and scoring the test. This will allow an individual to earn scores that are objective reflection of his or her level of understanding and not the scores that reflect the views of persons administering or scoring the test. Therefore, a test in which the test-taker will get the same score no matter who is administering the test is an objective test. In standardized tests, the questions, conditions for administering, scoring procedures, and interpretations are consistent. The use of a well-designed standardized test will help show the level of mastery of an individual area of knowledge or skill. In Nigeria, at the end of secondary school education, students are qualified to sit for certification examinations such as Senior School Certificate Examination (SSCE) conducted by the West African Examinations Council (WAEC) and National Examinations Council (NECO). The purpose of these examinations is to measure how much students have achieved the educational objectives of each subject. The respective certificates awarded by these examination bodies are officially recognized in Nigeria as equal. The certificates can also be used to accomplish some other purposes like securing employment in the appropriate cadres of public service, private companies and corporations and gaining admission into institution of higher learning. These certificates can also be merged where grades are weak or inadequate to secure admission into institution of higher learning both in Nigeria and abroad.

The importance attached to the SSCE both economically, socially and the privilege for admission into institutions of higher learning by the owner of the certificates make the awarding of these certificates one of the most important events in the Nigerian academic calendar. With a similar mandate, use of similar syllabus and standardized tests by WAEC and NECO to assess students' level of knowledge in various subjects, it is believed that the test items, conditions of administration, procedure for scoring and interpretations are consistent. In spite of this mandate, there are criticisms of different forms about the credibility of the examinations conducted by these bodies from major stake holders. The criticisms include: non-equivalence in the quality of examination items, disparity in performance, mass leakage of examination papers, overcrowding in examination halls and examination malpractices among others. According to Peter (2012, 15-18), the substandard nature of NECO made some Federal universities from 2002 to 2012 to have rejected NECO results. Ahmed (2014, 25-31) stated that NECO questions from 2011 to 2014 were of higher standard than those of WAEC. Ojerinde and Faleye (2005) stated that there was no difference between NECO and WAEC, when they were compared. Of all criticisms levied against these examination bodies those that gave the researcher much concern were the non-equivalence of the items in terms of difficulty and disparity in performance of students.

Seyi and Clement (2012, 23-36) reported that NECO has been alleged to be tougher than WAEC. This assertion may be due to the fact that WAEC being the first to be attempted by the students, is given all seriousness while they relax in NECO examination. Some others hold a view that NECO examinations have the easier items relative to WAEC. Considering the fact that NECO is administered immediately after WAEC, it is believed that students would have gained experience from the previous WAEC which helps them to perform better in NECO. These opinions have led to unreliability of examination results and believe in the superiority of one certificate over the other. For example, for many years now, Osun State government alongside some other States in Nigeria have played a major role in assisting students and parents in paying for WAEC SSCE registration. It is not the same with NECO. This may be an indication that the State also believes in the superiority of WAEC over NECO.

Ordinarily, if an examinee scores a high mark in a test, the test is described by the examinee as being easy and if otherwise it is described as being difficult. However, the properties of a test upon which an assessment of candidate is based is the psychometric characteristics of examinations. These properties include the difficulty and the discrimination indices, validity and reliability indices. At present, there are two popular statistical frameworks in educational measurement through which tests can be developed, validated, and finally used for assessing examinees performance. They are Classical Test Theory (CTT) and Item Response Theory (IRT). Classical Test Theory (CTT) pays attention to test level, its weak theoretical assumptions make it relatively easy to apply in many testing situations. It however has few weaknesses. Among these are: the person statistics is sample dependent, and the item statistics (item difficulty and item discrimination) are (examinee) sample dependent (Fan, 1998, 1-7). These pose some theoretical difficulty in some measurement situations. Under Classical Test Theory, examinee's test scores (usually called observed score) is made up of two components, the true score and the error score. The observed score is obtained by adding up the items answered correctly. Thus, under the CTT, examinee's test scores would be the sum of scores received on all items in a test. This method of scoring is referred to as number-correct scoring

(Adegoke, 2014; Tomkowickz and Wright, 2007, 181-190). Number right scoring occurs where the testee would obtain one (1) or zero (0) in a particular item (Kolawole, 2011).

The number-correct method of scoring produces maximum likelihood trait estimates based on raw scores. Obtained score (i.e. test scores) will always be an integer number and will range from 0 to N, the number of items in the test (Metibemu, 2016). Although, item statistics (i.e. item discrimination power and difficulty index) are important part of the Classical Test Theory, they are not used in aggregating examinee's test scores. However, they are used in the course of test development to identify and delete problematic items (Meyer and Zhu, 2013, 26-39). As a result, examinees who answered the same number of items irrespective of the items' level of difficulties and discrimination earn the same test scores.

Unlike the Classical Test Theory, Item Response Theory determines examinee's performances on a test based on the estimates of ability obtained from the pattern of item responses. An IRT model contains ability as well as item parameters, however the values for these are unknown. Only the test items are known, while the ability and parameters are unknown and therefore must be estimated (Hambleton Swaminathan, and Rogers, 1991). It attempts to model the ability of a test-taker and the probability of answering an item correctly based on the pattern of responses to all the items that constitute the test. Its ability parameter estimates are not test dependent and item parameter estimates are not group dependent. It overcomes the weaknesses of CTT with its ability to provide invariant item parameters. Under IRT, the primary interest is in whether an examinee gets an item correctly or not rather than in the raw test scores. According to Adegoke (2014, 181-190); Thorpe and Favia (2012), IRT uses logistic models to estimate the contribution of each item scored correctly by examinees on the latent trait of interest. This method of scoring is referred to as item-pattern scoring (Adegoke, 2014, 181-190; Thorpe and Favia, 2012). These can be done by procedures including maximum likelihood, Bayesian estimation procedure. The method of scoring takes into account not only a student total raw score as does the CTT-based Number-Correct, but in addition which test items the student answered correctly, as well as the psychometric characteristics of the items answered correctly. Since there are infinite number of logistic models that can be configured, and the parameters are difficult mathematically to produce, there are computer programs to complete such IRT models, especially one-, two-, and three- parameter logistic models**.** Estimates of ability under Item Response Theory have values that range from -∞ to + ∞. Although, in practice it usually ranges from -3 to +3, while under CTT test scores are integers that range from 0 to N, the total number of items contained in the test. Chemistry being a branch of science that deals with the study of the composition of a substance and how the properties relate to their composition. It has played a major role in science, technology and society, and it still does same till today. Hardly is there anything found in nature that chemistry does not have an influence or impact upon. To support this there's a saying that without chemistry there will be no life.

Weighing the usefulness and educational value of chemistry to the need of individual learner, economic and technological breakthrough of a nation and the effort of researchers to improve on its teaching and learning in Nigeria, it is essential that it is properly taught in schools and evaluated using equivalent standards. The procedure of item development, administration, scoring and interpretation of results of these examination bodies must be at equivalence. This way differences in examinees' scores will be as a result of individual academic effort. Since multiple-choice test is one of the type of test used by NECO and WAEC to set their questions in many subjects which include chemistry and there is no elaborate information about the equivalence of the item parameters of the examinations conducted by these two examination bodies, it is imperative to estimate examinees' scores in NECO and WAEC chemistry examination items in order to establish their equivalence or otherwise using CTT and IRT. This is because many of these criticisms are pointing at the credibility of WAEC and NECO SSCE questions which is dependent on their item parameters. Unlike IRT that has the capability to provide invariant item parameter and estimate adequately examinees' ability, CTT assesses items statistics such as difficulty and discrimination during test developments. These parameters have never been used in the estimation of examinees' score (Metibemu, 2016). Even though the CTT has been criticized for its inaccuracy in estimating test item characteristics and examinees' scores, students' achievement in Nigeria has been measured by teachers and public examining bodies based on the sum of their total scores which is typical of CTT. Research in other areas have shown that the procedures and frameworks for test development and how test items are scored can impact students' performance negatively or positively.

Situation of the Problem

Many researches had been conducted on comparison of performance of students in WAEC and NECO (using only the analytical method of CTT) in different subjects and states of the Federation with very few of such researches focused on comparison of students' performance in WAEC and NECO Chemistry using CTT and IRT in Osun State secondary schools.

Aim of the Study

The purpose of the study was to provide empirical explanation on the appropriateness of decisions made statistically based on examinee scores using classical test theory (number correct) and item response theory (item pattern) methods in WAEC and NECO SSCE Chemistry item in Osun State, Nigeria.The following research questions were raised for the study: 1) What are the parameters of WAEC and NECO SSCE chemistry items in terms of Unidimensionality? What are the test scores of examinees in WAEC chemistry items using number correct scoring (CTT) and item-pattern scoring (IRT) methods? One research hypothesis was also raised for this study: There is no significant difference in the test scores of examinees' in WAEC and NECO chemistry items using number correct scoring (CTT) and item-pattern scoring (IRT) methods.

METHOD

The study adopted the descriptive survey research design. Descriptive survey research design was employed since the data involved in the study were collected from the source without any manipulation. This survey approach was considered most suitable because the study sought information from a small segment of the population to make a generalization for all science students that sat for the 2017/2018 senior school certificate (SSCE) Chemistry examination paper 1 in Nigeria.

Study Group

The population for the study comprised all science students that sat for the 2017/2018 senior school certificate (SSCE) Chemistry examination paper 1 in Nigeria. A sample of 1,105 science final year candidates was randomly selected from a total population of 36,182 students who took the examination. Senior Secondary (SS III) chemistry students in the 32 selected schools, constituted the sample.

Material

The instruments for the study titled, "Chemistry Achievement Test Type 1" (NECO), and "Chemistry Achievement Test Type 2" (WAEC) were used to collect data. These were the adopted versions of June/July 2015 NECO and May/June WAEC 2015 Senior School Certificate Examination Chemistry (Objective) Paper 1. The NECO Type 1 paper was a five option objective test consisting of 60 items with each item having five options, lettered A-E while WAEC Type 2 paper was a four optioned test consisting of 50 items, and lettered A-D that was based on the senior school certificate chemistry curriculum in Nigeria. Correct response attracted a score of 1, while incorrect response attracted 0 based on the Senior Secondary School Chemistry curriculum.

Data Analyses

Data collected were subjected to SPSS, DIMTEST package, MIRT and equate IRT packages of R.

FINDINGS

Research Question One: What are the parameters of WAEC and NECO SSCE chemistry items in terms of Unidimensionality?

To answer this research question, the responses of examinees to the WAEC and NECO SSCE Chemistry items were subjected to Stout's Test of Essential Unidimensionality. This was done by separating the test into two subtests, the Assessment Subtest (AT) and the Partitioning test (PT). The AT are the items chosen as those that measure best along a dominant trait.

Table 1: Summary Statistics showing the Parameters of WAEC and NECO SSCE Chemistry Items in terms of Unidimensionality

	WAEC					NECO			
TL	TGbar	T	P value	decision	TL	TGbar	T	p-value	decision
7.6144	5.2854	2.3174	0.0102	Sig	9.2672	1.0304	8.1959	0.0000	sig

Significant <.05

Source: Own Analysis, 2019

Table 1 showed that the AT was dimensionally distinct from the remaining items of the test (t = 2.3174, p-value = 0.0102, one-tailed); therefore, the assumption of unidimensionality was rejected. The result shows

that more than one dimension accounted for the variation observed in examinees responses to the test items. Hence, the WAEC chemistry items violated unidimensionality assumption. Correspondingly, Table 1 showed that the AT was dimensionally distinct from the remaining items of the test (t = 8.1959, p-value = 0.0000, one-tailed). This indicated that more than one dimension accounted for the variation observed in student's responses to the test items. Thus, the NECO chemistry test violated unidimensionality assumption. Hence, the NECO chemistry items violated unidimensionality assumption. The results imply that the traits measured by the two groups of items were significantly different from one another (multidimensionality).

Research Question Two: What are the test scores of examinees in WAEC chemistry items using each of number correct scoring (CTT) and item-pattern scoring (IRT) methods? To answer this research question, the aggregate of the number of items that individual examinee correctly picked (i.e., the CTT scoring), ability estimates of the examinees (IRT scoring) and converted IRT scores to number correct scores were determined. The conversion was done using the M3PL model that fitted the data sets.

An abridged result of the performance of the examinees estimated using CTT and IRT framework respectively are presented in Table 2.

Table 2: Summary Statistics showing the Number Correct Scores of Examinees, IRT Ability Estimate and Converted IRT Scores of the Examinees in WAEC and NECO Chemistry Items using Number Correct Scoring (CTT) and Item-Pattern Scoring (IRT) Methods

	CTT		IRT				CONVERTED IRT	
EXAMINEE	WAEC	NECO	WAEC_AB1	WAEC_AB2	NECO_F1	NECO_F2	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

Source: Own Analysis, 2019.

Table 2 presents the aggregate number of correct items picked by individual examinees (i.e., the CTT scoring), ability estimates of the examinees (IRT scoring) and the IRT scores converted to number correct scores. The second and the third columns labelled WAEC and NECO represents the scores obtained by the examinees on the WAEC and NECO chemistry items respectively when CTT measurement framework was used in scoring their performance. Column 4 to 7 presents the scores of the examinees in the tests under IRT scoring approach. Column 4 labelled WAEC_AB1 is the estimated ability of the examinees in dimension 1 of the two dimensions underlying the WAEC chemistry test and column 5 with the label WAEC_AB2 is the ability estimate of the examinees in the second dimension underlying the WAEC data set. Column 6 labelled NECO_AB1 is the ability estimate of the examinees in dimension 1 of the two dimensions underlying the NECO Chemistry test and column 7 with the label NECO_AB2 is the ability estimate of the examinees in the second dimension underlying the NECO data set. Table 2 further revealed that examinees' estimated scores in WAEC and NECO chemistry items under IRT was higher ($\overline{X}$ = 22.23, SD = 9.88) than their scores under CTT ($\overline{X}$ = 19.65, SD = 8.15). This implies that the IRT item pattern method is consistent in estimating and interpreting examinee latent ability on the scale.

Hypothesis Testing

Hypothesis One: There is no significant difference in the test scores of examinees' in WAEC and NECO chemistry items using number correct scoring (CTT) and item-pattern scoring (IRT) methods. To test whether the difference observed in the scores of examinees' on WAEC items was significantly different using the number correct scoring method of CTT and item pattern scoring method of IRT, the scores obtained by the examinees under CTT and IRT was subjected to paired-samples t-test. The results are presented in Table 3

Table 3: Summary Statistics of Paired Samples t-test showing the difference in Test Scores of Examinees on WAEC Chemistry Items Estimated under CTT and IRT Measurement Frameworks

	Paired Samples Correlations			Paired Differences					
	N	Correlation	Decision	Mean	Std.Deviation	Std.Error Mean	t	Df	Decision
WAEC_CTT - WAEC_IRT	1105	0.749	NS	-2.58	6.574	0.198	13.06	1104	S

Significant <.05 (2-tailed)

Table 3 revealed that although the two measurement frameworks are significantly related (r = 0.749, p<0.05) there is a significant difference in examinees estimated scores in WAEC Chemistry items using CTT and IRT measurement frameworks respectively (t = 13.06, p=.000 < 0.05). This indicated that examinees estimated scores in WAEC Chemistry items generated using number correct scoring (CTT) measurement framework is different from the examinees estimated scores generated using item-pattern scoring (IRT) measurement frameworks.

Table 4: Summary Statistics of Paired Samples t-test showing the difference in Test Scores of Examinees on NECO Chemistry Items Estimated under CTT and IRT Measurement Frameworks

	Paired Samples			Paired Differences					
	N	Correlation	decision	Mean	Std.Dev.	Sd.Error Mean	t	df	decision
NECO_CTT-NECO_IRT	1105	0.53	NS	-2.40	10.396	0.313	-7.667	1104	S

Source: Own Analysis, 2019

Similarly, Table 4 showed that examinees estimated scores on the NECO chemistry items using CTT and IRT measurement frameworks was significantly related (r = 0.53, p<0.05). However, there is a significant difference in examinees estimated scores in NECO Chemistry items using CTT and IRT measurement frameworks respectively (t = 7.667, p < 0.05). This specified that examinees estimated scores in NECO Chemistry items generated using number correct scoring (CTT) measurement framework is different from the examinees estimated scores generated using item-pattern scoring (IRT) measurement frameworks. The results from table 3 and 4 implies that the number correct scoring method (CTT) and item pattern scoring

method (IRT) produced different scores for the same examinees on the test items, with IRT scoring procedure producing higher scores for the examinees.

DISCUSSION AND SUGGESTIONS

The CTT (number correct) and IRT (item pattern) scoring procedures produced different scores for the same examinees on a test, with IRT item pattern scoring procedure producing higher scores for the examinees. This may be due to the fact that IRT is more theory grounded and models the probabilistic distribution of examinees' success at the item level. As its name indicates, IRT primarily focuses on the item-level information in contrast to the CTT's primary focus on test-level information, Fan (1988). IRT models, in contrast to CTT, do not rely on sums or number correct scores to evaluate a person's performance, nor do they assume equal contribution of the items (questions) to the overall scores. Since items vary in their difficulty and persons vary in their trait level, this method may result in a more accurate assessment of respondents' latent traits because respondents with the same sum score may differ in their trait measurement (Le, 2013). This supports the findings of Fan (1998) and MacDonald and Paunonen (2002) from their empirical studies of the differences between these two test models that IRT differs considerably from CTT in theory and commands some crucial theoretical advantages over the CTT and it is expected that there will be appreciable differences between the IRT and CTT – based person statistic.

Conclusion and Suggestion

Within the limit of this study, IRT and CTT are not comparable in the estimation of examinees test scores. IRT estimates examinees' scores better than CTT considering its capability to provide item invariant parameter, which put the characteristics of the examinees' and the test itself into consideration in scoring examinees' responses to test items. Overall, there is a display of the superiority of the theoretical advantages of the IRT in the scoring method using the right data modelling approach. Determination of IRT dimension should come first before the application of a particular model. It is not always that a unidimensional latent variable may be appropriate as some set of test items may measure more than one ability. IRT methods of scoring estimates examinees scores better than CTT when attention is on the assessment of students' performance and in identifying individual performance in a test.

REFERENCES

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 Ile-Ife, Nigeria

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

Faleye, B. A. and 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).

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58 (3), 1-17.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. London, Sage Publications.

Kolawole, E. B. (2011). *Principle of tests construction and administration*. 2nd Revised Edition. Lagos: Bolabay Publications (Nig). Louis Cohen, Lawrence Manion, Keith Morrison 2007. *Research Methods in Education*. Abingdon, Oxon: Routledge.

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, Ibadan.

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

Okpala, P. N., Onocha, C. O., and Oyedeji, O. A. (1993). *Measurement in education*. Jattu-Uzairue: Edo State, Stirling- Horden Publishers (Nig.) Ltd.

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

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

Thorpe, G. L. & Favia, A. (2012). *Data analysis using Item Response Theory methodology: An introduction to selected programmes and applications psychology faculty scholarship paper 20*. Retreived 15th June, 2014 from http://digitalcommons. library.umaine.edu/psyfacupub/20.

Tomkowicz, J. T., & Wright, K. R. (2007). *Investigation of the effect of test equating and scoring methods on item parameter estimates and student ability scores*. A paper presented at the annual conference of *American Educational Research Association*, Chicago, April 10.