1 becomes

$m_y(x) = y = Ax + B$ ------ Equation 2

This was achieved using *equate* package of R language and Environment for statistical computing. The results are presented as in Table 3:

Table 3: slope and intercept of the linking

Slope	Intercept
7.1471	0.3754

On substitution for slope and intercept in equation 2 we have

$$m_{unified}(WAEC) = UNIFIED = 0.3754\,WAEC\,SCORE + 7.1471$$

Thus the equating function used in placing WAEC score on the scale of UNIFIED test for effective comparison of the test scores form the two tests. The WAEC and UNIFIED test equivalent using the equating function

The WAEC scores of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49 and 50 is equivalent 7.15, 7.52, 7.90, 8.27, 8.65, 9.02, 9.40, 9.77, 10.15, 10.53, 10.90, 11.28, 11.65, 12.03, 12.40, 12.78, 13.15, 13.53, 13.90, 14.28, 14.65, 15.03, 15.41, 15.78, 16.16, 16.53, 16.91, 17.28, 17.66, 18.03, 18.41, 18.78, 19.16, 19.54, 19.91, 20.29, 20.66, 21.04, 21.41, 21.79, 22.16, 22.54, 22.91, 23.29, 23.66, 24.04, 24.42, 24.79, 25.17, 25.54 and 25.92 of UNIFIED score respectively.

The results of the research question showed that the WEAC scores is equivalent to the Unified scores using linear equating method. Related research conducted by Adewale (2015) where he equated two year BECE results in Basic Science and Technology in Oyo State, Nigeria. when the scores are transformed using t-score, the scores from the two examinations (2013 and 2014) were very close when the raw scores are low. This study also corroborated that of Olatunji (2015) where she found out that linear equating method equated some of the scores obtained in WAEC and NECO

Hypothesis 1: There is no significant difference in the estimated difficulty index of the State Unified and WASCE Mathematics items

To test the hypothesis, t-test statistics was used to establish the difference in the obtained difficulty indices of the two Mathematics Examinations as in Table 4

Table 4 Difference in the estimated difficulty index of the State Unified and WASCE Mathematics items

Examination	N	$\overline{X}$	SD	t	df	Sig
WASCE	50	0.4513	00690	8.682	98	0.000
Unified Exam	50	0.3124	0.0897			

The result in Table 4 showed that the average difficulty indices of the Ekiti State Unified and the WASCE Mathematics Examinations respectively were 0.4513 and 0.3124 and with a t-test value ($t = 8.682$, $df = 98$, $p < 0.05$), the difference in the difficulty indices of both examinations was significant.

Hypothesis 2: There is no significant difference in the estimated discrimination index of the State Unified and WASCE mathematics items

To test the hypothesis, t-test statistics was used to establish the difference in the obtained discrimination indices of the two mathematics examinations as contained in Table 5

Table 5 : Difference in the estimated discrimination index of the State Unified and WAEC Mathematics items

Examination	N	$\overline{X}$	SD	T	df	Sig
WASCE	50	0.7381	0,1418	16.664	98	0.000
Unified Exam	50	0.2303	0.1622			

Table 5 showed that with t-test value (t = 16.664, df = 98, p <0.05), it was concluded that there was a significant difference in the discrimination power of the Ekiti State Unified and WAEC Mathematics Examinations.

Hypothesis 3: The difference in students' performance in the State Unified and the WAEC conducted Mathematics Examinations is not significant

To test this hypothesis, students' score in both Mathematics Examinations were compared using the paired sample t-test statistic and the result is as presented in Table 6.

Table 6 :- Difference in the performance of students in the State Unified and the WAEC mathematics items

Examination	N	$\overline{X}$	Sd	R	t	df	sig
WASCE	360	18.8111	14.29759	0.173	4.664	359	0.000
Unified Exam	360	15.2833	5.25482				

The result in Table 6 showed that students' class average performance in the two examinations were 18.81 and 15.28 for WAEC and the State Unified Examination respectively. The t-test value (t = 4.664, p < 0.05) showed that the difference in the performance was significant and with Pearson Product Moment Coefficient (r = 0.173, p = 0.001), the relationship in students' performance in the two examinations was significant.

Hypothesis 4:

Students' performance in the State Unified mathematics examination will not significantly predict their success in the WASCE mathematics examination.

To test this hypothesis, students' scores in both examinations were analysed with the use of linear regression. Scores in the State Unified Examination was used as the independent variable to predict scores in the WAEC mathematics that served as the dependent variable. The result is presented in Table 6

Table 6: Regression

R = 0.173 R^2 = 0.030 Adjusted R^2 = 0.027 F =11.00*	Unstandardized Coefficients B	 Std. Error	Standardized Coefficients Beta	T	Sig.
(Constant)	11.603	2.289		5.070	.000
State Unified Mathematics Examination	.472	.142	.173	3.330	.001

a. *Dependent Variable: State Unified Exam Score*

The result in Table 6 showed that R^2 = 0.030 which indicated that 3% of the variance in students' performance in the WAEC Mathematics Examination could be explained by the regression model.. With the F-ratio value (F = 11.00, p = 0.000), the regression model explains a statistically significant proportion of the variance and predicts WAEC mathematics success accurately. The regression coefficient (β = 0.472) as shown on Table 6 indicated that for every unit increase in students' performance in the State Unified Mathematics Examination the model predicted an increase of 47.2% in students' score in the WAEC Mathematics Examination. Thus the regression equation can be represented with the equation *WAEC Score = 11.603 + (0.472) Unified exam score*. Finally with the t-tests value (t= 3.330, p = 0.001) at 0.05 level of significant students' performance in the State Unified Mathematics Examination made a statistically significant contribution to students' success in the WAEC Mathematics Examination.

Discussion

The results of research showed that the WEAC scores is equivalent to the Unified scores using linear equating method. Related research conducted by Adewale (2015) where he equated two year BECE results in Basic Science and Technology in Oyo State, Nigeria revealed that . when the scores were transformed using t-score, the scores from the two examinations (2013 and 2014) were very close when the raw scores are low. This study also corroborated that of Olatunji (2015) where she found out that linear equating method equated some of the scores obtained in WAEC and NECO.

The results of hypothesis one revealed that there was a significant difference in the estimated difficulty index of the Ekiti State Unified Examination and the West Africa Senior Certificate Examination Mathematics items. From the result, the estimated difficulty indices of both Mathematics items were 0.4513 and 0.3124 respectively. The difference between 0.4513 and 0.3124 was significant.

Hypothesis two showed the estimated discrimination indices of both examinations Mathematics items (ESUE and WASCE) were 0.2303 and 0.7381 respectively. Even though the difference between the two indices was small, there was a significant difference in the discrimination power, also, the t-test value was t=16.664, df =98, p<0.05.

The third hypothesis showed that students class average performance in the both WAEC and ESUE were 18.81 and 15.28 respectively. The t-test value (t-4.664, p<0.05) showed that the difference in the performance was significant and with Pearson Product Moment Coefficient (r = 0, p = 0.001), the relationship in students' performance in the WAEC and ESUE were significant. It was positive, reflecting the fact that each case of the two variables were positively related and strong.

Hypothesis four showed students' performance in the ESUE Mathematics would significantly predict the success in the WASCE Mathematics ($R^2 = 0.030$).. Therefore, scores of the Unified Examination in Ekiti state could used to equate the scores of the WASCE.

Conclusion

. The study therefore concluded that Mathematics items that the Ekiti State Unified Examination and the WAEC are administering are equivalent. The scores yielded from the State Unified items and the WAEC items are symmetry confirming the fact that they are measuring the same construct; even though both the difficulty index and discriminating power of the two test items were significantly different, but the UNIFIED still serve as an effective way of weeding off the unqualified students. The result of the analysis revealed further that for every unit increase in student's performance in the State Unified Mathematics Examination, the model predicts an increase of 47.2% in student's score in WAEC Mathematics Examination. Therefore, students' performance in the State Unified Mathematics Examination made a statistically significant contribution to students' scores in the WAEC Mathematics Examination.

References

Adewale J.G. (2015) Equating two Year BECE Results in Basic Science and Technology in Oyo State, Nigeria. International Centre for Educational Evaluation (ICEE) Institute of Education, University of Ibadan – Nigeria

Doran, N.J. & Holland, P.W. (2000). Population Invariance and Equating of Tests: Basic Theory and the linear case. *Journal of Educational Measurement, 37:* 281 – 306.

Dorans, N.J. (2000). *Linking scores from multiple instruments.* Center for Statistical Theory and Practice Educational Testing Service. ETS.

Dorans, N.J., Pommerich, M. & Holland, P.W. (Eds.) (2007). *Linking and Aligning Scores and Scales.* New York: Springer

Dorans, N.J., & Liu, J., & Hammond, S. (2008). Anchor Test Type and Population Invariance: An Exploration Across Subpopulations and Test Administrations. *Applied Psychological Measurement, 32,* 81-97.

Dorans, N.J., Moses, T.P., & Eignor, D.R. (2010). *Principles and practices of test score equating* (ETS Research Report No. RR-10-29). Princeton, NJ: ETS.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.),*Educational measurement* (4th ed., pp. 187-220).Westport, CT: Praeger

Holland P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68,*123–149

Holland, P. W.,& Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.),*Educational measurement* (4th ed., pp. 187-220).Westport, CT: Praeger

Koretz, D. M., Bertenthal, M. W., & Green, B. F. (Eds.). (1999). *Embedding questions: The pursuit of a common measure in uncommon tests* (Report of the Committee on Embedding Common Test Items in State and District Assessments, National Research Council). Washington DC: National Academy Press

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag

Livingston, S.A. (2004). *Equating test scores (without IRT).* Princeton: Educational Testing Service.

Livingston, S.A. (2004). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30,* 23-39

Lord, F.M. (1980): Applications of item response theory to practical testing problems. Hillsdale, NJ: Eribaum

Olatunji, D. S. (2015). Analysis of Linear and Equipercentile Equating Of Senior School Certificate Examination Economics Multiple-Choice Papers in Kwara State, Nigeria. Unpublished Ph.D. Thesis, University of Ilorin.

Wright, B. & Stone, M. (1999): Measurement Essentials (2nd Ed.) Wilmington Delaware, Wide Range Inc