COMPARISON OF CLASSICAL TEST THEORY EQUATING METHODS USING THE NON- EQUIVALENT ANCHOR TESTS AMONG SECONDARY SCHOOL STUDENTS IN OSUN STATE

[1]OGUNSAKIN, I. B. & [2]ADEDIWURA, A.A.

[1&2]Department of Educational Foundations and Counseling,
Faculty of Education, Obafemi Awolowo University, Ile-Ife
sakinbamikole@yahoo.com, yemtoy20002000@gmail.com

Abstract

The study estimated the item parameters of the two forms of the non-equivalent anchor tests to be equated and examined the relative effectiveness of classical test theory equating methods in the non- equivalent anchor tests to be equated. The study adopted the Non-Equivalent Anchor Test (NEAT) design. The population of the study comprised 137,083 Senior Secondary two (SS II) students in Osun State. Sample for the study consisted of 1080 students that were selected using a multistage sampling procedure. A total of 45 Senior Secondary 2 (SS II) students were selected from each school using the simple random technique. Two adapted instruments titled Mathematics Achievement Test Form A (MATFA) and Mathematics Achievement Test Form B (MATFAB) were used to collect data for the study. The instruments were adapted from 2014 and 2015 West African Examination Council Mathematics objective items which served as MATFA and MATFAB respectively. The 2014 National Examination Council Mathematics objective test items were adopted. This served as the anchor items for both form A and B respectively. The reliability of the two instruments were established to be 0.79 for MATFA, and 0.75 MATFAB using Kuder-Richardson 20(KR-20) formula. Data collected were analyzed using IRTEQ software package, Common Item Program for Equating (CIPE) and R version (3.4.1) software. The results showed that the average difficulty and discrimination power of form A and form B Non-equivalent anchor test equated under CTT were (X = 0.53 ,X = 0.23) and (X =0.29, X = 0.27) respectively. The results further showed that Tucker's mean (TMEAN), Levine mean (LMEAN) and Braun-Holland Mean (BMEAN) were the most effective method of equating followed by Tucker's Linear (TLIN) and Levine Linear (LLIN) method and finally by equipercentile method. The study concluded that mean equating methods were more effective compared to linear and equipercentile equating methods.

Key words: Classical Test Theory, Test Equating, Non-equivalent Anchor Tests

Introduction

It is a common practice in high- stake examination such as the West Africa Examination Council (WAEC), the National Examination Council (NECO), National Board for Technical and Business Trades Examination (NABTEB), Graduate Records Examinations (GRE), Scholastic Assessment Test (SAT) and The Joint Admission and Matriculation Board (JAMB) to construct multiple forms of test for the purpose of administration. These test forms are constructed to measure the same construct. Each of these test forms is used as a testing instrument. They are nominally parallel and usually differ in difficulty. The measurement challenges associated with it is how to untangling the unintended

differences in difficulty among the test forms from the ability of the examinee. The purpose of untangling the unintended differences among the test forms is aiming towards enhancing the fairness of test or examination.

Nworgu (2011) opined that the fairness in test items can be achieved by using valid and reliable measurement instrument which largely depends on the quality of the items. Subsequently, important decisions are made based on outcome of the test. Moreover, the fairness in test items can be achieved using statistical tool called equating. This can be done by considering the degree of difficulty index across different forms of the tests (Form A and form B), and make necessary adjustment. The aim of the adjustment is to ensure that an examinee will not have undue advantages over others. In an ideal situation, test forms should be assembled to be strictly parallel. Based on this, they would have indistinguishable psychometric properties, hence, equating might be unnecessary. In reality, it is practically impossible to construct multiple forms of a test that are strictly parallel; therefore, equating is necessary to adjust the test construction process. (Agah 2013)

Practically, test equating/linking is a significant statistical tool that can be used to adjust scores amid two parallel test administered to students(Angoff, 1971; Kolen& Brennan, 1 995 & 2004; Dorans & Holland, 2000; and van Davier et al. 2004). Apart from using equating method for score adjustment, it can be used as a tool to amend score inflation. In recent time West African Examination Council came up with modality on registration of student's right from SS1. The body came up with the capturing of the Continuous Assessment (CA) scores of students every year so that at Senior Secondary 3 the number of candidates for its West African Senior School Certificate Examinations (WASSCE) would have been known. In the light of this, some schools might tend to inflate the Continuous Assessment (CA) scores that will be uploaded for each of the students on their website. (https://blueprint.ng/waec-to-register-candidates-from-ss1)

The issue of comparability of standards, which deals with uniformity and quality of assessment instruments, as well as honesty and integrity in reporting of assessment result among others, has been a problem right from the introduction of continuous assessment (Agah, 2013). In addition, this is no longer a news that the issue of score inflation on Continuous Assessment (CA) scores is a very big concern among schools. Those schools that are engaging in this menace are trying to help their students to pass their final examinations. In the light of this, test equating might be the appropriate statistical tool to amend the scores by placing the scores on the same scale, so that students from school A might not have undue advantages over another students from school B. Afemikhe (2007) suggested that enhancement strategies such as moderation, self-assessment and test scores equating can be used as educational standards control mechanisms in Nigeria.

Dorans and Holland (2000) affirmed that conducting a linking or an equating, an important issue in developing a common scale across different tests is whether the comparability of the test scores is dependent on particular groups of examinees. This is defined by characteristics such as gender, language, race, geographic region. Generally, this issue is commonly referred to as population invariance in linking and equating. This is further affirmed by Dorans and Holland (2000) that population invariance in equating exists when the relationship between two scales are the same for two or more subpopulations of examinees and hence the function used to equate the scales is not dependent on subpopulations. Huggins (2012) opined that a lack of equating invariance leads to a situation whereby examinees having the same score on one scale but belonging to different subpopulations have different expected test scores on the

corresponding equated scale. This situation results in an expected advantage for one or more subpopulations of students; hence, a lack of invariance in equating is a concern for fairness and equity in assessment, equating invariance focuses on fairness and equity at the reported (i.e., equated) test score level.

In order to achieve fairness among the examinees Test equating is necessary to be carried out through statistical models called Classical Test Theory and Item Response Theory. This will enhance fairness among the test taker by disallowing candidate A not to have unwarranted advantages over candidate B. Classical Test Theory (CTT) and Item Response Theory (IRT) are feasible theories to carry out item analysis. In this light, IRT might be the best model that can give concise and reliable results in test equating process, since it is a modern theory of tests.

Moreover, equating processes fall into two basic forms/categories observed scores and true scores. These two categories are serving different purposes in equating of test items. As pointed out by Kolen and Brennan (2004), the aim of observed score equating is to make an adjustment such that the properties of score distributions across parallel forms are as similar as possible. The result is to establish equating relationship between observed scores on two parallel forms. The goal of true score equating procedures is to map true scores on one form to true scores on another form (via the definition of true score in either Classical Test Theory or Item Response Theory) as opposed to mapping observed scores. in the process of mapping the scores, the role of anchor items are very important.

Anchor items refer to a single set of items that appear on two or more tests forms. These items, common to two or more forms, are said to serve as "anchors" that fix the measurement scale on which the test forms are connected (equated). The items common to two or more forms can also be described as the "links" which connect the forms together onto a common scale. However, an anchor test is a representative or a miniature version (i.e., a minitest) with respect to both content and statistical characteristics, of the tests being equated Kolen&Brennan, (2004). To ensure statistical representativeness, the usual practice is to make sure that the mean and spread of the item difficulties of the anchor test are roughly equal to those of the tests being equated (Dorans, Kubiak, & Melican, 1998).The requirement that the anchor test be representative of the total tests (i.e., the tests being equated) with respect to content has been shown to be important by Kleinand Jarjoura (1985) and Cook and Petersen (1987). Peterson, Marco, and Stewart (1982) demonstrate the importance of having the mean difficulty of the anchor tests close to that of the total tests. However, the literature does not offer any proof of the superiority of an anchor test for which the spread of the item difficulties is representative of the total tests.

Moreover, a minitest has to include very difficult or very easy items to ensure adequate spread of item difficulties, which can be problematic as such items are usually scarce (one reason being that such items often have poor statistical properties, such a slow discrimination, and are thrown out of the item pool). The importance of anchor items in equating are very essential in that they allow a new test to be used and equated at each successive operational test administration.

Furthermore, anchor items can be grouped into two categories namely: external which is also know as appended anchor test and internal which is also known as embedded anchor test. External anchor usually refers to items that are administered in a separately timed section and that do not count towards the examinee's score. One major advantage of external anchors is that they may serve multiple purposes, such as equating, pretesting, and trying out

of new item types of items. Items in an internal anchor test are part of the assessment and count towards each examinee's score. Internal anchor items are usually spread throughout the test. Some external anchors (i.e., items that are left out of or are external to the total score) are administered internally and consequently face some of the issues associated with internal anchors. In chosing the anchor items the following features must be considerd. They are; test length, content, statistical properties, invariance over time (i.e., lack of item parameter drift), and utility/placement of the anchor set and common items.

Litrature revealed that the length of a single test is highly correlated with its reliability. Longer tests are often more reliable than shorter tests that measure the same construct . Consequently, it would also be reasonable to assume that the length of an anchor set should be such that its reliability is of a respectable level; as high reliability is a necessary condition for valid interpretations of test scores. Most of the research suggests that the anchor set should represent at least 20% of the operational test; or for IRT equating methods. Conversely, 15 items should be used to properly serve this purpose as confiremed by Cook & Eignor, 1991; Fitzpatrick, 2008; Hambleton, Swaminathan, & Rogers, 1991; Kolen &Brennan, 2004; McKinley & Reckase, 1981; Vale, Maurelli, Gialluca, Weiss, & Ree, 1981). Studies such as Keller, Egan, and Schneider (2010) even suggest upwards of including 25% of the operational test as linking items when using an internal anchor, as one of their important findings in varying the number of items used for equating resulted in the ability to appropriately detect items as aberrant as the anchor set became sufficiently long. Conversely, some researcher suggest that there is no rule of thumb guiding anchor test selection.

Classical test theory equating method is a traditional method that have been in practice before the advent of IRT method of equating, several researchers have done much on test equating using CTT equating methods among which is the study carried out by Demir and Guler (2014) titled “Study of test equating on the common item non-equivalent group design” this research was aimed at testing the statistical equivalence of different forms of a test which are administered at the same time using non-equivalent neat design. The data collected from the 761 students of 15 age group who had answered the 3rd and 10th booklets of the science studies literacy test were analyzed through Tucker Linear equating, Levine linear equating, frequency prediction and Braun-Holland linear equating methods. The weighted mean error squares averages indices that were obtained through equating procedures were 0.046 for the Tucker- linear equating, 0.072 for the Levine- linear equating, 0.049 for frequency prediction, and 0.034 for the Braun-Holland linear equating. It was observed based on the Weighted Mean Squares Error (WMSE)coefficient that the Braun-Holland linear equating method was the most appropriate for the equating of booklets 3 and 10 in the PISA 2009 Science Studies sub-test.

In another research conducted by store (2013), titled “Item Parameter Changes and Equating” in this study, the conditioned for the four test designs were those that had moderate difficult items with reasonable or constricted difficulty variability. These conditions registered low bias and RMSE values. With the increase in item difficulty, it is expected that low ability examinees will incorrectly respond to items that are higher than their ability level. On the other hand, high ability examinees will get all the items correct because they are still lower than their ability level. In similar manner LaFlair, Daniel, Nicolas, Maria and Joan (2015) “examined the results of equating in small-scale language testing programs”. This study compared seven different classical test equating methods these are “mean”, “linear Levine”, “linear Tucker, “chained equipercentile”, “circle-arc”, “nominal weights mean”, and

"synthetic". A nonequivalent AnchorTest (NEAT) design was used to compare two listening and reading test forms based on small samples. The equating methods were evaluated based on the amount of error they introduced and their practical effects on placement decisions. It was found that two types of error (systematic and total) could not be reliably computed owing to the lack of an adequate criterion; consequently, only random error was compared. Among the seven methods, the circle-arc method introduced the least random error as estimated by the Standard Error of Equating (SEE).

In a similar study carried out by BurhanettinOzdemir (2017) titled " Equating TIMSS Mathematics Subtests with Nonlinear Equating Methods Using NEAT Design: Circle-Arc Equating Approaches" the purpose of the study was to equate Trends in International Mathematics and Science Study (TIMSS) mathematics subtest scores obtained from TIMSS 2011 to scores obtained from TIMSS 2007 form with different nonlinear observed score equating methods under Non-Equivalent Anchor Test (NEAT) design where common items are used to link two or more test forms. The results obtained from chained and frequency estimation based on equipercentile equating methods were compared to four different methods (Tucker, Levine, Braun-Holland and chained) based on a new nonlinear equating approach called circle-arc equating in order to see which method was the most appropriate for equating these forms. The results of different nonlinear equating methods were compared with respect to Root Mean Squared Error (RMSE) index, mean of bootstrap standard errors (MBSE) and mean of bootstrap bias. Results indicates that TIMSS 2007 mathematics tests were easier than TIMSS 2011 mathematics across the score scale which indicates that results were biased against to students participated to TIMSS 2007. Moreover, equating methods based on nonlinear circle-arc equating outperformed the equipercentile equating methods and presmoothing decreased both standard error and bias associated with each method.

The anchor set in the NEAT design can be either internal or external according to the contribution of the anchor items to the examinee's total score on the test. In practice, anchor items can be located anywhere on a test form and may or may not contribute to students' scores. Anchor items that are scattered throughout the test form and contribute to students' score is often called an embedded, internal anchor set. Anchor items that appear together as a block at the end of the test form and contribute to students' scores are sometimes referred to as appended, since the anchor set appears at the end of the test. In either case, the items are internal anchors because they appear as part of the test form and contribute to students' scores. Finally, anchor items that appear as a separate form or testing session and do not contribute to students' scores are referred to as an appended, external anchor set. (Joseph Ryan and Frank Brockmann 2018)

Test equating is an integral aspect of the test development process that ensures the comparability of scores across parallel forms. In order to practically construct multiple forms of a test that are strictly parallel, equating is necessary to adjust the test construction process. As such, it is expected that the various equating procedures should accurately estimate the relationship between scores on two parallel forms of tests. Otherwise, the consequences of not establishing equating evidence may be an effect on the reliability, validity and usability of the scores or test. Given the importance of test equating, stakeholders and some educators in countries like Nigeria will have to continue to make a direct comparison of the scores across cohorts to establish if the educational system is achieving its goals.Evidences from the Literature have shown that research using empirical data have provided little information on the consequences of not equating educational scores across years or occasions. Especially in

Nigeria, there have been little known research to explicitly show that decisions made based on scores that are not equated may be imperfect as well as that the examination standards for such tests have remained invariant over the years. As such, test practitioners are usually faced with the problem of fair comparison and the integrity of scores across different forms. Therefore, for high stakes examinations bodies such as the WAEC, NECO, NABTEB and JAMB, their tests items require further investigation.

In an effort to investigate or establish comparability of test score or test equating results, different kinds of transformation such as concordance, calibration methods had been proposed within the classical test theory. However, the current research is set to use the CTT equating methods to establish the statistical equivalent of the two forms of the tests by comparing the test scores across sub set of equating methods to ascertain the robustness of the sub-set of CTT equating method, hence this study.

This study seeks to compare the sub set of classical equating methods using non-equivalent anchor test. The specific objectives of the study are to;

a. estimate the item parameters of the two forms of the non-equivalent anchor test to be equated among Osun State Secondary School Students;
b. examine the relative effectiveness of sub set of Classical Test Theory equating methods (mean equating, linear equating and equipercentile equating)in the non-equivalent anchor tests to be equated

Research Questions:

In order to achieve the objectives stated above, the following research questions were raised:

1. What are the item parameters of the two forms of the non-equivalent anchor test to be equated?
2. What is the relative effectiveness of sub set of Classical Test Theory equating methods (mean equating, linear equating, and equipercentile equating) in the non- equivalent anchor tests to be equated

Methodology

The study adopted the Non-Equivalent Anchor Test (NEAT) design. The population of the study comprised 137,083 Secondary School Students in Osun State whose average age was 15 years. Therefore, the target population consisted of all Senior Secondary School 2 (SS2), in both public and private schools. The population sample for the research comprised of 1080 students. The students were selected using simple random sampling technique. From the population sample male were 475 representing (44%) and the female were 605 representing (56%) of the sample population. Two adapted instruments titled Mathematics Achievement Test Form A (MATFA) and Mathematics Achievement Test Form B (MATFAB) were used to collect data for the study. The instruments were adapted from 2014 and 2015 West African Examination Council Mathematics objective items which served as MATFA and MATFAB respectively. The 2014 National Examination Council Mathematics objective test items were adopted. This is made up of 15 items representing (30%) of the total items in each of the forms. This served as the anchor items for both form A and B respectively. The reliability of the two instruments were established to be 0.79 for MATFA, and 0.75 MATFAB using Kuder-Richardson 20(KR-20) formula. Data collected were analyzed using IRTEQ software package, Common Item Program for Equating (CIPE) and R version (3.4.1) software.

Results

Research Question One: What are the item parameters of the two forms of the non-equivalent anchor test to be equated?

In order to answer this research question, the Common Item Program for Equating (CIPE) softwarewas used to calibrate the responses of 1080 examinees to 50-items multiple choice mathematics test forms A and B respectively. The results are presented in Table 1 which shows the Item Parameter Estimates (form A and form B) and the parameter estimates of the anchor items contained in the two forms for CTT (P represents the difficulty of the items and D Represents the discrimination of the items)

Table 1: Item Parameter Estimates of form A and form B

	CTT			
	Form A		Form B	
ITEM	Difficulty	Disc	Difficulty	Disc
1	0.54	0.09	0.54	0.09
2	0.81	0.11	0.81	0.11
3	0.75	0.29	0.75	0.29
4	0.88	0.13	0.88	0.13
5	0.51	0.59	0.51	0.59
6	0.87	0.01	0.87	0.01
7	0.65	0.34	0.65	0.34
8	0.79	0.28	0.79	0.28
9	0.79	0.43	0.79	0.43
10	0.81	0.22	0.81	0.22
11	0.17	0.02	0.17	0.02
12	0.61	0.63	0.61	0.63
13	0.52	0.09	0.52	0.09
14	0.72	0.28	0.72	0.28
15	0.57	0.27	0.57	0.27
16	0.24	0.21	0.24	0.21
17	0.46	0.41	0.46	0.41
18	0.67	0.29	0.67	0.29
19	0.63	0.48	0.63	0.48
20	0.77	0.46	0.77	0.46
21	0.7	0.38	0.7	0.38
22	0.7	0.28	0.7	0.28
23	0.37	0.01	0.37	0.01
24	0.66	0.66	0.66	0.66
25	0.72	0.08	0.72	0.08
26	0.02	-0.02	0.02	-0.02
27	0.54	-0.17	0.54	-0.17
28	0.64	0.24	0.64	0.24
29	0.67	0.4	0.67	0.4
30	0.33	-0.3	0.33	-0.3
31	0.47	0.24	0.47	0.24
32	0.25	-0.12	0.25	-0.12
33	0.75	0.57	0.75	0.57
34	0.62	0.09	0.62	0.09
35	0.51	-0.02	0.51	-0.02
36	0.39	0.05	0.39	0.05
37	0.41	0.46	0.41	0.46
38	0.32	-0.03	0.32	-0.03
39	0.42	0.31	0.42	0.31
40	0.23	-0.23	0.23	-0.23
41	0.45	0.52	0.45	0.52
42	0.46	0.46	0.46	0.46
43	0.19	0.33	0.19	0.33
44	0.12	-0.21	0.12	-0.21
45	0.3	0.07	0.3	0.07
46	0.6	0.52	0.6	0.52
47	0.37	0.45	0.37	0.45
48	0.13	0.16	0.13	0.16
49	0.62	0.47	0.62	0.47
50	0.09	0.06	0.09	0.06
MEAN	0.52	0.23	0.52	0.23
STD	0.22	0.24	0.22	0.24

Table 2: Presents the Parameter Estimates of the Anchor Items Contained in the two Forms

	CTT			
	P	D	P	D
IT3	0.75	0.29	0.23	0.37
IT6	0.87	0.01	0.33	0.49
IT9	0.79	0.43	0.52	0.36
IT13	0.52	0.09	0.45	0.43
IT15	0.57	0.27	0.21	0.34
IT18	0.67	0.29	0.42	0.28
IT23	0.37	0.01	0.23	0.02
IT26	0.02	-0.02	0.15	0.07
IT29	0.67	0.40	0.24	0.19
IT32	0.25	-0.12	0.29	0.13
IT36	0.39	0.05	0.22	0.17
IT39	0.42	0.31	0.26	0.22
IT42	0.46	0.46	0.26	0.17
IT46	0.60	0.52	0.30	0.44
IT49	0.62	0.47	0.26	0.35
MEAN	0.53	0.23	0.29	0.27
STD	0.22	0.21	0.10	0.14

Table 2 shows the parameters of the common items estimated in the two independent samples. The estimates presented in the table were obtained using Classical Test Theory and Item Response Theory. The table shows that under CTT, difficulty and discrimination (P, X = 0.53, STD = 0.22; D,X = 0.23, STD = 0.21) of the common items in Form A were respectively different from the values of the difficulty and discrimination of the common items in Form B (P, X = 0.29, STD = 0.10; D, X = 0.27, STD = 0.14) were different from one another.

Furthermore, the table shows that under IRT framework, the discrimination, difficulty and vulnerability to guessing of the common items estimated in Form A (a, X = 1.21, STD = 0.94; b, X = 1.28, STD = 2.67 and c, X = 0.30, STD = 0.20) were different from the respective values of the estimates under Form B (a, X = 0.15, STD = 0.34; b, X = 2.79, STD = 3.40; and c,X = 0.18, STD = 0.08). These results shows that there were differences in the characteristics of the commons items. Therefore, the rescaling of Form A item parameter estimates to the scale of Form B is needed.

Research Question Two: *What is the relative effectiveness of sub-set of Classical Test Theory equating methods (mean equating, linear equating, equipercentile equating)i n the non- equivalent anchor tests to be equated?*

In order to answer this research question, the test scores emanating from the tests were linked with the aid of CIPE Package. To achieve this, Form A test scores were transformed to the scale of Form B test using linear equating, mean equating and equipercentile equating methods. According to Kolen and Brennan (2014), linear equating is represented by

Form B test using linear equating, mean equating and equipercentile equating methods. According to Kolen and Brennan (2014), linear equating is represented by

$l_y(x) = \frac{\sigma(Y)}{\sigma(X)}x + [\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)]$ .....(Eqn.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 eqn 1 becomes

$l_y(x) = Ax + B$.....(Eqn.2)

In practice, there are three methods used in equating test scores under the linear equating approach: Tucker's linear equating method (TLIN), Levine linear equating method (LLIN) and Braun-Holland linear method (BLIN)

For the mean method of equating, the function for equating is expressed as

$m_y(x) = x + [\mu(Y) - \mu(X)]$ .....(Eqn.3)

In equation 3, the coefficient of x (i.e., 1) is the slope and $\mu(Y) - \mu(X)$ is the intercept. On substitution, eqn 3 becomes

$m_y(x) = Ax + B$.....(Eqn.4)

To identify the most effective methods of equating used under CTT, the standard error of the equating methods were obtained and compared. The most effective method, according to Kolen and Brennan (2014), is the equating method that produced the smallest standard error of equating. Standard error of equating is obtained by taking the standard deviation of the equated scores under CTT framework. The standard errors of equating of the mean, linear and equipercentile equating methods are presented in table 3

Table 3: Standard error of Equating of the Equating Methods of CTT

Sub set of CTT equating methods	Std. Error of equating
EQUATED_SCORE_TMEAN	5.29
EQUATED_SCORE_LMEAN	5.29
EQUATED_SCORE_BMEAN	5.29
EQUATED_SCORE_TLIN	5.49
EQUATED_SCORE_LLIN	5.56
EQUATED_SCORE_BLIN	6.82
EQUATED_SCORE_EQUI	6.98

Table 3 shows that under CTT measurement framework, TMEAN, LMEAN and BMEAN are the most effective method of equating with (STD error of equating = 5.29), follow by TLIN (STD error of equating = 5.49) , LLIN method (STD error of equating = 5.56), BLIN (STD error of equating = 6.82) and then by Equipercentile method (STD error of equating = 6.98). These results shows that the mean methods of test score equating were more effective than the linear and equipercentile methods of equating under the CTT equating approach.

Discussion of the findings

The result of the study shows that under CTT, difficulty and discrimination of the common items in Form A were respectively different from the values of the difficulty and discrimination of the common items in Form B.Based on the differences observed in the characteristics of the commons items, mainly the difficulty index, there is need for rescaling of Form A item parameter estimates to the scale of Form B. This finding is in tandem withthe observation of Store (2013) where it was discovered that the best study conditions for the four

test designs were those that had moderate difficult items with reasonable or constricted difficulty variability. These conditions registered low bias and RMSE values. With the increase in item difficulty, it is expected that low ability examinees will incorrectly respond to items that are higher than their ability level.

Also, the result of the second objective of the study establishes therelative effectiveness of Classical Test Theory equating methods in the non- equivalent anchor. To identify the most effective methods of equating used under CTT, the standard error of the equating methods were compared. The most effective method according to Kolen and Brennan (2014), is the equating method that produced the smallest standard error of equating. It was established in this present study that under CTT measurement framework, TMEAN, LMEAN and BMEAN were the most effective method of equating followed by TLIN and by LLIN method. Finally, by equipercentile method, it can be established empirically that the mean methods of test score equating are more effective than the linear and equipercentile methods of equating under the CTT equating approach. This is contrary to the study carried out by Demir and Güler (2014) while examining the statistical equivalence of different PISA 2009 science tests administered at the same time with different equating methods under NEAT design. In the study,Program for International Student Assessment (PISA) 2009 science tests were equated with Tucker linear equating, Levine linear equating, and frequency prediction and Braun-Holland linear equating methods. The author found that among these linear equating methods, Braun-Holland linear equating method was the most appropriate for PISA 2009 science tests which is contrary to the results of the present study. In similar manner, in the study carried out by Burhanettin (2017) titled Equating TIMSS Mathematics Subtests with Nonlinear Equating Methods Using NEAT Design: Circle-Arc Equating Approaches when compared to the circle-arc (Tucker, chained, Levine and Braun-Holland) equating methods, results indicates that the new circle-arc methods outperformed the equipercentile methods and yielded more consistent results with smaller MBSE, RMSE and bias values.

Conclusion

The study concluded that mean equating methods were more effective compared to linear and equipercentile equating methods and also different forms of tests should not be equated when the difficulty index of the tests are different. In a nut shell, the major contribution of the current study is that it provided comprehensive information concerning the performance of the CTT equating methods.

Recommendations

- The characteristics of the common's items which served as the link between the two forms of the test should be rescaled, if difficulty index of the two forms are different so that the two forms will be on the same scale.
- Test scores equating methods should be used to standardize students' mathematics examination scores. Particularly mean equating methods follow by linear and equipercentile equating methods under CTT equating method.
- Statistical equivalence of mathematics test items should be checked with appropriate test theory model before administration

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 the comparison of studentscontinuous assessment measure*. (Unpublished Ph.D. Thesis), Retrieved January, 1, 2018

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed., pp. 508 -600). Washington, DC: American Council on Education. (Reprinted as W.H. Angoff, Scales, Norms, and Equivalent Scores. Princeton, NJ: Educational Testing Service, 1984).

Blueprint an online newspaper, https://blueprint.ng/waec-to-register-candidates-from-ss1/, Retrieved August, 6, 2018

BurhanettinOzdemir (2017) Equating TIMSS Mathematics Subtests with Nonlinear Equating Methods Using NEAT Design: Circle-Arc Equating Approaches. *International Journal of Progressive Education*, 13 (2) 116-132

Cook, L. L., &Eignor, D. R. (1991). IRT equating methods. *Educational measurement: Issues and practice*, *10*(3), 37-45.

Demir, S., &Güler, N. (2014). Ortakmaddelidenkolmayangruplardesenineilişkin test eşitlemeçalışması. *International Journal of Human Sciences*, 11(2), 190-208. doi: 10.14687/ijhs.v11i2.2870

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.

Fitzpatrick, A. R. (2008). NCME 2008 presidential address: The impact of anchor set configuration on student proficiency rates. *Educational Measurement: Issues and Practice*, 27(4), 34-40.

Hambleton, R.K.Swaminathan, H.& Roger, H. J. (1991). Fundamentals of item response

Huggins, Anne C., (2012). The effect of differential item functioning on population invariance of item response theory true score equating" open access dissertations.724. retrieved from http://scholarlyrepository .miami.edu/oa dissertations/724

Joseph R & Frank (2018) A Practitioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory

Keller, L. A., Egan, K. L., & Schneider, M. C. (2010). Item parameter drift in anchor items-detection and consequences: An analysis of simulated and operational test data. CTB/McGraw-Hill: Monterey, CA.

Kolen M.J., & Harris, D.J. (1990). A comparison of item pre equating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*, 27, 27 - 39.

Kolen, M.J. & Brennan, R.J. (1995). Test equating: Methods and practices. New York: Springer- Verlag.

Kolen, M.J., Brennan, R.L. (2004). Test Equating, Linking, and Scaling: Methods and Practices, 2nd ed. Springer-Verlag, New York.

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

LaFlair, G. T., Isbell, D., May, L. N., Gutierrez Arvizu, M. N., & Jamieson, J. (2017). Equating in small-scale language testing programs. *Language Testing*, *34*(1), 127-144.

McKinley, R. L., &Reckase, M. D. (1981). *A Comparison of Procedures for Constructing Large Item Pools* (No. RR-81-3). Missouri univ-Columbia Tailored Testing Research lab.

Nworgu, B. G. (2011). Differential item functioning: A critical issue in regional quality assurance. *Journal of Educational Assessment in Africa*, *6*, 112-123.

Store (2013) *Item Parameter Changes and Equating: An Examination of the Effects of Lack of Item Parameter Invariance on Equating and Score Accuracy for Different Proficiency Levels*. Unpublised Ph.D thesis, The Graduate School at The University of North Carolina at Greensboro.
theory. Newbury Park. C.A: Sage

Vale, C. David, Vincent A. Maurelli, Kathleen A. Gialluca, David J. Weiss, and Malcolm James Ree. *Methods for Linking Item Parameters*. Assessment Systems Corp St Paul Mn, 1981.

van Davier, A. A., Holland, P.W. & Thayer, D.T. (2004). Statistics for social science and public policy: The kernel method of test equating. New York: Springer Verlag.