ASSEREN
Journal of Educational Research and Development
VOLUME 11, JUNE 2024 PAGE 81-92
ISSN: 2536-6899 E – ISSN: 2814 – 3248 Website: www.theasseren.org.ng

CONCURRENT VALIDITY OF THE SENIOR SCHOOL CERTIFICATE EXAMINATIONS CHEMISTRY ITEMS IN NIGERIA

Babatimehin, Temitope, Okunola, Aghogo G. & Ogunsanmi, Olawale Ayoola

Submitted: 16 October, 2023 Revised: 29 May, 2024 Accepted: 1 June, 2024

Abstract

The study utilised the scores of examinees in WAEC and NECO Senior School Certificate Examination (SSCE) chemistry items under CTT and IRT. Non-experimental design of descriptive research type was adopted. The population was 36,182 students that registered for 2017/2018 WAEC SSCE Chemistry in Osun State. Two instruments were used to collect data; Chemistry Achievement Test Types I and II which were the adopted versions of May/June WAEC and June/July NECO Paper I 2015 respectively. A sample of 1,105 students was randomly selected. Simple random sampling was used to select 5 from 10 Local Governments Areas in the 3 senatorial districts in Osun State. Purposive sampling procedure was used to select two co-educational (one public and one private) from each LGA (30 schools) and two co-educational Federal unity schools. Data were analyzed using CTT and IRT methods of scoring, mean, standard deviation, correlation matrix and scatter plots. Result indicated that students approximated scores were high in WAEC Chemistry Items (WCI) using IRT (X=22.23, SD=9.88) compared to CTT (X =19.65, SD=8.15). While students approximated scores in NECO Chemistry Items (NCI), using IRT was high (X =26.95, SD=11.69) compared to CTT (X= 24.55, SD = 9.32). Also, moderate correlation exists between scores of examinees' in WCI and NCI under IRT (0.61) and CTT (0.63). The study concluded that there was a fair concurrent validity between scores of examinees in these examinations under IRT and CTT. The study recommended that IRT method be utilise for scoring items and concurrent validity in item validation.

Keywords: Concurrent validity, Item Response Theory, Classical Test Theory

Introduction

Testing is a crucial component of instruction and learning. After the teaching-learning process is complete, it is utilized to grade the pupils. According to Afolabi (2012), a test is described as a predetermined set of items that are particular catalyst meant to elicit responses from participants and may be scored. The curriculum's development and improvement, including needs evaluation, challenges associated with learning, proficiency level, with differences of students, are all aided by test. It also serves as a teaching guide. Data collection is accomplished through testing. Its power is in the way the information is arranged and the tools provided by testing technologies for determining how accurate the information is. The Senior School Certificate Examination (SSCE), administered by the West African Examinations Council (WAEC) and the National Examinations Council (NECO), and the National Business and Technical Certificate Education (NBTCE), administered through the National Business and Technical Examination Board (NABTEB), are certification examinations that students in Nigeria are required to take after completing secondary school. These examinations are used to gauge how far students have come in realizing the aims of each subject's curriculum. These examination bodies' respective credentials are formally acknowledged as being identical in Nigeria. They might be utilized to land jobs in the proper corporate, private, and public service echelons. Additionally, candidates have the option of combining their scores from any two examination sessions. Whether or not a candidate will qualify to be admitted into tertiary institution in Nigeria and other countries is largely determined by the caliber of the certificates issued by any of these examination organizations.

The process of certification is a significant activity in the academic timetable in Nigeria because of the economic and social significance associated with SSCE alongside the opportunities to be admitted into tertiary institutions for the possessor of such credentials. With a comparable mandate and the use of comparable standardized examinations in evaluating the knowledge of students in a range of areas, there is the assumption that the items of the test, the rules for administration, awarding of scores, and scores interpretated are uniform. Despite this mandate, there are many types of criticism concerning the validity of the examinations carried out by these agencies from significant stakeholders. The following are only a few of the aspersions: Unequally balanced quality of examination items; discrepancy in scores; widespread exam paper leaks; congestion in exam rooms; and exam misconduct. Peters (2012) claims that from 2002 to 2012, certain Federal institutions rejected NECO results due to the NECO's subpar quality. According to Ahmed (2014), WAEC questions from 2011 to 2014 were of worse quality than NECO questions. When NECO and WAEC were compared, Ojerinde and Faleye (2005) claimed that the two examinations are equal. The researcher's concern was particularly about the imbalance in the difficulty if the items and the variance in students' scores among the critiques levelled at these assessment organizations.

According to Seyi and Clement (2012), NECO is reputed to be more difficult than WAEC. This claim may be true since students treat the WAEC test, which is the first one they take, seriously while treating the NECO exam more casually. Others believe that the NECO examinations are simpler than the WAEC because of their test questions. The fact that WAEC is conducted before NECO, is an indication that experience from the prior would have made a difference which invariably will help improve the performance of students in NECO. The belief in the superiority of one certificate over another have made examination results unreliable. For instance, the Osun State government, along with certain other States in Nigeria, has long helped students and parents with the cost of paying for WAEC SSCE registration. With NECO, it's different. This might be a sign that the State likewise thinks WAEC is better than NECO.

Test theory, according to MacDonald (1999), is only a group of mathematical concepts that codify and explain certain queries about creating and utilizing tests. Two of the most often used test-theoretical frameworks are IRT and CTT. The majority of methods, including CTT, were developed in the 1920s, when the majority of techniques were first developed. Theories of validity, reliability, objectivity, test analysis, item analysis, and other concepts are some of the component theories that make up this theory. The bulk of the processes were subsequently extended to incorporate educational examinations after initially being confined to psychological testing. Long established testing circumstances, whether in individual or group situations, the same items are administered to every element of a population, such as students applying for admission into colleges or for employment, are best suited for CTT (Natarajan, 2009). The examination taker may be presented these item sets using paper and pencil or a computer, according to Oyebola (2016).

Quantitative item analysis, can be achieved using CTT or IRT, uses a scientific technique (Ojerinde, 2013). A crucial phase in the test-creation process is item analysis (Hernandez, 2009). Ojerinde (2012) defines item analysis to be methods by which validity and usefulness of items can be assessed. Erguven and Erguven (2014), and Kline (2005), believe a test developer's first responsibility is to make sure results from test are dependable indications of test takers' ability, more importantly when judgments will be based on the results. This method is frequently carried out by using statistical analysis known as item analysis on examinees' score responses to trial versions of the items in a test (Erguven & Erguven, 2014; Kline, 2005). Krishnan (2013) argued that the goal is to help the instrument's designers improve the tool by modifying or deleting components that fall short of a minimally acceptable standard. Techniques in Standard item analysis include item evaluations complexity, indices of discrimination, and item distractors. Largely the dependence of item statistics is on the characteristics of the examinee sample utilised in carrying out the analysis.

Therefore, the population should be represented by the examinee sample that the test is meant for (Hambleton & Jones, 1993). This is a serious problem for test developers who use CTT.

In CTT, a bad item is defined as one with high item difficulty index ($p > 0.80$) and $p < 0.20$), or one with discrimination index that is very low (rpbis 0.20). These criteria (item difficulty index, item discrimination power and p-value are considered in choosing test items to be administered. The arithmetic average of a group of scores is known as the mean, which measures central tendency. The square of a set of scores' standard deviation represents variance, which is a measure of dispersion (variability) (Adegoke, 2012). The following equations are often used to estimate the mean and variance of a collection of scores:

$$mean = \frac{\Sigma fx}{\Sigma f} \text{ .... } [i]$$

$$S^2 = \frac{\Sigma fd^2}{N-1} \text{ .... } [ii]$$

where

Σ = sum

f = frequencies of occurrences of scores

S^2 = variance

d = mean deviation

N = Total number of sample (Adegoke, 2012; McCall, 1975).

Despite the fact that Metibemu (2016) emphasised that the means and variances of dichotomous test items can be estimated in the same manner as other means and variances, there are deviations that offer much simpler formulae, such as calculating the number of responders who responded rightly and number of responders did not endorse the keyed option. According to Kline' (2005) the test item's mean using these derivations indicate number of students that got the item right (denoted by p), where variance is calculated as the product of (p) and aggregate of students that responded wrong (denoted by q). The test items' mean and variance are represented symbolically in the manner provided by Metibemu (2016);

***(p)** = (number of test takers who successfully answered the question)/(number of test takers who attempted the question)…………………..[iii]*

Item Response Theory generally is a statistical approach concerning examinees, items in test, and test performance, also the relationship between performance and abilities under testing by the items, according to Hambleton and Jones (1993). As the name suggests, IRT links test takers' performance to the characteristics of the items. Ojerinde (2016) noted that relating item responses and underlying capabilities can be stated in many ways (models). One or more capabilities can underlie performance in test, plus item answers can be scored discretely, dichotomously or polytomously where item scores groupings might be sorted or not ordered. IRT places emphasis regarding item level information while CTT does so at the test level. According to Hambleton and Jones (1993), item analysis under IRT necessitates (i) using relatively difficult mathematical techniques (like maximum likelihood estimation) and large sample sizes to determine sample-invariant (independent) item parameters and (ii) applying goodness-of-fit standard to identify items falling short of a particular response model. Baker (2001) and Hambleton and Jones (1993) state that IRT item guideline for estimations are not dependent on the examinees' level of competence while answering to the item. Therefore, to calibrate item settings, test creators will not require a sample representation of the candidates intended for the final edition of the test. To assure accurate item parameter estimate, will however, require diverse and high candidate sample (Hambleton and Jones, 1993). According to Guler, Uyanik, and Teker (2014), the IRT proposes three distinct models with the description three parameter, two parameter, and one parameter models. IRT models are referred to as unidimensional models since they only take into account one attribute, or characteristic, of the examinee. The data must satisfy three fundamental assumptions in order to calibrate item parameters from examinee responses to dichotomous test items (data) using the IRT model. These include (a) unidimensionality (b) local independence (c) and item characteristic curves (ICC), which may be used to define each item individually (Wiberg, 2004). In the opinion of Guler et al. (2014), unidimensionality presupposes that each test item only measures one ability. According to Hulin, Drasgow, and

Parson (1983), breaking this assumption would provide seriously misleading findings since the unidimensionality idea mandates that every item on a test must assess one underlying construct of a candidate. Unidimensionality can be done using Cronbach analysis and Factor Analysis among others (Ojerinde & Ifewulu, 2012).

Dimiter (2012) asserts that measurements are not complete consistent and precise in the social, behavioural sciences, as well as in physical sciences. According to measurement terminology, a measurement's reliability increases with its precision and consistency. Meaning that the reliability of scores indicates the extent to which they do not have random error. Most importantly, the validity of score interpretations and data-driven judgments in counselling, education, and other sectors depend on the dependability of scores. Anastasi (1988) defines test's validity as a measure of both what it is measuring and how well it is doing it. According to Afolabi (2012), validity refers to how suitable or correct the interpretations drawn from test results in connection to its application. Numerous authors (Cronbach, 1951; Freeman, 1971; Field, 2005) have used the term "concept validity" to denote something similar.

The term "coefficients" refers to the estimated correlation coefficients between the test and the ideal criteria. The definitions make it very obvious that a comparison must be made using appropriate, independent criteria in order for validity to be established. The appropriate measure (criterion) must be correlated with the test in order to obtain the coefficient of validity. Following is a brief discussion of four validity forms. content, face, construct and criterion-related validity (concurrent and predictive). The term "content validity" describes how well items in a test reflect perfectly the subject matter of the test. This includes the language used and if it is appropriate for the reading level it is intended for. Are there enough items to adequately capture the construct-relevant and construct-underrepresented variation (Afolabi, 2012)? Whatever form of validity procedures concentrating on the connection between the test under validation and any other associated external test (s) with objectives that are similar are typically referred to as criterion-related validity. Both predictive and contemporaneous features are present. According to Cohen and Swerdlik (1999), the criterion-related validity means how well the score on a test predict a person's best advantageous standing on a certain measure of interest (criterion). According to Okpala, Onocha and Adedeji (1993), criterion-related validity is more specifically the degree to which test results (such as achievement test results) are consistent with current criterion measures (concurrent validity) or forecast future criterion measures (predictive validity).

When evaluating a test's predictive validity, we look at how well it can really predict something that it should theoretically be able to. The timing of the criteria scores is what distinguishes predictive validity from concurrent validity. Concurrent validity is concerned if the test and criteria scores are acquired almost at the same time. Okpala et al. (1993) claim that both concurrent and predictive validity can be achieved with the CTT measurement framework. A correlation value of 0.75 denotes a criterion-related valid test, according to Okpala et al. (1993).

The Pearson product moment correlation coefficient (r_{xy}) is a popular method of criterion-related validity reporting. However, there are a number of drawbacks to the correlation coefficient, which many researchers have chosen as the best method for providing internal consistency estimates of validity and reliability within the Classical Test Theory (CTT) measurement framework (Skurnik and Nuthall, 1968). Similar to this, Coffman (1971) made the case that using the correlation coefficient to create measures of internal consistency (validity and reliability) overstates the internal consistency since it disregards the means and standard deviations of scores. However, William (2000) noted that since seen test results are impacted by the unreliability of the tests, it is incorrect to interpret the validity coefficient as the correlation between observed test scores. Meadows and Billington (2005) discovered an alternative and additional measure of accuracy to the internal consistency (the standard error of measurement) in order to address the issues with correlation coefficient, as a measure of internal consistency coefficient as indicated above.

Moore, Notz and Flinger (2013) however suggested a very useful graph for showing relationship between two variables quantitatively measured for a particular candidate is a scatter plot. By this the values of one variable shows on the horizontal axis, and the vertical axis shows the values of the other variables. This way every individual in the data shows as a point on the graph. According to Cronbach and Meehl (1955), construct validity is how well a test measures the psychological feature of interest. Construct validity is supported by all evidence of validity, including content and criterion-related validity.

Construct validity gives proof of how much a test's score explains how consistently it measures the particular characteristic or construct that it purports to assess. Metibemu (2016) stated that construct and trait are unobservable behaviours of test subjects that are often assessed via a test. The specific attribute that a test assesses depends on its intended use. In reality, it is challenging, if not impossible, to validate a construct in its entirety. Campbell and Fiske (1959) in Afolabi (2012) and Dibu-Ojerinde (2012) create the following statistical approach for construct validation:

1. Testing for "divergence" between measurements or manipulations of related but conceptually separate behaviours or characteristics as opposed to;
2. Testing for "convergence" across multiple measures or manipulations of the same trait or behaviour. As a result, convergent validity occurs when a test shows a strong correlation when compared with a test measuring comparable qualities, and discriminant validity occurs when a test possesses a weak correlation when compared to a test measuring similar but conceptually unrelated features.

A significant problem that has to be addressed by WAEC and NECO is standard comparison in relation to equality in assessment instrument and objectivity in scoring and reporting results. Contrary to IRT, which may supply invariant item parameters and estimate examinees' abilities accurately, evaluation of item statistics is carried out at the item development stage under CTT. It is crucial to make test results similar because it is nearly impossible to construct parallel test forms. Findings from research, from other sources, show that the process of developing alongside the guidelines for scoring items in a test can positively influence students' performance or otherwise. In order to establish the equivalence of these examinations students' scores in WAEC and NECO Chemistry items are correlated in order to determine their concurrent validity, hence, this study

Research Objectives

1. determine the scores of the examinees on the chemistry items using CTT and IRT Scoring methods.
2. establish the concurrent validity of the SSCE chemistry items.

Research Questions

1. What are the scores of the examinees using CTT and IRT?
2. Is there a difference in the concurrent validity of WAEC and NECO chemistry examinees' scores?

Methodology

The research design used for the investigation was non-experimental design of descriptive research type. This design was adopted because answers to multiple choice WAEC and NECO Chemistry items were elicited from students who were sampled out of the entire 2017/2018 session Senior Secondary three students in the State. The population for the study comprised the 36,182 students registered to sit for Chemistry WAEC and eligible to write NECO SSCE in Osun State during 2017/2018 academic year. This constituted the research population of which 18,106 are males and 18,076 females. The use of a multi-stage sampling approach was adopted. From the three senatorial districts in Osun State having 10 Local Governments Areas (LGA) each, five LGAs, totaling 15, were chosen using a simple random sampling procedure. Purposive sampling was utilized to choose two co-educational schools (one public and one private) from each LGA (30 schools) and two co-educational Federal unity schools to ensure an all representation of schools using purposive sampling procedure (32 schools altogether). 1105 chemistry SS3 students from intact classes in the school that were picked formed the sample. The sample selection format is shown in Table 1.

Data for the study were gathered using the adopted versions of the Objective Paper I SSCE Chemistry adopted from June/July 2015 NECO (Type 1) and May/June 2015 WAEC (Type 2) were available. Adequate syllabus coverage was the reason for choice of instruments. Type 1 paper has sixty items, every item having five alternatives (A-E) while there are 50 items in Type 2 and every item has four options (A-D) from where candidates picked the right answer. The key was scored 1, while wrong option was scored 0. These two research instruments are standardized external examinations presumed moderated and verified by the testing boards, and they reflect the NECO and WAEC tests, respectively. However, the contents were used to determine the validity. To ensure that the test items were written in accordance with the requirements of the curriculum, the researcher carefully examined each instrument's test items in relation to its corresponding syllabus.

Two steps of data collection were used for this investigation. One involves gathering data from the Type 1 test and two involves gathering data from the Type 2 test during a two-week period. The administration of the tests took place under typical testing circumstances. With the support from the chemistry teachers in the selected schools, research assistants and approval from the school principal, study's instruments were administered to the students. The reason for carrying out the study was explained to the students, and they were also made aware that the data would be kept private. The assessments were meant to gauge the students' degree of readiness for their final exams, according to the research assistants who are graduates of several universities. They also helped with the process' overall monitoring. Data were analyzed using number correct and item pattern scoring method of CTT and IRT respectively, Inter- Correlation matrix, Scatter plot, Mean and Standard deviation.

Table 1: Sample Selection Format

Senatorial District	Selected Local Government Area	Number of Selected Schools	Population of Students
Osun West	5	10	384
Osun Central	5	10	245
Osun East	5	10	254
Co-educational Federal Unity School		2	167
Total	**15**	**32**	**1105**

Results

Research Question One: What are the scores of the examinees using CTT and IRT approaches of scoring?

Table 2 presents the summary of examinees correct item, the examinees' selection of the correct items (number correct scoring), estimation of ability (item pattern scoring), and the IRT scores translated into the number of correct scores.

Table 2: Number Correct and Converted IRT Ability Estimate and IRT performance of the Candidates

	CTT		IRT				CONVERTED IRT	
EXAMINEE	WAEC	NECO	WAEC_AB1	WAEC_AB2	NECO_AB1	NECO_AB2	WAEC_IRT	NECO_IRT
1	14	36	0.21	-0.36	-2.11	0.68	17.03	46.83
2	27	36	-1.46	-1.42	-2.04	0.97	39.46	24.29
3	32	28	-1.96	-1.80	-0.38	0.03	42.26	43.06
4	31	38	-1.73	-1.82	-1.82	0.92	28.05	13.19
5	17	17	-0.11	0.13	1.91	0.15	15.40	18.23
6	19	22	0.22	-0.90	0.40	-0.35	33.12	46.42
7	27	38	-1.51	-0.99	-2.23	0.21	33.24	17.49
8	31	23	-1.47	-1.64	0.51	0.12	39.04	44.38
9	29	40	-1.62	-1.63	-1.93	-0.15	33.13	45.51
10	24	38	-1.00	-1.15	-2.08	0.81	29.68	23.93
11	21	31	-0.98	-0.70	-0.34	-0.28	27.68	42.66
12	26	36	-1.07	-1.02	-1.75	1.15	35.54	46.50
13	28	36	-1.67	-1.45	-2.15	0.67	33.14	44.94
1091	24	34	-0.17	-1.90	-1.75	2.23	31.38	20.35
1092	27	20	-0.47	-1.93	0.23	-0.03	23.84	27.38
1093	14	26	0.67	-1.00	-0.54	-0.23	21.72	46.43
1094	18	39	-0.06	-0.41	-2.24	0.36	18.24	23.35
1095	24	29	0.02	-1.78	-0.24	-1.43	23.68	18.50
1096	13	18	0.55	-0.36	0.68	-0.85	17.51	50.59
1097	20	47	0.11	-1.77	-3.03	1.44	24.52	49.28
1098	21	43	0.39	-2.37	-2.67	1.76	35.81	49.24
1099	27	43	-0.42	-1.87	-2.79	1.57	32.63	22.32
1100	20	25	-0.64	-0.59	-0.18	-0.90	15.56	25.88
1101	11	25	0.81	0.38	-0.29	-1.83	16.77	27.43
1102	17	25	-0.17	-0.54	-0.26	-2.06	23.09	27.77
1103	22	24	-0.59	-0.67	-0.25	-2.16	19.14	48.95
1104	24	42	0.07	-1.82	-2.76	1.62	29.13	49.37
1105	23	42	-0.26	-1.65	-2.77	1.58	26.24	20.22
Mean	19.65	24.55					22.23	26.95
SD	8.15	9.32					9.86	11.69

Examinees' results from the type I and II tests using CTT measurement guideline was utilized to rate candidates' abilities, are shown in colomns2 and 3, titled WAEC and NECO. Results of the candidates' from IRT scoring are shown in columns 4 through 7. The approximated performance of candidates in dimension 1 of the two dimensions underpinning the WAEC chemistry test is shown in column 4 titled WAEC_AB1, and the estimated ability of the test takers in dimension 2 of the two dimensions underpinning the WAEC data set is shown in column 5. The estimated performance of the candidates in the first of the two dimensions underpinning the NECO Chemistry test is in column 6 with the label NECO_AB1, and the ability estimate of the examinees in the second dimension underpinning the NECO data set is in column 7 with the label NECO_AB2. The IRT scores were transformed into number correct scores in order to compare the CTT and IRT results more effectively. As shown in Table 2 candidates approximated scores in WAEC chemistry items using CTT was lower (X= 19.65, SD = 8.15) than their scores using IRT (X = 22.23, SD = 9.88) While estimated scores for candidates in NECO Chemistry items under CTT (X= 24.55, SD = 9.32) was lower than their scoring under IRT (X= 26.95, SD = 11.69).

Research Question Two: Is there any difference in the concurrent validity of WAEC and NECO SSCE chemistry examinees scores?

To answer this question, examinees scores from WAEC and NECO Chemistry items under CTT and IRT were correlated. The result is shown on Table 3.

Table 3: Correlation Matrix of WAEC AND NECO Items' Scores

	NECO_CTT	WAEC_IRT
WAEC_CTT	.63	
NECO_IRT		.61

Table 2 displays a correlation matrix for the results of the WAEC and NECO items. It demonstrates that utilizing the CTT method of estimation the correlation between examinees' scores on the WAEC and NECO chemistry items was 0.63, whereas using the IRT technique of estimation, it was 0.61. The significance of these findings is that there was a moderate correlation between examinees' scores on the WAEC and NECO chemistry questions. The sets of scores acquired using the number-correct (CTT) and item-pattern (IRT) methods of estimation in the WAEC and NECO chemistry questions, respectively, were plotted using scatter plot choices in order to further investigate the degree of association between examinees' results. This scatter plot is presented in Figures 1 and 2

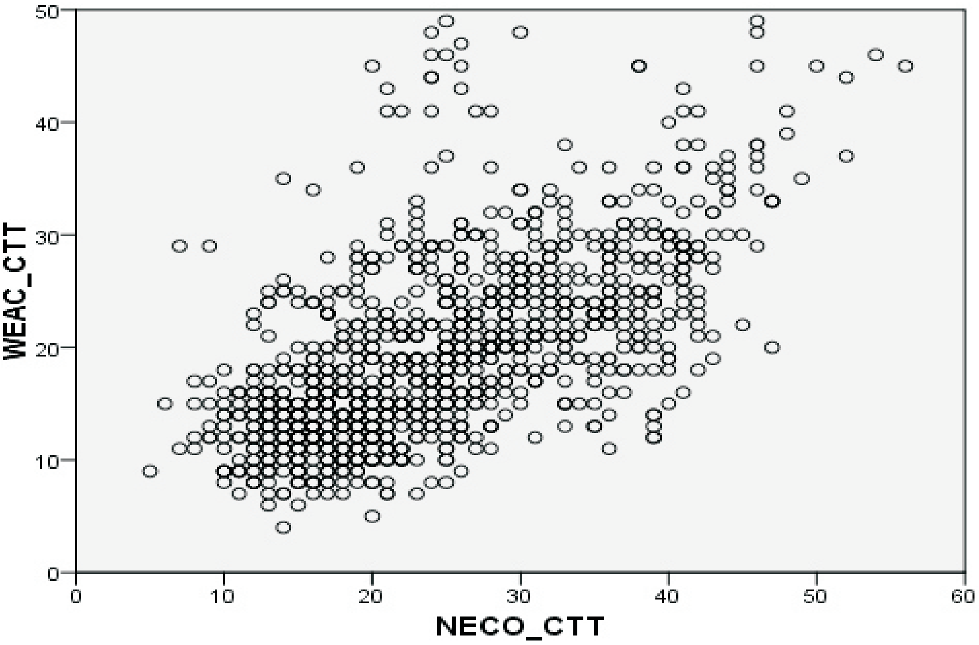

Figure 1: Scatter Plot of WAEC and NECO Chemistry Items Estimated under Number -Correct Scoring Method of CTT

Figure 1 shows that the scores obtained on the WAEC and NECO chemistry items by the examinees using scoring method of CTT were quite at variance with one another. In the same vein

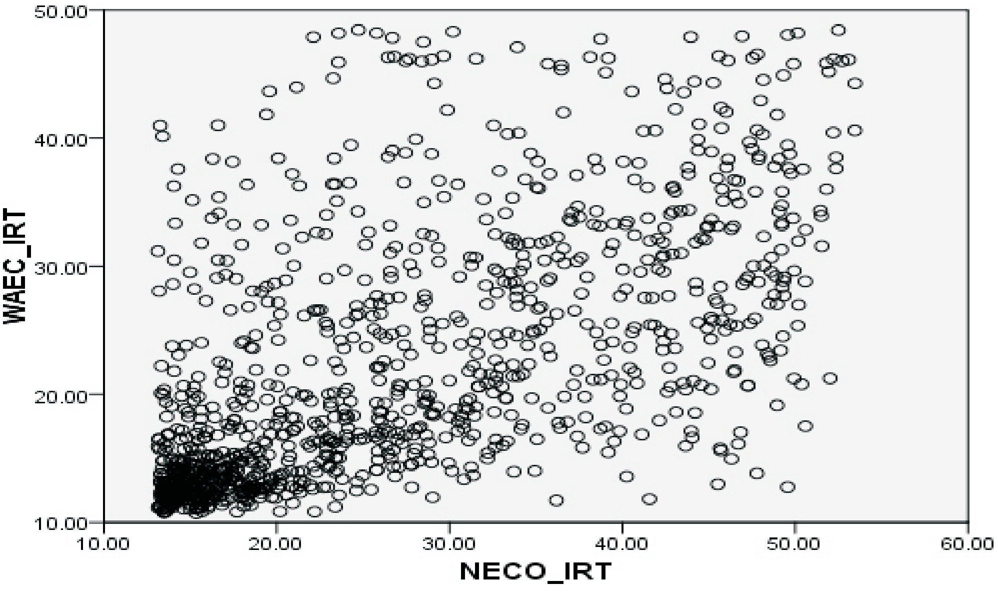

Figure 2: Scatter Plot of WAEC and NECO Chemistry Items Estimated under Item Pattern Scoring Method of IRT

Figure 2 shows that the scores obtained in WAEC and NECO chemistry items by the examinees using IRT method of scoring is quite at variance with one another.

Figures 1 and 2 show that candidates' performance in WAEC and NECO chemistry items were moderately related. Also, there was fair relationship between the candidates obtained scores on the tests. Although, considerable amount of differences existed in the scores obtained on the two tests. This result implied that the items in examination showed fair concurrent validity.

Discussion

The findings of research question one reveals that the candidates' approximated scores in NECO was significantly different (higher) from their scores in WAEC chemistry. Meaning that the scoring approaches of the two theories produced different results for the same candidate on each examination. It is also an indication that NECO may be easier for the candidates than WAEC. This may not be farfetched from the fact that as its name implies, IRT's attention is on the item-level information unlike CTT primary focus on test-level information (Fan, 1988). IRT models unlike CTT do not rely on sums or number correct scores to evaluate a candidates score nor do they assume equal contribution of the items to the overall scores (Metibemu, 2016). Valipour and Zoghi (2014) had a similar result in their comparative study of CTT and IRT in estimating test item parameters in linguistic test. Their result found that CTT and IRT are comparable. In the same vein Awopeju and Afolabi (2016) also in their comparative analysis of CTT and IRT based item parameters approximation of NECO mathematics examination established that the two frameworks are comparable. The outcome of the inter-correlation matrix and scatter plots of examinees scores using CTT and IRT scoring method showed moderate relationship between the examinees test scores (WAEC and NECO under CTT was 0.63 and 0.61 under IRT). There is presence of considerable difference on the scores obtained on the two tests which in essence made the concurrent validity to be fair.

Furthermore, research question two showed that candidates scores obtained in WAEC and NECO were quite at variance under the two theories. According to Wiberg (2004) if the plot is not showing a straight line is an indication that the estimates are not at variant. This result is supported by *Hassana and Abuh (2016) in their study on the relationship between students' achievement in Mathematics examinations conducted by WAEC and NECO using the Pearson Moment Product Correlation showed that there is a weak relationship that existed between students' achievement in WAEC and NECO mathematics results. Also in support of this finding is Bernadine and Augustine (2022) who submitted that no significant relationship was observed in the distribution of items across the various levels of cognitive domain by WAEC and NECO. Although,* this result negated Kolawole (2007) findings that WAEC and NECO 2005 Mathematics objective questions have high significant relationship. At the same time the result is not at par with the findings of Oluwatayo (2007) where the two examinations are significantly related. This finding also negated the findings of Salako Adegoke and Ogundipe (2017) who compared the performance of students on WAEC and NECO mathematics and physics and reported statistically significant difference in the mean between the two groups.

Conclusion

This study concluded that IRT approximates candidates' scores higher than CTT considering its ability to provide item invariant parameters. Also, NECO items are easier for the candidates compared to WEAC items considering their approximated scores. The concurrent validity between WAEC and NECO (2015) chemistry items using the two test theories was only moderate.

Recommendations

It is therefore recommended

1. That teachers, examiners and examination bodies should adopt IRT scoring method which take into account invariant item parameters as this will help

improve the performance of students.

2. Concurrent validity should also be considered at the item development stage.

References

Ahmed, M.F. (2014). Difficulty index of mathematics multiple-choice items of West African Examinations Council and National Examinations Council senior secondary school certificate examinations from 2006 – 2010 *Journal of ATIP,* ***13***, 25 – 31.

Adegoke, B. A. (2014). Effect of item-pattern scoring method on Senior Secondary School Students' Ability Scores in Physics Achievement Test, *West African Journal of Education*, *24*, 181-190.

Afolabi, E.R.I. (2012). *Tests and Measurement*: *A tale bearer or true witness?* An inaugural lecture series 253. Obafemi Awolowo University press limited.

Anastasi, A. 1988. Psychological Testing. Macmillan.

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal*, *12*(28), 263-284.

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation. Pp. 176.

Bernadine, N. N., & Augustine, U. O. (2022). Comparative Analysis of 2021 and 2022 WAEC and NECO Chemistry Multiple Choice Questions in Enugu State, Nigeria. *British Journal of Education, 10*(14), 7-14.

Cohen, R. J., & Swedlik, M. E. (1999). *An introduction to tests and measurement*. Psychological testing and assessment. 4th ed. Mayfield Publishing House.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological test. *Psychological Bulletin, 52*, 281-302.

Dibu-Ojerinde, O. O. (2012). General principles of test planning. *Educational Tests and Measurement,* Obafemi Awolowo University Press.

Dimiter, M. O. (2012). Statistical methods for validation of assessment scale data in counselling and related fields. American Counselling Association, www.counseling.org

Erguven, C., & Erguven, M. (2014). An empirical study on assessment of item-person statistics and reliability using Classical Test Theory measurement measures. *Journal of Technical Science and Technology*. ISSN 2298 – 0032: Pp. 25-33.

Faleye, B. A. & Dibu-Ojerinde, O. O. (2005). Some outstanding issues in assessment for learning. Paper Presented at the 2005 Annual Conference of the International Association for Educational Assessment. (IAEA), Hilton Hotel, Abuja (Nigeria).

Guler, N., Uyanik, K. G., & Teker, G. T, (2013). Comparison of classical test and item response theory in terms of item parameters. *International Journal of Social Sciences Research*, *2*(1), 1-6

Hambleton, R., & Jones, R. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*, 38 - 47.

Hambleton, R. K., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In H. Tinsley & S. Brown (Eds.). Handbook of applied multivariate statistics and modelling. Academic Press.

Hassana, O. A., & Abuh, E. E. (2016). Correlational analysis of student achievement in West Africa Examination Council and National Examination Council of Nigeria in mathematics, *Journal of Research in National Development, 14*(1), 1-13.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory:* Application to psychology www. Brain bench .com

Kline, T. J. (2005). *Classical test theory assumptions, equations, limitations, and item analyses*. Psychological testing. Sage Publications, *5,* 91-106.

Kolawole, E. B. (2007). A comparative analysis of the psychometric properties of Nigerian two examining bodies for Senior

Secondary School mathematics. *Research Journal of Applied Sciences*, *2*(8), 913-915

Krishnan, V. (2013). The early child development instruments (EDI): *An item analysis using Classical Test Theory (CTT) on Alberta's data.* Early child development mapping (ECMap) project Alberta, community University partnership (CUP), Faculty of extension, University of Alberta, Edmonton, Alberta.

McDonald, P. (1999). *Test theory: A unified treatment*: Mabwab, N. J: Lawrence Erlbaum Associates

Meadows, M., & Billington, P. (2005). *A review of the literature on marking reliability*. Report produced for the National Assessment Agency

Metibemu, M. A. (2016). *Comparison of classical test theory and item response theory frameworks in the development and equating of physics achievement tests in Ondo state, Nigeria*. Unpublished Ph.D. Thesis, Institute of Education University of Ibadan.

Moore, D. S., Notz, W. I, and Flinger, M. A. (2013). The basic Practices of statistics (6th ed.). W. H. Freeman and Company

Natarajan, V. (2009). *Basic principles of IRT and application to practical testing and assessment*. MeritTrac Services (P) LTD. India

Ojerinde, O.O. & Faleye, B. A. (2005). Do they end at the same point? Journal of Social Science, (3), 239-241.

Ojerinde, D., & Ifewulu, B. C. (2012). *Item unidimensionality using 2010 Unified Tertiary Matriculation Examination Mathematics pretest.* A paper presented at the 2012 international conference of IAEA Kazastan.

Ojerinde, D. (2016). *Lecture Modules on Item response theory (IRT),* Joint Admission and Matriculation Board (JAMB). Pp. 72.

Okpala, P. N., Onocha, C.O., & Oyedeji, O. A. (1993). Measurement in Education. Jattu-Uzairue: Stirling-Horden Publishers (Nig.) Ltd.

Olutayo, J. A. (2007). Comparative Analysis of Students' Performance in Chemistry in WAEC and NECO Senior School Certificate Examination. *International Journal of Research in Education, 4*(1&2), 184-200.

Olutola, A. T. (2015). *Empirical analysis of item difficult and discrimination indices of senior school certificate multiple choice biology tests in Nigeria*. A paper presented at the 41st annual conference of International Association of Educational Assessment (IAEA) held on 11th - 15th October, 2015 at University of Kansas, Lawrence, Kansas, USA.

Peter K. (2012). A study of the attitude of some Nigerian science students towards NECO and WAEC. *Journal of Professional Science and Vocational Teachers Association of Nigeria, 12*(1), 15-18

Salako, R. J., Adegoke, B.O., & Ogundipe, L. O. (2017). Performance appraisal of NECO and WAEC SSCE: An empirical evidences from mathematics and physics. *International Journal of Innovative Social & Science Education Research, 5*(3):1-10.

Seyi, A. I., & Clement, A. A. (2012). A correlational analysis of students' achievement in WAEC and NECO Mathematics. *Journal of Education and Practice*, *3*(1), 23-36

Skurnik, L. S. and Nuttal, D. L. (1968). Describing the reliability of examinations. *The Statistician,* 18, 119-128.

Valipour, V., & Zoghi, M. (2014). A comparative study of classical test and item response theory in estimating test item parameter in a linguistic test. *Indian Journal fundamental and Applied Life Sciences. 4*(54), 424-435.

Wiberg, M. (2004). *Classical test theory vs item response theory: An evaluation of the theory test in Swedish driving license test (No. 50).* Kluwer Academic Publications

William, D. (2000). Reliability, Validity and all that Jass Education, *29*(3), 9-13