PRINT: ISSN 0975-1122 ONLINE: 2456-6322
Int J Edu Sci, 3(2): 137-144 (2011)
DOI: 10.31901/24566322.2011/03.02.09

A Psychometric Analysis of Two Major Examinations in Nigeria: Standard Error of Measurement

A. D. E. Obinne

Department of Educational Foundations and General Studies, University of Agriculture, Makurdi, Nigeria

KEYWORDS Students. Senior Secondary School. Item Response Theory. Measurement Effectiveness

ABSTRACT This study deals with the psychometric analysis of the two major examinations conducted in Nigeria by NECO and WAEC. The objective was to compare the standard error of measurement of Biology examinations conducted from 2000 – 2002 using the one-parameter model of Item Response Theory (IRT). Standard error of measurement (SEM) is commonly used to produce confidence interval and it is an estimate of how much error there is in a test. Instrumentation research design was used for this study. Benue State, Nigeria was the study area. The population for the study comprised all year three (SSIII) senior secondary school students who enrolled for May/June/July 2006 Biology senior secondary school certificate examination of NECO and WAEC in the three education zones of Benue State. The sample for the study was one thousand eight hundred (1800) students. Multi-stage stratified sampling technique was used to get this sample. NECO and WAEC 2000 – 2002 objective Biology questions were the instruments for the study. The maximum likelihood estimation techniques of the BILOG MG Computer Programme and the SPSS were used for data analysis. The results showed significant differences in the SEM of Biology examinations conducted by NECO and WAEC in 2000, 2001 and 2002. This implied that Biology examinations conducted by NECO had smaller SEM (high reliability) than those of WAEC. It has recommended that IRT analysis should be employed by Nigerian Examination bodies.

INTRODUCTION

Evaluation is seen as a qualitative description of pupil behaviour (Itsuokor 1986). Mehrens and Lehman (1981) opine that no matter how efficient the teacher is, how intelligent the students are and how adequate the audio-visual equipment, if no provision is made for the evaluation of the students' progress, the teaching effort may be completely invalidated. Evaluation concerns determining the quality of the curriculum, facilities and performance of pupils, using various tools. Test is one such tool for evaluation.

Test consists of a set of uniform questions or tasks to which a student is to respond independently and the result of which can be treated in such a way as to provide a quantitative comparison of the performance in different students (Nworgu 1992). Testing is a fundamental part of the teaching-learning process used not only as a basis for ranking students at the end of the teaching – learning process but to guide teaching, and aid in the development of curriculum, as well as in the assessment of needs, learning difficulties, level of mastery and differences among students. Based on the level of performance criterion, there are three types of test namely, the General Mental Ability Tests, Separate Ability Test and Achievement Test (Nkemakolam 2003).

The practical relevance of these tests and their testing is largely dependent on their levels of reliability, validity, difficulty and discrimination. Validity is, according to the Standards for Educational and Psychological Testing, "the most fundamental consideration in developing and evaluating tests" (cited in Hogan and Agnello 2004).

One kind of support for the validity of the interpretation is that the test measures the psychological trait consistently. This is known as the reliability of the test. Reliability, that is, a measure of the consistency of the application of an instrument to a particular population at a particular time, is a necessary condition for validity. A reliable test may or may not be valid, but an unreliable test can never be valid. (http://www.wordengine.jp/research/pdf/IRT_reliability_and_standard_error.pdf).

All these add up to the psychometric properties of a test. The development of achievement, ability, aptitude, interest and personality tests is generally a multi-step process that can follow one of two distinct measurement frameworks. These are usually called the Item Response Theory (IRT) and Classical Test Theory (CTT) measurement strategies (Macdonald and Paunanen 2002). In education, psychometricians apply IRT in order to achieve tasks such as developing and refining examination items, maintaining banks of items for exams, and equ-

ating for the difficulties of successive versions of examination (for example, to allow comparisons between results over time).

This study sought to confirm one of the psychometric properties (reliability) of the test items in the Biology examinations of two examining bodies in Nigeria (West African Examinations Council and National Examinations Council) using the Item Response Theory (IRT) measurement framework.

This study was necessitated by public complaint and conception of the superiority of the test items of one examination over the other.

The IRT measurement framework was chosen for this study because it takes care of the limitations of the Classical Test Theory (CTT). The major objective of the study was to determine the standard error of measurement of Biology examinations conducted by NECO and WAEC.

The results of this study would help to indicate the overall quality of the examinations conducted in Biology by WAEC and NECO. A confirmation of the reliability and validity of the examinations conducted by WAEC and NECO would help to establish public confidence and acceptability of results from their examinations. The public needs to be convinced that the examinations conducted by both WAEC and NECO are of relatively equal standards and that no one is of inferior quality. Thus, the public will be enlightened on the interpretation of students' results from the examinations conducted by these two examining bodies. They will not just assume high quality of either of the examining bodies. Presently, the performances of students in the examinations conducted by these two bodies are interpreted based on the sum of their total scores which is typical of Classical Test Theory (CTT). The use of these scores just summed across all items to consider the performance of examinees hides the characteristics of both the examinee and the test. For objective and adequate decisions to be taken on the performance of students in examinations by WAEC and NECO, the psychometrics of the tests needs to be determined. Examining bodies need to consider the psychometric properties of tests in taking decisions on the observable performance of candidates in order to improve upon test construction, administration and analysis. In addition, from this study, WAEC and NECO would have a clearer understanding of their performance in test construction and be appropriately guided from now on as they, hopefully, accept and adopt Item Response Theory (IRT) evaluation framework.

Standard Error of Measurement (S.E.M)

The latent trait models do not use the concept of reliability; what is used is the concept of standard error of measurement or precision of measurement (Korashy 1995). Standard error of measurement, according to Biirbaum (1968), is defined as the measurement effectiveness of a test item at each level of the trait ability being measured. While the reliability of a test is clearly important, it is probably one of the least understood concepts in testing. One of the purposes of the reliability coefficient of a test is to give us a standard index with which to evaluate the validity of a test. More importantly, the reliability coefficient provides us with a way to find the (SEM) the Standard Error of Measurement. SEM allows practitioners to answer the question, "If I give this test to this student again, what score would she achieve?" (http://www.wordengine.jp/research/pdf/IRT_reliability_and_standard_error.pdf).

Psychometricians describe the Standard Error of measurement of a test differently but meaningthe same (From the Virginia Department of Education. www.doe.virginia.gov/VDOE/Assessment/sem.html.) The standard error of measurement (SEM) is a statistical phenomenon and is unrelated to the accuracy of scoring. All test results, including scores on tests and quizzes designed by classroom teachers, are subject to the standard error of measurement. If a student were to take the same test repeatedly, with no change in his level of knowledge and preparation, it is possible that some of the resulting scores would be slightly higher or lower than the scores that precisely reflect the student's actual level of knowledge and ability. The difference between a student's actual score and his highest or lowest hypothetical score is known as the standard error of measurement.

In the same light, SEM is described as the standard deviation of test scores that would have been obtained from a single student had that student been tested multiple times. It is a measurement of the "spread" of scores within a student had the student been tested repeatedly (www.tea.state.tx.us/student. assessment/taks/standards/s...).

The most common use of the SEM is the production of the confidence interval. The SEM

is an estimate of how much error there is in a test. The SEM can be looked at in same way as standard deviations (www.ed.Sc.edu/caw/technical/sido13.htm). The SEM can be added and subtracted to a student's score to estimate what the student's true score would be.

Ability estimates and each item validity are accomplished by its standard error in the latent trait model. The standard error of measurement is independent of the particular examinee sample and it is an indication of the amount of error in ability estimate at different points of the ability continuum (Lord and Norvick 1968). Each item contributes to the standard error independently of other items in the test. This is not so with the classical reliability estimates where the contribution of each item to the test reliability and validity depends upon what other items are in the test (Wood 1980). The contribution of one item is not easily identified. The standard error associated with a test item or person measurement (ability estimate) is evaluated using the PROX procedure (Izard and White 1980). Wright and Wainer (1980) talked about the Rasch standard error as another measure of standard error and the use of the maximum likelihood technique.

Other methods of estimating SEM include using a reliability coefficient and the tests standard deviation:

$SEM = S (\sqrt{1 - r})$

where s = the standard deviation for the test and r = the reliability coefficient for the test (web.sau.edu/Water streetMaryA/NEW%20 intro%20to%20test%2.).

Research Questions

The study answered the following research questions:

1. What are the standard errors of measurement of the test items in the Biology examinations conducted by NECO?
2. What are the standard errors of measurement of the test items in the Biology examinations conducted by WAEC?

Hypothesis

1. There is no significant difference in the mean standard errors of measurement of the items in the Biology examinations conducted by NECO and WAEC from 2000 – 2002 based on One, parameter model of IRT.

METHODOLOGY

Research Design: Instrumentation research design was deemed appropriate for this study. Instrumentation research is seen as a study which aims at introducing new concepts, procedures, technologies or instruments for educational practices (ICEE 1982).

Area of the Study: The study area was Benue State, Nigeria. Benue is within the North Central Zone of Nigeria. It is made up of 23 local government areas with three educational zones (A, B and C). Educational Zone A is made up of eight local government areas which are Tarka with six (6) secondary schools, Ukum with twenty (20) secondary schools, Ushongo with seventeen (17) secondary schools, Vandeikya with twenty-six (26) secondary schools, Konshisha with twenty-one (21) secondary schools, Kwande with forty-six (46) secondary schools and Logo with ten (10) secondary schools.

Zone B has six (6) local government areas which are: Buruku with twenty (20) secondary schools, Gboko with forty-five (45) secondary schools, Guma with fourteen (14) secondary schools, Gwer West with eleven (11) secondary schools and Makurdi with forty-five (45) secondary schools.

Zone C has nine (9) local government areas which are: Ado with six (6) secondary schools, Agatu with twelve (12) secondary schools, Apa with fourteen (14) secondary schools, Obi with six (6) secondary schools, Ogbadibo with nineteen (19) secondary schools, Ohimini with eleven (11) secondary schools, Okpokwu with fourteen (14) secondary schools, Otukpo with thirty-four (34) secondary schools and Oju with thirteen (13) secondary schools.

Population of the Study: The population of the study comprised all year three (SSIII) senior secondary school students who enrolled for the May/June/July 2006 Biology Senior Secondary School Certificate Examination of WAEC and NECO in the three education zones of Benue State. This population was chosen because it was assumed they should have covered the WAEC and NECO Biology syllabuses. There were 35,000 students that sat for WAEC 2006 Biology examination and 42,193 students for NECO Biology examination (WAEC and NECO sources).

Sample and Sampling Technique: One thousand eight hundred (1800) students formed the

sample for this study. The multi-stage stratified sampling technique was used for the study.

Instrument for Data Collection: The instruments for this study consisted of WAEC and NECO 2000 – 2002 objective Biology questions respectively. The Biology objective questions for each year are made up of 60 items for both WAEC and NECO.

Method of Data Collection: The instruments were administered to the students by trained research assistants and senior Biology teachers of the selected schools under the supervision of the researcher. The instruments were administered under similar conditions as given by the examination bodies. The multiple matrix sampling technique was used in the data collection.

Method of Data Analysis: The Maximum Likelihood Estimation technique of the BILOG MG Computer Programme was used to analyze the data collected. This technique was used to answer the research question while the hypothesis was tested using the two (2) independent groups t-test analysis of the SPSS computer programme at 0.05 level of significance.

RESULTS AND DISCUSSION

Research Question One

What are the standard errors of measurement of the test items of the Biology examinations conducted by NECO based on one-parameter model?

The standard errors of measurement (S.E.M) of the test items of Biology examination conducted by NECO for the years 2000, 2001 and 2002 based on IRT one – parameter model are shown in Table 1. The standard errors (SE) ranged from 0.30 of item 26 and 27 to 0.49 of item 55 for the year 2000. The S.Es are generally low with all the items (100%) having S.E below 0.50, indicating high reliability. For year 2001, the standard errors of measurement of the items ranged from 0.23 to 0.40. All items (100%) have S.E below 0.50 (high reliability). In the year 2002, the S.Es. of the items ranged from 0.38 to 0.79. 85% of the items have S.E of 0.50 and below, this also indicates high reliability. Typically, S.E.M of 0.50 and below is described as low (high reliability), while S.E.M. above this value is described as high (low reliability).

Table 2 shows the standard errors of measurement of the test items of Biology examinations

Table 1: Standard errors of measurement of biology examinations conducted by NECO for the years 2000, 2001 and 2002 based on one-parameter model

Item	*S.E 2000*	*S.E 2001*	*S.E 2002*
1	0.31	0.21	-
2	0.34	0.25	-
3	0.32	0.25	0.46
4	0.31	0.23	0.46
5	0.33	0.41	0.45
6	-	0.25	0.49
7	0.40	0.25	0.41
8	0.32	0.25	0.44
9	0.41	0.23	0.45
10	0.31	0.27	0.41
11	0.32	0.24	0.43
12	0.31	0.26	0.53
13	0.44	0.25	0.79
14	0.31	0.22	0.55
15	0.38	0.29	0.43
16	0.30	0.25	0.63
17	0.31	0.24	0.43
18	0.34	0.24	0.43
19	0.33	0.29	0.47
20	0.33	0.33	0.54
21	0.31	0.23	0.45
22	0.32	0.25	0.70
23	0.35	0.24	0.58
24	0.30	0.25	0.44
25	0.36	0.26	0.45
26	0.30	0.23	0.42
27	0.30	0.47	0.46
28	0.30	0.29	0.45
29	0.31	0.25	0.43
30	0.35	0.24	0.43
31	0.29	0.25	0.56
32	0.29	0.24	0.49
33	0.32	0.36	0.56
34	0.33	0.27	0.45
35	0.33	0.24	0.65
36	0.40	0.23	0.42
37	0.30	0.23	0.43
38	0.32	0.27	0.41
39	0.31	0.23	0.67
40	0.32	0.25	0.45
41	0.31	0.26	0.72
42	0.33	0.25	0.42
43	0.31	0.25	0.44
44	0.31	0.24	0.46
45	0.32	0.31	0.55
46	0.36	0.29	0.45
47	0.28	0.24	0.65
48	0.34	0.24	0.45
49	0.32	0.27	0.48
50	0.30	0.26	0.44
55	0.49	0.27	0.40
51	0.33	0.24	0.44
52	0.37	0.26	0.45
53	0.36	0.27	0.45
54	0.36	0.27	0.44
56	0.30	0.24	0.41
57	0.30	0.25	0.45
58	0.28	0.30	0.45
59	0.32	0.24	0.42
60	0.30	0.27	0.43

conducted by WAEC for the years 2000, 2001 and 2002 based on the one – parameter model of IRT.

The results for the year 2000 showed that the items have their S.E. ranging from 0.01 for item number 4 to 0.74 for item 53. However, most items had the S.E. within the range of 0.20 to 0.30 (54% and 43% respectively) indicating high reliability.

Biology items for the year 2001 have their S.Es. ranging from 0.22 (item 3) to 0.74 (item 53). Most items were found with S.E between 0.20 and 0.30 (56% and 43% respectively) indicating high reliability.

It was, also, found that test items for the year 2002 had the S.E. range from 0.63 (item 33) to 1.35 (item 40). Majority (65%) of the items were however, with S.E. of 0.60 (low reliability).

There is no significant difference in the mean standard errors of measurement of the items in the Biology examinations conducted by NECO and WAEC from the year 2000 – 2002 based on one parameter model of IRT.

To test this hypothesis, the t–test statistics was used. The result for the year 2000 based on one–parameter model of IRT is shown on Table 3 from the result, it was seen that the t-statistic (t-stat = 2.04) was higher than the t-critical (t-critic = 1.98) at 0.05 level of significance. By this, the null hypothesis of no significant difference in the mean standard errors of measurement of items in the Biology examinations conducted by NECO and WAEC is rejected. So, there is a significant difference.

Table 4 has the result of the t-test of the standard errors of NECO and WAEC for the year 2001 based on one-parameter model of IRT.

The result in the Table shows that the t-statistic (t stat = 3.26) was higher than the t-critical (t-critic = 1.98). This means that there is a significant difference in the mean S.E. of Biology items of NECO and WAEC of the year 2001 based on the one–parameter model.

In Table 5 is the result of the t-test of the S.E. of Biology test items administered by NECO and WAEC for the year 2002 based on one – parameter model of IRT.

The result revealed that the t-statistic (t-stat = 11.04) of the items was higher than the t-critical (t-critic = 1.98) which showed that there is a significant difference in mean S.E. of Biology test items administered by NECO and WAEC.

Table 2: Standard errors of measurement of biology examinations conducted by WAEC for the years 2000, 2001 and 2002 based on one-parameter model

Item	*SE 2000*	*SE 2001*	*SE 2002*
1	0.23	0.23	0.87
2	0.31	0.31	0.65
3	0.22	0.22	0.64
4	0.01	0.01	0.72
5	0.25	0.25	0.80
6	0.24	0.24	0.63
7	0.22	0.22	0.75
8	0.22	0.22	0.79
9	0.26	0.26	0.73
10	0.27	0.27	0.62
11	0.35	0.35	0.66
12	0.30	0.30	0.76
13	0.25	0.25	0.64
14	0.33	0.33	0.86
15	0.26	0.26	0.68
16	0.25	0.25	0.92
17	0.24	0.24	0.68
18	0.26	0.26	0.63
19	0.30	0.30	0.7
20	0.27	0.27	0.73
21	0.24	0.24	0.65
22	0.30	0.30	0.63
23	0.33	0.33	0.64
24	0.26	0.26	0.68
25	0.29	0.29	0.71
26	0.32	0.32	0.63
27	0.28	0.28	0.65
28	0.35	0.35	0.91
29	0.23	0.23	0.61
30	0.35	0.35	0.65
31	0.26	0.26	0.65
32	0.29	0.29	0.67
33	0.37	0.37	0.63
34	0.23	0.23	0.83
35	0.31	0.31	0.65
36	0.31	0.31	0.64
37	0.32	0.32	0.63
38	0.26	0.26	0.69
39	0.23	0.23	0.65
40	0.31	0.31	1.35
41	0.26	0.26	0.67
42	0.37	0.37	0.65
43	0.28	0.28	0.69
44	0.30	0.30	0.72
45	0.23	0.23	0.70
46	0.26	0.26	0.65
47	0.33	0.33	0.71
48	0.32	0.32	0.64
49	0.33	0.33	0.74
50	0.36	0.36	0.71
51	0.31	0.31	0.67
52	0.35	0.35	0.71
53	0.74	0.74	0.66
54	0.46	0.46	0.71
55	0.31	0.31	0.68
56	0.33	0.33	0.63
57	0.37	0.37	0.66
58	0.36	0.36	0.64
59	0.30	0.30	0.65
60	0.36	0.36	0.67

Table 3: 2000 S.E. one-parameter model t–test: Two–sample assuming equal variance

	NECO	*WAEC*
	Variable 1	*Variable 2*
Mean	0.32	0.30
Variance	0.00	0.01
Observation	60	60
Pooled Var.	0.01	
Hypothesiz	0	
df	118	
t Stat	2.04	
P (T<=t) one	0.02	
t Critical one	1.65	
P (T<=t) two	0.04	
t Critical two	1.98	

Table 4: 2001 S.E. one-parameter model t-test. Two-sample assuming equal variances

	NECO	*WAEC*
	Variable 1	*Variable 2*
Mean	0.23	0.30
Variance	0.02	0.01
Observation	60	60
Pooled Var.	0.01	
Hypothesize	0	
df	118	
t Stat	-3.32	
P (T<=t) one	0.00	
t Critical one	1.66	
P (T<=t) two	0.00	
t Critical two	1.98	

Table 5: 2002 S.E. one-parameter model t –test: two–sample assuming equal variances

	NECO	*WAEC*
	Variable 1	*Variable 2*
Mean	0.47	0.70
Variance	0.02	0.01
Observation	60	60
Pooled Var.	0.01	
Hypothesiz	0	
df	118	
t Stat	-11.04	
P (T<=t) one	3.31	
t Critical one	1.66	
P (T<=t) two	0.04	
t Critical two	1.98	

Discussion of Research Questions One and Two

S.E. – NECO: The difficulty index of every item in a test is accompanied by its standard error in latent trait analysis and the smaller the standard error the better the item (Baumgartner 2002).

Based on this model, the S.E. of items for the three years were low (high reliability) with items for 2000 and 2001 having the lowest S.Es. However, 71% of items in 2001 have S.E. of 0.20, while 75% of items in 2002 have S.E. of 0.30. Thus, the 2001 NECO items were the best (most reliable).

S.E. WAEC: Research question two is on the standard errors of items in the Biology examination conducted by WAEC. As revealed in Table 2 which is based on one–parameter model, the items S.Es ranged from 0.01 to 0.74 for the year 2000. This range shows that the items (97%) had low standard errors, which makes the items reliable. Items for 2001 had their S.Es ranging from 0.22 to 0.74. These items, (99%) also, had relatively low S.Es. The S.E.s for the 2002 items were from 0.63 to 1.35. No item in this year has desirable S.E. value. The S.E.s of the items in the year 2002 were relatively high, they were highest among the three years (lowest reliability).

The various ranges of item S.Es for NECO and WAEC for the three years reflect a good precision of the test items, and implying that the items are sufficiently reliable. These results are supported by the findings of Gallini (1983), Harrison (1986), Korash (1995), Nkpone (2001) and Akindele (2003). These researchers suggested that test developers should increase the number of relevant test items in order to improve the accuracy of estimation. This is evident in Table 1 as, item 6 was not included in the item calibration for year 2000 just as items 1 and 2 for year 2002.

According to Baumgartner (2002), standard error of measurement is an estimate of how one should expect a test score to vary due to measurement error, that is, lack of reliability. There are various sources of measurement errors. Bock (1972) was of the view that the estimated standard error becomes exact, that is zero, as the sample size increases, which means that sample sizes can influence the standard error of items in a test. Other sources of measurement errors, according to Livingston (1988), include selection of specific items to test a general ability, time of testing and the person who grades or scores the test. Livingston (1988), also, suggested ways to determine these influences, such as giving the same subject two different variants of the test (alternate forms reliability), splitting the test into two similar halves (split–half reliability), and analyzing the data from the individual test items

(internal consistency). For time influence, he suggested that the subjects be given the test two or more different times (stability) and two or more graders should score the same test (inter-rater reliability).

Hypothesis One

The purpose of this hypothesis was to compare the standard errors of measurement of items of the Biology examinations conducted by NECO and WAEC from 2000 – 2002 based on one parameter model of IRT.

The test of difference in the standard errors of measurement of items of the Biology examinations conducted by NECO and WAEC for 2000 based on the one –parameter model (shown in Table 3) revealed that there is a significant difference in the standard errors of measurement of items of the Biology examinations conducted by NECO and WAEC.

The S.Es of NECO and WAEC Biology items for 2001 based on one-parameter model had significant difference as seen on Table 4. There was, also, an observed difference in the S.E. values of NECO and WAEC items as shown on Tables 1 and 2, with NECO items having S.E. values ranging from 0.23 to 0.40 while WAEC items ranged from 0.22 to 0.74; indicating wider range with WAEC items. Thus, the reliability levels for NECO items were higher than those of WAEC.

The test for difference of the S.Es in the NECO and WAEC Biology items on Table 5 for 2002, based on one parameter model, showed that there was no significant difference in the S.Es of Biology items in the examinations conducted by NECO and WAEC. This was, also, evident in the S.E. values of the items on Tables 1 and 2 where the items from the two examination bodies had relatively high S.E. values.

CONCLUSION

The answer to research question one revealed that the standard errors of items of the Biology examination conducted by NECO were small and this high reliability. This means also the case with research question two. Therefore, based on the data analyzed in this study, the conclusion is that the items of the Biology examinations conducted by NECO and WAEC were reliable with NECO items having the lowest range of S.Es indicating very high reliability.

RECOMMENDATIONS

It is recommended that the item response theory (IRT) analysis should be adopted by all examination bodies in Nigeria. This will enable the examiners place their examinees on the correct ability level because the IRT analysis is able to describe the test items and the abilities of the examinees. The IRT analysis could take care of the issue of superiority of one examination over the other.

REFERENCES

Akindele BP 2003. *The Development of Item Bank for Selection Tests into Nigerian Universities: An Exploratory Study.* Ph.D Thesis, Unpublished. Ibadan: University of Ibadan, Nigeria.

Assessment in Education 2007. From <www.doe.virginia.gov/VDOE/Assessment/sem.htm/> (Retreived 14/01/09).

Baunngartner TA 2002. *Conducting and Reading Research in Health and Human Performance.* 3rd Edition. New York: MC-Gram Hill High Education.

Birnbaum A 1968. Some latent trait models and their use in inferring an examinee's ability, In: FM Lord, MR Norvick (Eds.): *Stastitical Theories of Mental Test Scores.* Reading MH: Addison-Wesley.

Bock RD 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometricka,* 37: 29-51.

Gallini JA 1983. A Rasch analysis of Raven item data. *The Journal of Experimental Education,* 1: 27-32.

Harrison DA 1986. Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics,* 11(2): 97-115.

Hogan TP, Agnello J 2004. An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement,* 64: 802-812.

International Centre for Educational Evaluation ICEE 1982. *Education Research in Nigeria.* Institute of Education. University of Ibadan.

Item Response Theory, Reliability and Standard Error. From <http://www.wordengine.jp/research/pdf/IRT_reliability_ and_standard_error.pdf> (Retrieved on February 3, 2011).

Itsuokor DE 1986. *Essentials of Tests and Measurements*. Ilorin: Woye and Sons.

Izard JF, White JD 1980. The use of latent trait models in the development and analysis of classroom tests. In: D Spearitt (Ed.): *The Improvement of Measurement in Education and Psychology: Contributions of Latent Trait Theories.* Australia: Australian Council for Educational Research, pp. 65-72.

Korashy AF 1995. Applying the Rasch model to the selection of items for mental ability test. *Educational and Psychological Measurement, National Council on Measurement in Education,* 55(5): 753 – 763.

Lord FM, Novick MR 1968. *Statistical Theories of Mental Test Scores.* Massachusetts: Addison-Wesley.

Livingstone SA 1988. Reliability of Test Results. In: JP

Keeves (Ed.): *Educational Research, Methodology and Measurement. An International Handbook.* New York: Pergamon Press, pp. 115-127.

MacDonald P, Paunonen SV 2002. A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, National Council on Measurement in Education,* 62(6): 91-943.

Mehrens WA, Lehamm IJ 1981. *Measurement and Evaluation in Education and Psychology*. 3rd Edition. New York: Holt, Rinehart and Winston.

Nkemakolam E 2003. Taxonomy of educational measures. In: BG Nworgu (Ed.): *Educational Measurement and Evaluation: Theory and Practice*. Revised Edition. Nsukka: University Trust Publishers, pp. 66-83.

Nkpone HL 2001. *Application of Latent Trait Models in the Development and Standardization of a Physics Achievement Test for Senior Secondary Students*. Ph.D Thesis, Unpublished. Nigeria: University of Nigeria, Nsukka.

Nworgu BG 1992. *Educational Measurement and Evaluation: Theory and Practice*. Awka: Hallman Publishers.

Student Assessment Standards 2007. From <www.tea.state.tx.us/student.assessment/taks/standards/sem.pdf (Retreived 05/03/09).

Technical Education 2007. From <www.ed.sc.edu/caw/tecnical/sido13.htm> (Retrieved 14/01/09).

Wright B, Wainer H 1980. Robust estimation of ability in the Rasch model. Psychometrika, 45 (3), 373-391. From <web.sav.edu/waterstreetMaryA/WEW%20intro% 20to%20test%2> (Retrieved 20/01/2009).

Wood R 1980. Item analysis. In: JP Keeves (Ed.): *Educational Research Methodology and Measurement.* New York: Pergamon Press, pp. 50-55.