EQUATING 2015 AND 2016 BASIC EDUCATION CERTIFICATE EXAMINATION ON CIVIC EDUCATION USING CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY IN OYO STATE, NIGERIA

By

TERHEMBA GODWIN ATSUA
Institute of Education,
University of Ibadan, Ibadan,
Oyo State.

IFUNANYA V. UZOESHI
Institute of Education,
University of Ibadan, Ibadan,
Oyo State.

PEACE OLUDI
Institute of Education,
University of Ibadan, Ibadan,
Oyo State.

And

EBERE SAMPSON WAGBARA
Community Secondary School, Rumuolumeni,
Port Harcourt, Rivers State.

Abstract

Classical Test Theory (CTT) and Item Response Theory (IRT) represent two frameworks used for test equating.They have provided researchers and examining bodies the choice of working within CTT or IRT or both frameworks. Several identified theoretical short comings of CTT have limited its applications. IRT is therefore, recommended for this task because of its theoretical advantages. These claims have rarely been empirically investigated. Hence, the studycompared CTT and IRT using the linear equating method. Five research questions guided the study. Survey design using single group equating method was adopted. The

target population was JSS III students in Ibadan North. The total population as at the time of the study was unknown because not all students had registered for the Basic Education Certificate Examination (BECE). However, ten schools in Ibadan North were selected using simple random sampling technique. From the sampled schools, six hundred and nineteen students participated in the study.Instrument used was 2015 and 2016Civic Education BECE tests. Data was analysed using BILOG-MG3 and IRTPRO software. Findings revealed that the equating functions used in the conversion of 2015Civic Education BECE test scores to the scale of 2016 under CTT and IRT were not the same. However, CTT and IRT approach produced similar results. It was concluded that both theories are good approaches for test equating. Both CTT and IRT approach were recommended for test equating.

Keywords: Test Theories, Linear Equating Model

Classical Test Theory (CTT) and Item Response Theory (IRT) represent two frameworks used for test equating.They have provided test developers and examining bodieswith the choice of working within CTT or IRT or both frameworks. Among the applications of the two frameworks is test equating. Test equating is a statistical procedure that is used in adjusting the difficulty difference in tests. It employs several approaches under the two contrasting frameworks thereby leaving the measurement community with most appropriate approach for equating a test. These methods are generally subsumed under CTT and IRT frameworks. They include: linear equating, mean/sigma equating, equipercentile equating, the Stocking-Lord Test Characteristics Curve (Stocking-Lord TCC) method, the Haebara TCC method, and the concurrent calibration among others. While linear, mean, and equipercentile equating methods are common to both CTT and IRT frameworks, the Stocking-Lord Test Characteristics Curve (Stocking-Lord TCC) method, the Haebara TCC method, and the concurrent calibration apply only to the IRT framework. Since the focus of this study was on the comparison of CTT and IRT frameworks, emphasis was on linear equating method.

Under CTT and IRT frameworks, linear equating is achieved by setting the means and standardized scores of the two forms of tests equal (Kolen, 1988; Cook & Eignor, 1991). In this method, Form 1 scores are converted so as to have the same mean and standard deviation as scores on Form 2.This conversion is achieved by setting the standardized item parameter estimates/scores of Form1 and Form2 equal. Mathematically, this is expressed as:

$$\frac{x_1 - \mu_1}{\sigma_1} = \frac{x_2 - \mu_2}{\sigma_2} \quad \ldots \text{Eqn. (1)}$$

Where: σ_1 and σ_2 are the standard deviations of Form 1 and Form 2; μ_1 and μ_2 are the means of Form 1 and Form 2; x_1 and x_2 are scores on Form 1 and Form 2 respectively. Finding Form 2 equivalent of Form 1 requires making x_1 the subject of equation 2, this gives:

$x_1 = \frac{\sigma_1}{\sigma_2} x_2 + [\mu_1 - \frac{\sigma_1}{\sigma_2} \mu_2]$ …Eqn. (2)

Where $\frac{\sigma_1}{\sigma_2}$ = slope (A), $[\mu_1 - \frac{\sigma_1}{\sigma_2} \mu_2]$= Intercept (B)

Therefore, equation 2 becomes:

$x_1 = Ax_2 + B$ …… Eqn.(3)

This equation represents the model for placing Form 2 on the same scale of Form 1 (Kolen, 1988; Cook & Eignor, 1991).

Equating test requires the conversions to be independent of the group used. This requirement according to Cook and Eignor(1991) potentially discredits CTT methods of equating, as test scores obtained under the CTT framework are sample dependent. An alternative is the IRT equating methods which are ability score based and sample independent(Cook & Eignor, 1991). As a result, it does not matter if an examinee takes an easy test or a hard form of a test. The examinee's ability estimate developed from either test form will be identical, within measurement error (Meyer & Zhu, 2013) provided the parameter estimates for both forms have been placed on the same IRT ability scale (Rupp & Zumbo, 2006; Bond & Fox in Meyer & Zhu, 2013).

Three different scenarios can be conducted when equating test using Classical Test and Item Response Theories. These scenarios are referred to as test equating designs. Equating designs refers to the various methods or frameworks of collecting data for equating analysis. Generally there are three equating designs. They are: single group or common subject design, random or equivalent group designs and the common item non-equivalent group designs (Kolen, 1988; Dorans, Moses & Eignor, 2010). In common subject equating, the alternate parallel form is administered to all examinees in a given sample. The random or equivalent groups design is characterized by randomly selecting two equivalent groups from a population, and administering one form to the second group. Lastly, the non-equivalent group design is executed by administering each of the forms of a test containing common items to different groups of examinees. The equating design has advantages and disadvantages that make it more or less useful for different situations. For example, the single subject design requires the smallest sample sizes and the equivalent group designs require that largest sample sizes to achieve the same level of accuracy (Dorans et al, 2010). Therefore, in this study, the single group design was used.

Empirical evidence on the effectiveness of CTT approach to test equating is largely unknown. This perhaps may be due to the fact that testing in the developed nations is done within IRT framework.The focus of studies in test equating is solely on the IRT framework. In fact Nworgu and Odili, (2005) pointed out that since many large assessment programs use IRT models to develop and calibrate tests, the use of IRT

based equating is often a logical choice. Among studies that compared CTT and IRT based equating methods are (Ogbebor, 2012; Ogbebor & Onuka, 2013; Ogbebor, 2017).For instance, Ogbebor (2017) found that the two approaches to test equating produced similar results. Nworgu and Agah(2012) found similar results with linear equating, concurrent calibration (CC) using the Rasch model and the threeparameter IRT model, and separate calibration using the three parameter IRT model with fixed common item parameter (FCIP) equating and mean/sigma (M/S) equating.It therefore, becomes imperative to test the tenability of the hypothesis that the two frameworks produced nearly the same results.

Olatunji (2015) analysis of linear and equipercentile equating of senior school certificate examination Economics multiplechoice test found that the different methods produced similar results when the tests to be equated were parallel and the groups in the two years were equivalent. Olabode and Adeleke (2015)comparative analysis of item local independence of WACE and NECO 2012 Mathematics objectives test items using different linear equating methods, the IRT 1PL and3PL found that the different methods produced similar results. Zannu (2016)equated 2012 and 2013 Physics test of unified tertiary matriculation examination (UTME) of the Joint Admission and Matriculation Board. Analysis of data revealed that the test items were of average difficulty and discrimination but the mean difficulty and discrimination index of the 2013 form of the test was greater than that of 2012. A poor correlation was observed between the difficulty and discrimination indices of the two forms of the tests. Also a correlation observed between the examinee score was found to be moderate.

Bichi and Bichi (2016) analysed dichotomously scored science achievement test items using item response theory framework. The findings of the study revealed that the test measured a single trait thus satisfying the condition of unidimensionality. Similarly, the Person Chi-square goodness of fit test revealed that the two parameter IRT model was more suitable since no misfit item was observed and the test reliability was .86. The mean examinee ability was set at 0 (SD = 0.61). The mean item difficulty was 0.48 (SD =0.95) and mean item discrimination was 0.91 (SD = 0.19). 15(37.5%) items were identified as problematic having failed to meet the set criteria and 25(6.5%) were good.

Adegoke (2016)determined the comparability of WAEC 2010 and 2011 Physics objectives using linear equating methods. Findings revealed that the relationship between the two was low and was not statistically significant. When the mean of the examinees' ability was compared, the result confirmed that items in form one were relatively more difficult than form two. Person correlation coefficient of the liner relationship between the two forms of the tests showed that the two forms of the tests were linearly related, however, the value of the relationship was low.

Introduction of IRT has provided test developers and examining bodies with the choice of working within either CTT or IRT or the combination of the two frameworks. Several identified theoretical short coming of CTT have limited its application to test

equating in particular and IRT has always been recommended for this task because of its theoretical advantages over CTT. However, these claims have rarely been investigated empirically, thus, they are largely unknown. The inference drawn from the results of the studies reviewed is that findings are inconsistent. Furthermore, none has compared the equating methods using (a) data that consist of dichotomously and polytomously scored items; (b) data that are not necessarily normally distributedand (c) nonequivalent groups.It is on the basis of these that the study compared CTT and IRT equating methods using the linear equating method. The study therefore answered the following questions:

i. Are the test items of BECE 2015 and 2016 Civic Education unidimensional?

ii. To what extent are the test items of BECE 2015 and 2016 Civic Education locally independent?

iii. To what extent do the test itemsof BECE 2015 and 2016 Civic Education fit IRT assessment of model-data fit?

iv. How comparable are the equating function used in placing the scores of 2015 and 2016Civic Education BECE tests using CTT and IRT?

v. How comparable are the equated scores of 2015 and 2016Civic Education BECE tests obtained under CTT and the ones obtained under IRT?

Methodology

The study adopted the single group equating design of the survey type. The population of the study comprised all JSS IIIstudentstaking Civic Education in Ibadan North Local Government Area of Oyo State. Simple random sampling technique was used to select 10 schools in Ibadan North Local Government Area and all JSS III students taking Civic Education in the selected schools that participated in the study. The total sample of students that participated in the study was 619 in number.

Two instruments were adopted and used for data collection. The instruments were 2015 and 2016 BECE Civic Education objective tests developed by the National Examination Council. Since the test items were adopted from the examination body, they were assumed to have undergone standardization process.They were therefore adopted without modification. In the design the two tests are required to be administered to the same student. Therefore,the objective test items for 2015 and 2016 BECE Civic Education were administered on every student at two different occasions.

The process of data collection began with the presentation of an official letter of introduction from the researchers to the principals of the sampled schools. The purpose and modality for the study was discussed with the principals. The approval to conduct the study led to the second stage whereCivic Education teachers from each school were requested to serve as research assistant. The researchersmet with the Civic Education teachers and students in each school and created rapport with them, got them fully informed about the purpose and how to go about the administration of the instruments. The researchers and the teachers agreed on a time framewith the students to administer

the instruments. The instruments were administered on the students in their classroom and all were retrieved. Out of the six hundred and nineteen instruments administered and retrieved, five hundred and ninety eight were dully completed while the remaining twenty one were returned uncompleted.

The administered instruments were collated and scored using the predetermined scoring key. A correct answer to an item was scored 1 mark while a wrong answer was scored 0. The student's testsscoreswere subjected to CTT and IRT scoring framework. Under CTT, the total number of items answered correctly were obtained for all the students. This total score on the two tests were transformed and equated. Under IRT, the students' scores were subjected to IRT assumptions and model data fit to establish which IRT model best fits the test. Thereafter, the best model was used to estimate the ability parameter of the students on the two tests. For effective comparison of the scores with those estimated under CTT, the estimated IRT ability scores were converted to the number correct as the CTT. Data were analysed using IRTPRO and BILOG-MG3 software.

Results and Discussion

Research question one

Are the test items of BECE 2015 and 2016 Civic Education unidimensional?

To test theIRT assumption of unidimensionality of 2015 BECE Civic Education multiple choice test, Stout's test of essential unidimensionality implemented in DIMTEST 2.0 was used. To perform the test, items were divided into two sub-tests that were as dimensionally distinct as possible: the partitioning sub-test and the assessment sub-test. Items that formed a secondary dimension, the assessment sub-tests, were selected empirically using the HCA/CCPROX cluster procedure and DETECT statistic in DIMTEST, and the candidate cluster was tested to see if it was dimensionally distinct from the remainder of the test. The null hypothesis is that the responses are unidimensional (the average covariance within groups = 0).Therefore, failure to reject the null hypothesis indicates that the assumption of unidimensionality is tenable. A random sample of 30% of the examinees was used to select the assessment sub-test, and the remaining sample was used for the dimensionality test. T = 1.3334 (P-value = 0.0926, one-tailed).The assumption of unidimensionality was therefore, not rejected. This indicated that there was only one dominant dimension that accounted for the variation observed in students.

To test the IRT assumption of Unidimensionality of 2016 BECE Civic Education multiple choice test, Stout's test of essential unidimensionality implemented in DIMTEST 2.0 was used. To perform the test, items were divided into two sub-tests that were as dimensionally distinct as possible-the partitioning sub-test and the assessment sub-test. Items that form a secondary dimension and the assessment sub-test were selected empirically, using the HCA/CCPROX cluster procedure and DETECT statistic in DIMTEST, and this candidate cluster was tested to see if it was

dimensionally distinct from the remainder of the test. The null hypothesis is that the responses are unidimensional (the average covariance within groups = 0). So failure to reject the null hypothesis indicates that the assumption of unidimensionality is tenable. A random sample of 30% of the examinees was used to select the assessment sub-test, and the remaining sample was used for the dimensionality test. T = 1.6334 (P-value = 0.0842, one-tailed); therefore the assumption of unidimensionality was not rejected. This indicated that there was only one dominant dimension that accounted for the variation observed in students.

The findings of the study collaborates those ofOgbebor(2012); Ogbebor and Onuka,(2013), Bichi and Bichi (2016) and Ogbebor (2017) respectively. The result however disagreed with Marco et al. (1983)whose study found that the two approaches to test equating could not produce similar results. Ogbebor (2017) for instance found similar results with linear equating. In fact, Kolen and Brennan (1995) pointed out that since many large assessment programs use IRT models to develop and calibrate tests, the use of IRT based equating is often a logical choice. The implication is that examining bodies need to subject test items to IRT models during item development process. Hence, items that are subjected to IRT models are capable of measuring specific traits for test takers.

Research question two

To what extent are the test items of BECE 2015 and 2016 Civic Education locally-independent?

To test for assumption of local independence of 2015 BECE Civic Education multiple choice test items, inter-correlation matrix of the tetrachoric correlation was computed and the result is presented in Table 1.

Table 1: Summary of Tetrachoric Correlation for 2015 BECE Civic Education Multiple Choice Test Items

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	++	++	60
1	1															++	++	
2	0	1														++	++	
3	-0	0	1													++	++	
4	0.1	0	0.1	1												++	++	
5	-0	-0	-0	0.1	1											++	++	
6	0	0.2	0.1	-0	0.1	1										++	++	
7	0.2	0.1	0	0	-0	-0	1									++	++	
8	0.1	0	0	-0	-0	0	0.2	1								++	++	
9	0.2	0	0.1	-0	-0	0.1	0.1	0.2	1							++	++	
10	0.1	0	0.1	-0	-0	0	0.1	0.2	0.2	1						++	++	
11	0.1	0.1	-0	0	-0	0	0.2	0	0.1	0.2	1					++	++	

12	-0	-0	0.2	-0	-0	-0	0.1	0.1	0.1	0	0.1	1				++	++	
13	0.1	0	-0	0.1	-0	0.1	0.1	0.1	0.1	0	0.1	0	1			++	++	
14	0.1	0.1	0	0.1	0	-0	0	0.1	0.1	0.1	0	0.1	-0	1		++	++	
15	-0	-0	0	-0	0.1	0	-0	-0	0	0.1	-0	-0	0.1	-0	1	++	++	
16	0.1	-0	0.1	-0	0	0.1	0.1	0.1	0.2	0.1	0	0.1	0	-0	-0	++	++	
17	0.1	0	0	0	-0	0	0.1	0.1	-0	-0	0.2	-0	-0	-0	-0	++	++	
18	-0	-0	-0	0.1	0	0.1	0.1	0	-0	-0	-0	0.1	-0	0	-0	++	++	
19	0.3	0	-0	0	-0	0	0.2	0.1	0.1	0.1	0.1	-0	0	0.1	-0	++	++	
20	0	0.1	-0	-0	-0	0	0.1	0.1	0.1	0	0	0	0.2	0.1	-0	++	++	
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	++	
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	++	
60	**0**	**-0**	**0**	**-0**	**0.1**	**-0**	**0**	**0.1**	**-0**	**0**	**0**	**0.1**	**0.1**	**0**	**0.1**	**0**	**0**	**0**

Table 1 indicated that the inter-correlation matrix of the tetrachoric correlation for 2015 BECE Civic Education multiple choice test items observed among the items was 0.2 which is not greater than the minimum standard of 0.3 correlation coefficient set for adjudging an item to be locally dependent. Hence the assumption of item local independence was not violated by the test items.

Table 2: Summary of Tetrachoric Correlation for 2016 BECE Civic Education Multiple Choice Test Items

Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**++**	**++**	**60**
1	1.0															++	++	
2	0.1	1.0														++	++	
3	0.0	0.0	1.0													++	++	
4	0.0	0.1	0.2	1.0												++	++	
5	0.0	0.0	0.0	0.0	1.0											++	++	
6	0.1	0.0	0.0	0.1	-0.1	1.0										++	++	
7	0.2	0.0	0.0	0.0	0.0	0.1	1.0									++	++	
8	0.1	0.0	0.1	0.1	-0.0	0.0	0.1	1.0								++	++	
9	0.2	0.1	0.0	0.0	0.2	0.1	0.1	0.0	1.0							++	++	
10	0.1	0.0	0.1	0.0	0.0	0.0	-0.0	0.0	-0.0	1.0						++	++	
11	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	1.0					++	++	
12	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.1	1.0				++	++	
13	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.1	0.0	1.0			++	++	
14	0.1	0.0	0.0	-0.1	-0.1	0.0	0.1	0.1	-0.0	0.1	0.0	0.1	0.1	1.0		++	++	
15	0.0	0.0	0.1	0.0	-0.1	0.1	0.1	0.0	-0.1	0.1	0.1	0.1	0.1	0.0	1.0	++	++	

16	0.2	0.1	0.0	0.1	-0.1	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.0	**0.3**	0.0	++	++	
17	0.1	0.1	0.1	-0.1	-0.1	0.0	0.1	0.2	0.0	0.1	0.1	0.2	0.0	0.1	0.1	++	++	
18	0.1	0.1	0.1	0.1	0.0	0.0	-0.1	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.1	++	++	
19	0.1	0.0	0.1	0.0	-0.1	0.0	0.1	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	++	++	
20	0.1	0.1	0.1	0.1	-0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.1	++	++	
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	++	
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	++	
60	**0.0**	**0.0**	**0.0**	**0.1**	**0.1**	**0.1**	**0.0**	**0.0**	**-0.1**	**0.1**	**0.1**	**0.0**	**0.0**	**0.1**	**0.1**	**0.0**	**0.0**	**0.1**

A cursory look at Table 2 showed that the maximum tetrachoric correlation observed among the items was 0.3 which is not greater than the minimum standard of 0.3 correlation coefficientset for adjudging an item to be locally dependent. Hence the assumption of item local independence was not violated by the test items. The findingis consistent with Olabodeand Adeleke (2015) findings that the WAEC and NECO 2012 items were locally independent.The result supports that of Joshua, Ubi and Abang (2011) who found that UME Mathematics items for 2000, 2001, 2002 and 2003 years were to a great extent, locally independent. Oke(2012) study on item local independent in WAEC Economics found similar result that the items were locally independence. The result is consistent with that of Ogbebor (2017) finding when Mock examination and WAEC items were compared. The implication of the finding is that test items of any measurement instrument should be locally independent irrespective of the purpose and objective to which it is set to achieve.

Research question three

To what extent do the test items of BECE 2015 and 2016 Civic Education fit IRT assessment of model-data fit?

To determine if the test items of BECE 2015 and 2016 Civic Education fit IRT assessment of model-data fit, the items were subjected to binary response to test model fit. For binary response tests items, there are three models through which the item parameters and pearson parameters could be estimated: These are 1-PL, 2-PL, and, 3-PL. Chi-square likelihood ratio Goodness of fit statistics is used to test whether or not it is reasonable to model the tests items according to the one, two or three parameter logistic model. A Chi-square value with the smallest value indicates the best model that fit the data. To estimate this, the data was subjected to the three models and the result is presented in Table 3.

Table 3: Model-Data Fit Assessment for 2015 and 2015 BECE Civic Education Test Items

Model	-2 Log Likelihood Value for 2015	-2 Log likelihood Value for 2016
1PL	49856.8524	39551.8418
2PL	48234.2351	38214.2354
3PL	45172.9670	36205.3254

Result from Table 3 presents the model-data fit assessment of the 2015 and 2016 BECE Civic Education. Column one presents the parameter model while column two represent the Chi-square value of the data set for 1PL, 2PL, and 3PL that was used to calibrate the test. The values were obtained from phase two of BILOG MG software. The result showed that when the data set were modeled using 1PL, 2PL, and 3PL, the smallest Chi-square value was observed when the data set were modeled with the 3PL, hence the data set was modeled using the 3PL model. This finding aligned with that of Ogbebor (2017) who found that Mock Economics achievement test were modeled into the 1PL, 2PL, and 3PL and the smallest Chi-square value was observed. Similarly, Bichi and Bichi (2016) Pearson Chi-square goodness of fit test revealed that the data fit the two parameters IRT model, and IRT model was more suitable since no misfit item was observed. Oke(2012) found similar result when WAEC test items were modeled using 1PL, 2PL, and 3PL and the smallest Chi-square value was obtained when the data set were modeled with the 3PL, hence the data set was considered to fit the 3PL model. The implication of the result is that since the Civic Education BECE is meant to be a large assessment exercise and that since many large assessment programs use IRT models to develop and calibrate tests, the use of IRT-based equating is a logical choice.

Research question four

How comparable are the equating functions used in placing the scores of 2015 and 2016Civic Education BECE test under CTT and IRT?

Table4: Equating Function of Scores for 2015 and 2016 Civic Education BECE Test under CTT and IRT

	CTT	IRT
Slope	1.04097	1.08039
Intercept	-1.74789	-3.2233

Result from Table 4showsthat under the CTT approach, a slope of 1.04 emerged and under the IRT approach, a slope of 1.08 was obtained. Furthermore, the intercept emerging from the conversion of 2015 BECE Civic Education test score to the scale of 2016was -1.75 under the CTT approach and -3.22 under the IRT approach. The results

indicated that the equating functions used in the conversion of the 2015Civic Education BECE test scores to the scale of 2016Civic Education BECE test under the CTT framework and IRT framework were not the same. The findings of the study aligned with that of Zannu (2016) whose analysis revealed that the comparable differences in the test items were of the average difficulty and discrimination but the mean difficulty and discrimination index of the 2013 form of the test was greater than that of 2012. The result confirmed that of Ogbebor (2017) when CTT and IRT approaches were employed to comparable Mock Achievement test. Differences in the two approaches were found where low relationship in the parameter estimate was produced for CTT and IRT model.The implication of the finding is that the IRT model is most preferred to the CTT model because of the low relationship in the parameter estimate CTT produced.

Research Question five

How comparable are the equated scores of 2015 and 2016Civic Education BECE test under CTT and IRT?

In order to evaluate the comparison of the equated scores of 2014 to 2015 Civic Education BECE test under CTT and IRT the equating functions presented in Table 5 were used alongside equation:

$$x_1 = Ax_2 + B$$

In the equation x_1 is the 2015 test score of a particular student placed on the scale of 2015 score, and x_2 is the score obtained by the student in 2016 test. Table 6 presents the 2015 equivalent of 2016Civic Educationtest.

Table 5: Equated Scores of 2015 and 2016 Civic Education Tests

2015Score	CTT 2016Equivalent	IRT 2016Equivalent
22	21	
23	22	
24	23	
25	24	
26	25	
27	26	
28	27	
29	28	
30	29	29
31	31	30
32	32	31
33	33	32

34	34	34
35	35	35
36	36	36
37	37	37
38	38	38
39	39	39
40	40	40
41	41	41
42	42	42
43	43	43
44	44	44
45	45	45
46	46	46
47	47	
48	48	
49	49	
50	50	
51	51	

Result from Table 5 showed that the minimum score obtained by the students on the 2015BECE Civic Education test was 22. When this score was placed on the scale of 2016, it was found to be equivalent to a score of 21 on the 2016BECE Civic Education test when the CTT method was used. This same trend was repeated until a score on the 2015 BECE Civic Education test was equivalent to the score on the 2016BECE Civic Education For example, a score of 23, 24, 25, 26, 27, 28, 29, and 30 on 2015 were respectively equivalent to 22, 23, 24, 25, 26, 27, 28, 29 on 2016 edition of the test. However, a score of 31 on 2015 version of the test was equivalent to 31 on the 2016 version of the test. From this point (i.e., 31), the scores on 2015 were equivalent to the scores on 2016. Although for IRT methods the obtained scores started from 30, a score of 30 on 2015 version of the test was equivalent while 29 and 31 were equivalent to 30 and this trend continues until when a score of 34 on the 2015 edition of the test was equivalent to the same score. These results showed that the classical test theory and the item response theory approach to test produced similar results to a very large extent. The result revealed that the two tests are equated in their ability estimate. The study affirms the findings of Adewale (2015) who equated two year BECE results in Basic Science and Technology. The finding is in consonance with that of Ogbebor (2017)

whose finding revealed that placing the examinees ability score on a common scale gave almost the same ability estimate. This implies that linear method is somehow stable in producing equal results for a large number of different scores.

Conclusion

Conclusions drawn from the findings of the study are that, the use of classical test theory and item response theory are very essential in test equating. Although item response theory is more effective in equating test items as it enhances selecting of items that best measure student's ability giving adequate information concerning the behaviour of an item as well as the examinees. These features depict IRT as a better option in giving acceptable information regarding the behaviour of an item as well as the examinees. However, both CCT and IRT could be used for test equating.

Recommendations

Based on the findings of the study, the following recommendations were made:

i. Examining bodies should use both CTT and IRT approaches for test equating.

ii. Examination bodies that are working within the CTT framework should incorporate IRT framework into their test development process.

iii. Examination bodies should ensure that the test to be used should be equated with an existing test to ensure that they measurethe same construct.

References

Adegoke, B. A. (2016). Assessment of the comparability of WAEC 2010 and 2011 Physics objective tests. In B. A. Adegoke, O. Popoola & O. E. Babatunde (Eds.). *Public examining in sub-Saharan Africa: Issues, challenges and prospects* (pp27 – 38). Garki Abuja: Marveolus Mike Press Ltd.

Adewale, J.G.(2015).Equating two year BECE results in Basic Science and Technology in Oyo State, Nigeria. Available at www.buse.ac.zw/downloads/J.%20Gbenga%20Adewale.doc on 12th January, 2017.

Bichi, A. A. & Bichi, A. A. (2016). Analysis of dichotomously scored science achievement test items using item response theory framework. In B. A. Adegoke, O. Popoola & O. E. Babatunde (Eds.). *Public examining in sub-Saharan Africa: Issues, challenges and prospects* (pp51 – 66). Garki Abuja: Marveolus Mike Press Ltd.

Cook, L. L.& Eignor, D. R. (1991). An NCME module on IRT equating methods. *Educational Measurement, Issues and Practice, 10*(3), 191 – 199.

Dorans, N. J., Moses, T. P.& Eignor, D. R. (2010). Principles and practices of test score equating. A Research Report of Educational Testing Service RR-10-29.

Hills, J. R., Subhiya, R. G.& Hirsch, T. M. (1988). Equating minimum-competency tests: Comparison of methods. *Journal of Educational Measurement,25*(3), 221–231.

Joshua, M. T., Ubi, B.& Abang, I.(2011). Item local independence in selection examination in Nigeria: Implications for assessment for regional integration. An Unpublished Paper, University of Calabar.

Kolen, M.J. (1988). Traditional equating methodology. *Educational Measurement, Issues and Practice,7*(4), 29–36.

Kolen, M.J. & Brennan, R.L. (1995). *Test equating methods and practices*. New York: Springer.

Marco, G. L., Peterson, N. S. & Stewart, E. E. (1983). A test of adequacy of curvilinear score equating models. In D. White (Ed.). *New horizons in testing*(Pp147–177). New York: Academic.

Meyer, J. P. & Zhu, S. (2013). Fair and equitable measurement of student learning in massive open online courses (MOOCs): An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment,8*, 26–39.

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

Nworgu, B. G. & Agah, J. J.(2012). Application of three parameter logistic model in the calibration of a Mathematics achievement test. *Journal of Educational Assessment in Africa, 7.*

Nworgu, B.G. & Odili, J.N. (2005). Analysis of Oil SSCT Biology multiple choice test. *Review of Education, 16*(2), 140 – 152.

Ogbebor, U. C. (2012). Differential item functioning in Economics question paper of National Examinations Council in Delta State Nigeria. Unpublished M.Ed. Project, University of Ibadan.

Ogbebor, U.C. & Onuka, A.O.U. (2013). Differential item functioning as an item bias indicator. *Journal of International Educational Research, 4*(4), 367 – 373.*http://www.interesiouriials.org/ER.*

Ojerinde,D. (2013).Classical test theory (CTT) VS item response theory (IRT):An evaluation of the comparability of item analysis results. A Guest Lecture Presented at the Institute of Education, University of Ibadan on 23rd May.

Oke, W.N.(2012). Item local independence in WAEC Economics in Ajeromi-Ifelodun Local Government Area of Lagos State. Unpublished Master Project University of Ibadan.

Olabode, J.O.& Adeleke, J.O.(2015). Comparative analysis of item local independent of WACE and NECO 2012 Mathematics objectivetest items. *Journal of Educational Research and Development,2,*112 – 120.

Olatunji, D. S. (2015). Analysis oflinear and equipercentile equating of senior school certificate examination Economics multiple-choice papers in Kwara State, Nigeria. Unpublished Ph.D. Thesis, University of Ilorin.

Olufemi, A.S. & Oluseyi, I.A. (2015). Differential item functioning of senior secondary school uniform promotion English language multiple choice examination questions in Ekiti State. *International Advanced Journal of Teaching and Learning, 1*(2), 1 – 6.

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.

Zannu, B. G. (2016). Equating 2012 and 2013 Physics test of unified tertiary matriculation examination of the Joint Admission and Matriculation Board.In B. A. Adegoke, O. Popoola & O. E. Babatunde (Eds.). *Public examining in sub-Saharan Africa: Issues, challenges and prospects* (pp197 – 212). Garki Abuja: Marveolus Mike Press Ltd.