

COMPARISON OF PSYCHOMETRIC PROPERTIES OF MULTIPLE-CHOICE TEST USING CONFIDENCE AND NUMBER RIGHT SCORING AMONG SENIOR SECONDARY SCHOOL STUDENTS IN IBADAN METROPOLIS

BY

**Fehintola, J.O. Ph.D & Akingbade Olumide: Department of Guidance and Counselling,
University of Ibadan, Nigeria**

E-Mail: jof677@yahoo.com/joseph.fehintola@gmail.com/fehintola.j@dlc.ui.edu.ng

Abstract

This study investigated the comparison of psychometric properties of multiple-choice test using confidence and number right scoring among senior secondary school students in Ibadan metropolis. The study adopted a descriptive design of survey type. The population for the study consisted of Senior Secondary School two (SSS II) students in Ibadan Metropolis, Oyo state, Nigeria. A sample of 400 Agricultural science students was selected across 4 Local Government Areas in Ibadan metropolis, using purposive (mainly Agricultural Science Students) and random sampling techniques. The instrument used for the study was Agricultural science Multiple-choice Test. The 50 items Agricultural Science 4-option test was administered on the Students. Data collected were analyzed using paired samples t-test, Kuder-Richardson (KR-21), Cronbach alpha, and Fisher z-test. The results obtained revealed that significant difference existed in the difficulty indices with Number Right (NR) and Confidence Scoring Method (CSM) with mean of 55.42 and 44.01 respectively. Also, there was a significant difference in the CSM and NR in the discrimination indices with NR and CSM has mean of 0.57 and 0.52 respectively. It was found that NR significantly improved the difficulty and discrimination indices. Furthermore, the finding revealed that there was no significant difference between NR and CSM in the reliability coefficient. Based on these findings, it was recommended that number right scoring method should be used to assess Agricultural science students' performances because it makes test item appear moderate in terms of difficulty level and is very easy for students to guess the items right.

Keywords: Comparison, Psychometric Properties, Multiple Choice Test, Confidence and Number Right Scoring

Introduction

Assessment is a very important component of the educational process because it measures and gives dimension to students' learning outcome. Effectiveness of instruction is determined by the attainment of set goals and one major tool or device for such determination is test. In order to obtain information about the mastery knowledge of students, there are different categories of test used on different parameters; however, category based on test format becomes majorly used.

Today objective test has become popularly used assessment tool for measuring mastery ability and learning outcomes particularly the multiple-choice test because it is one of the most versatile and widely applicable test items for measuring different types of cognitive attributes effectively. It does so by measuring different types of learning outcomes in the areas of knowledge, comprehension, application, analysis and synthesis. Today, multiple choice tests are the most

profoundly respected and broadly utilized sort of objective test for estimation of information, ability, or accomplishment (Ben-Simon, Budescu, & Nevo, 2009; Lee & Winke, 2013).

A typical multiple-choice item made up of an introductory part otherwise known as stem which present problematic situation and several options (alternatives or choices) among which provide possible solutions to the problem. The alternatives include the correct answer and several plausible answers called distractors. More so, this type of test can be scored quickly, accurately, discourage arbitrary allocation of marks and with ease by teachers and even clerks. However, in spite the numerous significances associated with multiple choice tests, the unique limitation to its uses could be traced to its susceptibility to greater propensity to cheat or do blind guessing. Guessing enables examines without the ability to solve a particular item to be credited with unmerited scores where test-wise students would be also score higher scores more than the mastery level they possess in the subject matter in question. Consequently, this makes reliable and valid appraisal about the performance of students to be difficult (Ajayi, 2013). As a way to ensure the continuity of merit of objective test in general and that of multiple-choice test in particular and hence to sustain their continued usefulness. However, a number of scoring procedures have been actualized (Rojas, 2012). Examples of those scoring methods include among others; negative marking, partial-credit method, retrospective correction for guessing, number right, logical-choice weight, confidence scoring etc.

Ben-Simon *et al.*, (2009) opined that multiple-choice items are scored using the non-conventional partial-credit scoring (PCS) method which allows a more accurate measurement of student knowledge. PCS is a method that captures information about a student's degree of level of knowledge with respect to each choice presented in relation to a test item. Although a variety of different partial-credit scoring methods exists. The three main formats include; (a) the liberal multiple-choice test which allows students to select more than one answer to a question if they feel uncertain of the correct one (Ajayi, 2013). Duchy, Kyndt, Baeten, Pottier and Veestraeten (2009), reviewed that for a 5-choice test, award 3 points for a single correct answer. Candidates who choose two answers including the correct 1 get 2 points; candidates who choose three answers including the correct 1 get only 1 point. (b) Elimination Testing (ET), in this, candidates are asked to mark as many incorrect options as they can identify in a question with N options. When students' responses are to be scored by using ET, the instruction is stated as cross out all the alternatives that are considered to be incorrect. As such, 1 point is awarded for each incorrect choice that is identified, but N-1 points are deducted if the correct option is identified as incorrect. (c) Confidence weighting (CW) in CW, students have to indicate what they believe is the correct answer and how confident they are about their choice.

Ben-Simon *et al.* (2009) compared seven different scoring methods awarding partial credits. However, none of the approaches could be regarded as the best methods, considering the validity and reliability. Despite the fact that partial credit scoring methods can measure partial knowledge; student's ability cannot be precisely measured as it doesn't take care of guessing. Even though they award students for partial knowledge, they proffer no solution for guessing. Partial-credit scoring methods do not seem to affect test reliability as compared to any of the conventional scoring method. As noted by Jennings and Bush, (2006), no partial-credit scoring method has been identified as superior compared to the conventional multiple-choice scoring

methods. What characterized the empirical studies reviewed perhaps the most are their frequently contradictory results. What appears to work well in a particular context does not in another.

Conventionally, multiple choice tests have been scored using a 'rights minus wrong' correcting models proposed by Kurz (2009), otherwise known as conventional Negative Marking (NM) which is predominant and aimed to penalize the student for incorrect responses. The fundamental idea behind this scoring method is that students acknowledge they will lose marks for incorrect answers (Betts, Elder, Hartley & Trueman (2009). As a result, students are discouraged to guess, and this is expected to increase test reliability and validity because the test score is a truer reflection of a student's ability.

This method could have been the best approach to correction for guessing but study conducted by (Barton, 2004; Karandikar, 2010) shown that slight improvement was accounted for and also, NM specifically examine true/false items. Because of the high susceptibility to guessing, the results cannot simply apply to items with more alternative choices. Moreover, there is no consistency concerning applied analysis criteria, test length, knowledge domains and test instructions in these studies. Clarity about the amount of the penalty given for incorrect answers is ambiguous. Some authors (Bar-Hillel, Budescu, & Attali, 2005; Espinosa & Gardezabal, 2010) state that an effective penalty that discourages guessing should exceed the standard penalty of $1/(n-1)$. Beyond this approach, studies (Barton, 2004) report an increase in reliability or validity when negative marking is implemented.

Hence, attempts have been made to develop other scoring procedures which would have fewer defects. Since no empirical evidence is available in the literature that helps to direct the choice between either approach, the researcher therefore focuses on alternative approach to non-conventional partial-credit scoring and conventional negative marking. A possible alternative scoring procedure, for example, van der Vleuten., (2010) reported on an alternative method to conventional negative marking, a Number right, (NR) where students are instructed to choose an option as the answer and one point is awarded for each correct answer. According to Ajayi (2013) reported that, in a situation whereby an objective item has four options A, B, C, D and that the key is C such the correct answer would be assigned one mark. He proposed that the simplest way of assigning mark to objective test item is number right scoring method. Under this method, we are signaling to students that partial knowledge is not important, guessing is an acceptable practice and it is all right to have misconceptions. Number right scoring does not actually leave partial information unrewarded as majority could think.

However, in reality, we rarely have the opportunity to guess in our decision making. Also, many important and irreversible decisions are mostly made based on partial knowledge because time does not permit us to delay decision making till full knowledge is attained. Errors due to misconceptions can be disastrous in critical fields such as medicine, aerospace and engineering. In reality, a person could face termination, lawsuit or jail sentence for errors made due to misconceptions (Barton (2004). Therefore, the aim of the researcher is to investigate the psychometric properties of multiple-choice test using only number right and confidence scoring

methods. This procedure as it would be employed in this study is known as Confidence Scoring method

Statement of the Problem

The issue of identifying the most reliable scoring method of multiple-choice test presents a difficult situation for effective measurement of complex learning outcome. Though the multiple-choice test is a task of human decision making of selecting the correct answer amongst several alternatives, yet there are no findings identifying the best scoring method for use. It is expected that a good test should be able to measure that which is purport to measure effectively and consistently. And if test must be given the kind of seriousness it deserves, it is widely discovered that the marking system determine the precision of the reliability and validity of any given test. There is need for assessment reform that seeks measures that will be better informed teaching and learning and provide more useful and reliable feedback regarding the outcomes. Therefore, the researcher decided to determine the comparison of psychometric properties of Multiple-Choice Test Using Confidence and Number Right Scoring among Senior Secondary School Students in Ibadan Metropolis.

Purpose of the Study

The study was designed to examine the comparison of psychometric properties of Multiple-Choice Test Using Confidence and Number Right Scoring among Senior Secondary School Students in Ibadan Metropolis. Based on this, the specific objectives of this study were to: (i) determine which of difficulty index is better using scores generated from number right and confidence scoring methods (ii) ascertain which of discriminating index is better using number scoring and confidence scoring procedures (iii) compute reliability of Agricultural science paper using scores generated from number right and confidence scoring methods (iv) compute validity of Agricultural science paper using scores generated from number right and confidence scoring methods and (v) determine which of reliability and validity coefficient is better using the Number right and Confidence scoring methods.

Research Questions

Research question 1: What is the difference in the difficulty index of number right and confidence scoring methods?

Research question 2: What is the difference in the discrimination index of number right and confidence scoring methods?

Research Question 3: Would there be any difference in the reliability of the two scoring methods on Agricultural science Multiple-Choice Questions?

Research question 4: Would there be a difference in the validity of the two scoring methods on Agricultural Science Multiple-Choice Questions?

Research question 5: Which of the Scoring method is better for reliability test using the Number right and Confidence scoring methods?

Research question 6: Which of the Scoring methods is better for validity test using the Number right and Confidence scoring methods?

Methodology

Descriptive research design of survey type was used for this study. The design aided the researcher to make effective comparison of the difficulty and discrimination levels of confidence scoring and number right scoring procedure on multiple-choice test for their reliability and validity. The population for the study consisted of all Senior Secondary School two (SSS II) students in Ibadan Metropolis in Oyo State, Nigeria who offered Agricultural Science in the metropolis. A sample of four hundred (400) students was drawn across 4 LGAs in Ibadan Metropolis out of 11 LGA that are made up of Ibadan metropolis. 100 candidates were purposively picked in the sense that they must be students that are offering Agricultural Science subject and randomly picked because they are more than the researchers can cope with for the study in each LGA selected for this study. A simple random technique was used to select 4 LGA which included; Ibadan North, Ibadan South West, Akinyele and Ibadan North West. All the schools are mixed and had been accredited by WAEC of having minimum requirement for offering Agricultural Science, in terms of facilities and human resources. There were 266 and 134 male and female students respectively and their age ranges between 12.5 to 14.0 years.

Research instrument used for collecting data consist of a multiple-choice format of a 4-alternative (A, B, C, D) options. The instrument were made up of fifty (50) test items which were adapted from past West African Examination Council (WAEC) question of 2018. Since the items had been validated by the WAEC and standardized, they were administered on the 400 SSII already selected for the study. The items covered the entire Agricultural Science syllabus for SS I to III.

The 50 items Agricultural Science Multiple-Choice Tests were administered on the selected students in selected Senior Secondary Schools in four L.G.A. of Ibadan Metropolis with the assistance of the subject teachers in each school. The time frame for the test was 45minutes. The researcher collected the 400 scripts and scored the testees using confidence and number right scoring methods. In using number right, 2marks was assigned to correct answer each and zero mark to incorrect option chosen by testees. In confidence scoring procedure, the testee indicates on his answer script the level of degree of certainty he has on the answer to the item. The discrimination and difficulty indices for each item when scored with number right and confidence scoring methods were determined using item analysis formulae and their mean differences were compared using Paired samples t-test. The internal consistency reliability and concurrent validity of the test was found using K-R₂₁formular and Cronbach alpha, respectively, with the aid of SPSS version 21.0. The reliability values obtained from the confidence and number right scoring methods now compared using Fisher z- Transformation of r-values. Z-scores (standard scores) are raw scores that have been adjusted for mean and standard deviation. Fisher z- transformation is an approximate variance-stabilizing for r-values. More so, the two validity indices obtained by using confidence scoring and number right scoring were compared using Fisher z-test.

Results

Research question 1: What is the difference in the difficulty index of number right and confidence scoring methods?

Table 1: Result of Multiple-Choice Test scored with Number Right and Confidence Scoring Methods

Item no	NR-difficulty	CSM-Difficulty	Item No	NR-difficulty	CSM-Diff.
1	53.20	39.80	26	65.70	48.10
2	62.50	43.50	27	39.80	20.80
3	60.60	67.10	28	47.70	39.40
4	53.70	45.80	29	46.30	39.80
5	47.70	32.40	30	62.50	50.90
6	47.20	44.40	31	48.80	31.90
7	61.60	46.30	32	60.60	46.30
8	63.90	46.30	33	56.00	44.00
9	57.90	50.50	34	60.60	49.10
10	51.40	41.70	35	42.60	27.80
11	46.30	33.30	36	78.70	57.40
12	53.20	43.10	37	69.40	53.20
13	52.80	42.60	38	60.20	51.90
14	51.40	38.40	39	58.80	44.40
15	54.60	44.00	40	55.10	40.30
16	60.20	43.10	41	75.50	61.60
17	50.90	30.10	42	47.20	60.60
18	48.10	38.00	43	69.00	58.30
19	48.60	44.90	44	38.00	31.00
20	56.00	54.20	45	52.70	43.50
21	71.80	49.10	46	51.90	42.10
22	46.30	25.00	47	54.20	48.60
23	51.40	53.70	48	49.50	42.10
24	68.10	42.60	49	51.90	39.80
25	63.90	49.50	50	44.90	38.40

Table 1 shows the result of items description based on examinees' responses to each item when Number right and Confidence scoring methods were used. Both number right and Confidence scoring shown that none of the items is very difficult for the examinees. Two items (27 and 44) appeared difficult for examinees when Number right was used while sixteen items (1, 5, 11, 14, 17, 18, 22, 27, 28, 29, 31, 35, 40, 44, 49 and 50) were considered difficult with Confidence scoring methods for examinees. With number right, Item 2, 7, 8, 21, 24, 25, 26, 30, 36, 37, 41 and 43 appeared easy to testees while only item 3 and 41 recorded to be easy when scored with confidence method. This implies that only 36 items were considered suitably moderate with Number right and 32 items only for Confidence Scoring method. Paradoxically, highest number of performance was recorded with Number right scoring over Confidence scoring procedure

since majority of the items appeared to be moderate at measuring cognitive capacity of examinees.

Table 2: T-test Analysis summary showing Difficulty Index difference based on Confidence and Number Right Scoring Methods

Variable	Scoring-Method	N	Mean	SD	t	df	p	η^2
Difficulty Index	NR	50	55.4180	8.93835	11.257	49	<0.05	0.72
	CSM	50	44.0140	9.26137				

N = number of test item (50) and Number of students used =216

Table 2 reveals that there was a significant difference in the difficulty index of number right and confidence scoring method; $t(49) = 11.257$, $p < 0.05$, $\eta^2 = 0.72$. Thus, the null hypothesis was rejected. The table further reveals that NR (mean= 55.4180) displayed higher difficulty index than CSM (mean= 44.0140). Size of effect ($\eta^2 = 0.72$) reveals that scoring method had large effect on difficulty index of agricultural science achievement; that is, scoring method it accounted for 72% change in scoring method discriminating index.

Research question 2: What is the difference in the discrimination index of number right and confidence scoring methods?

Table 3: Summary of Item Discrimination Analysis Table showing the Result of Multiple-Choice Test scored with Number Right and Confidence Scoring Methods

Item no	NR-Discrimination	CSM-Discrimination	Item No	NR-Discrimination	CSM-Discrimination
1	.64	.54	26	.57	.76
2	.66	.41	27	.37	.34
3	.66	.51	28	.75	.56
4	.63	.40	29	.54	.54
5	.79	.52	30	.58	.52
6	.65	.69	31	.60	.40
7	.64	.54	32	.49	.52
8	.52	.65	33	.69	.66
9	.71	.75	34	.62	.69
10	.69	.63	35	.50	.39
11	.54	.37	36	.31	.57
12	.44	.42	37	.30	.27
13	.56	.52	38	.48	.52
14	.69	.51	39	.38	.43
15	.70	.64	40	.79	.64
16	.67	.73	41	.36	.38
17	.72	.44	42	.78	.47
18	.57	.56	43	.42	.48
19	.64	.58	44	.41	.36
20	.69	.46	45	.50	.46

21	.47	.74	46	.78	.64
22	.48	.30	47	.66	.62
23	.69	.20	48	.55	.56
24	.23	.56	49	.24	.33
25	.57	.56	50	.55	.51

Table 3 reveals item discrimination analysis of Agricultural science four options test using Number right and confidence scoring procedures. The result shows that number right scoring was effective at measuring cognitive status of examinees with 43 items were considered to be **very good**, while confidence scoring measured 40 test items to be very good. Five items (27, 36, 37, 39, and 41) were measured to be good and two items were considered fair items with number right. Whist seven items (11, 22, 27, 35, 41, 44, and, 49) were measured of good items and three items (23, 37, and 38) were considered fair items with confidence scoring.

Table 4: t-test Analysis summary showing Confidence and Number Right on test Discrimination Indices

Variable	Scoring-Method	N	Mean	SD	t	df	p	η^2
Discriminating Index	NR	50	0.5694	.14348	2.426	49	<0.05	0.11
	CSM	50	0.5170	.12943				

N= number of item (50) and Number of testees used = 216

Table 4 reveals that there was a significant difference in the discriminating index of four-option multiple-choice test using number right and confidence scoring method; $t(49) = 2.426$, $p < 0.05$, $\eta^2 = 0.11$. The table further reveals that NR (mean= 0.5694) displayed higher discriminating index than CSM (mean= 0.5170). Size of effect ($\eta^2 = 0.11$) reveals that scoring method had moderate effect on discriminating index of agricultural science achievement; that is, scoring method it accounted for 11% change in scoring method discriminating index.

Research Question 3: Would there be any difference in the reliability of the two scoring methods on Agricultural science Multiple-Choice Question?

Table 5: Comparison of the reliability of Confidence Scoring and Number right Methods

Scoring Method	N	r	z	z^2	P
Number-right method	400	0.643	.767	-1.735	>0.05
Confidence Scoring method	400	0.844	.890		

From table 5, the number of items is 50 and the number of examinees is 400. The result reveals that there is no significant difference in the reliability of the two scoring methods on Agricultural science Multiple-Choice Question; $z = -1.735$, $p > 0.05$. Therefore, number right method did not significantly differ in reliability from confidence scoring method using multiple-choice test. By implication, none of the method is superior in-terms of the internal consistency of a test; that is, the use of any of the methods will not affect the reliability of any multiple-choice test. However,

by considering their r-values, Agricultural science multiple-choice test shows better reliability index when scored with Confidence scoring method (0.844) than Number right (0.643).

Research question 4: Would there be a difference in the validity of the two scoring methods on Agricultural Science Multiple-Choice Question?

Table 6: Comparison of the validity of Confidence Scoring and Number right Methods

Scoring Method	N	R	z	z ²	p
Number-right method	400	0.422	0.454	2.214	<0.05
Confidence Scoring method	400	0.542	0.611		

To answer this question the r-value obtained from concurrent validity report on the two scoring methods were compared using z-test. The result is presented in Table 6. From table 6, the numbers of items are 50, number of respondents are 400 and r-values are 0.422 and 0.542 for Number Right and Confidence Scoring methods, respectively. Using the z-table, the r-values when transformed become $z = 0.454$ and 0.611 , respectively. Comparing these z-values using Fisher z- test gave a z-value of 2.214 which is significant at the 0.05 level of significant ($p < 0.05$). This reveals that there is a significant difference in the validity of the two scoring methods on Agricultural science Multiple-Choice Question. Therefore, number right method significantly differs in validity from confidence scoring method using multiple-choice test.

Research question 5: Which of the Scoring method is better for reliability test using the Number right and Confidence scoring methods?

Table 7: Reliability coefficient summary showing variances in internal consistencies of scoring methods

Reliability Methods	Scoring Methods	
	CSM	NRM
Guttman's L2	0.8514	0.7096
Coefficient Alpha/KR21	0.8444	0.6434
Feldt-Gilmer	0.8463	0.6900
Feldt-Brennan	0.8459	0.6861
Raju's Beta	0.8444	0.6826

Table 7 reveals that CSM displayed a superior reliability value (α -0.844) than NRM (KR21-0.6434). This indicates that CSM have higher internal consistency than NRM. On this premise, CSM have better reliability.

Research question 6: Which of the Scoring methods is better for validity test using the Number right and Confidence scoring methods?

Table 8: Concurrent validity summary showing validity variances in the two scoring methods

Scoring Method	N	R	r ²
Number-right method	400	0.422	0.178
Confidence Scoring method	400	0.542	0.294

Table 8 reveals that confidence scoring method (CSM) ($r = 0.542$) recorded high concurrent validity over number right method (NRM) ($r = 0.422$). Size of effect according to Cohen (1988) reveals that when $r = 0.5$, it implies large effect and when $r = 0.3+$, it implies moderate effect. By implication CSM had large concurrent validity with standardized WAEC Agricultural science test. Coefficient of determination ($r^2 = 0.294$) reveals that Agricultural Science Multiple-Choice test and standardized WAEC test co-variance accounted for 29.4% increase in the test concurrent validity.

Discussion

The first research question examined the difference in the discriminating index of number right and confidence scoring method on agricultural science fifty (50) multiple-choice items. The result shows that number right scoring method displayed higher discrimination index than the confidence scoring method. This suggests that number right has a significant effect on discrimination index of achievement test. This study disagreed with Awodele, Faremi, Adetunji and Bamidele (2013) when they investigated the difficulty and discrimination indices of Logical-Choice Weight and Confidence Scoring Methods on Chemistry multiple choice test. Also, Odeyemi (2003) discovered that when responses are to be made with increased in confidence level, multiple-choice test become more and more difficult.

The second research question which was based on discrimination indices at the different confidence levels showed that the means increases as the level of confidence of examinees increase (Boyinbode, 1986). The study further showed that Number right scoring procedure has a significance effect on the difficulty index of Agricultural Science Multiple-Choice tests than confidence scoring method. Previous researchers have reported that tests that have items mostly of middle difficulty are of good discrimination, which mean that items that are not too difficult or too easy can reveal good discrimination when the upper and lower difference index is used. The point here is the fact that most of the items that constituted the test in this study have difficulty level from the middle value downward. This aligned with the report of Roja and Nazli, (2012) reported that different degrees of difficulties for the test taker are produced from different testing method.

On the third research question, to ascertain the difference in the reliability of the two scoring methods used in this study, the result showed that none of the method is superior in terms of the internal consistency of a test. Although by considering their r-values, confidence scoring has a

significant contribution than number right scoring method on the reliability of Agricultural science multiple-choice test. However, there have been conflicting reports in this regard. Awodele et al (2013) reported that confidence scoring method of multiple-choice test has a significant effect; z^2 statistic of 5.21 obtained at $p < 0.05$. Similarly, this study made use of a conventional and non-conventional scoring procedure. Paradoxically, this supported Rahimi, (2007) who found that when different test formats are used to measure certain ability, they lead to obtain different findings. In other words, the way of test administration may have some effects on the learner's performance and test results. Consequently, test performance seems to be greatly affected by test method.

Likewise, on the fourth research question which addressed the validity effect of the two scoring methods on Agricultural Science multiple-choice test, result reveals that Number Right has a better concurrent validity than Confidence scoring. This result corroborates with Ajayi (2013) who found that in the overall performance of agricultural science students in number right and confidence scoring methods, the Number Right had the greatest effect compare to Confidence Scoring Method which implies that students performed best in number right scoring method than the confidence scoring method.

The fifth research question examined the best scoring procedure among Confidence scoring and Number Right scoring procedures for reliability test. The finding showed that Confidence scoring approach displayed a superior internal consistency index than Number Right Method. Even still when other reliability methods were used, such as Guttman'L2, Feldt-Gilmer, Feldt-Brennan and Raju's Beta methods Confidence scoring still displayed very high reliability indices compared to Number right. This implies that CSM is better for reliability test. This result agreed with Awodele, Adediwura, Adetunji, Bamidele and Ajeigbe (2013) who reported in the Comparative Effectiveness of Logical-choice Weight and Confidence Scoring Methods on Reliability and Validity of Chemistry Multiple-choice Test Items, that Chemistry test yielded a greater reliability index with Confidence scoring method. Similarly, Odeyemi (2003) went further that the effect of confidence scoring on the reliability of multiple choices was significant at 0.05 level. It was found that confidence scoring method least favours the students. In order for learners to fully demonstrate his or her cognitive ability, Advocators of confidence testing have stated that knowledge is neither a dichotomous nor a trichotomous affair, which conventional multiple-choice tests seem to imply, but is continuous in the sense that there are varying degrees of knowledge.

Lastly, on account of better scoring a method which is the basis of sixth research question, the validity test, according to the result of this study, Confidence scoring method yielded high concurrent validity over Number right method of four option Multiple-choice test. This implies that Confidence scoring method is better for validity test than Number right scoring procedure. For precision and accuracy, it is expected that the schools should have enough valid and reliable tests for assessing their students' cognitive ability when they have been exposed to curriculum content areas as well as to prepare them for external examinations such as WAEC, NECO etc. therefore, CSM displayed higher which makes it more suitable for validity test than the latter.

Conclusion

Based on the findings of this study, the following conclusions were drawn. Number right scoring method has a considerably significant effect on difficulty and discrimination indices of achievement test. This implies that when learners' cognitive skills are to be captured in four-option questions, number right scoring is more effective than confidence scoring. Discrimination indices at the different confidence levels showed that the means increase as the level of confidence of examinees increase. However, Confidence scoring method is better method of scoring for reliability and validity test.

Recommendations

Based on the findings the following recommendations were made:

1. Public Examination Bodies should adopt Number right method in setting and scoring Agricultural Science Multiple-choice Achievement tests. This method is found to be the best scoring approach at discriminating between high intelligent and low intelligent students as well as very easy to score four-option test items
2. Teachers and Lecturers should develop skills on how to conduct and score Agricultural Science Multiple-choice tests using Confidence Scoring Methods since this approach is not widely used across Nigerian schools. This can be greatly achieved through organizing workshop on Testing and Scoring Approaches. This procedure could be better for reliable and valid tests.
3. Due to the fact that Confidence scoring captures cognitive status of students in four-option tests, Consultancy firms who conduct Aptitude tests for employees and promotion exercise for their workers should as well make use of Confidence scoring method
4. In order to help in proffering a suitable tool for knowledge evaluation, Confidence scoring method is recommended for formative and summative evaluation of learners by classroom teachers at all levels of Nigeria educational system as it has been found to be effective and efficient in reducing contribution to blind or random guessing of examinees on multiple-choice tests.

References

- Ajayi B. K (2007). The Effect of Four Scoring Methods on Multiple Choices Agricultural Science Test. An Unpublished Ph.D Thesis Institute of Education, University of Ibadan, Ibadan. 208 pages.
- Ajayi B. K. (2013). The Effect of Number Right and Confidence Scoring Methods on Multiple Choice Agricultural Science Test Scores. Institute of Education, Ekiti State University. *International Journal of Education and Research*. Vol. 1 No 8, 34 - 40.
- Awodele B.A., Adediwura A.D., Adetunji A.A., Bamidele E.F., and Ajeigbe T. O (2013). Comparative Effectiveness of Logical-choice Weight and Confidence Scoring Methods on Reliability and Validity of Chemistry Multiple-Choice Test Items in Nigerian Secondary Schools. *Journal of Educational and Social Research*, Vol. 3 (2), 47- 52.
- Barton, L.F. (2004). Modern Language testing at the turn of the century: Assuring that what we count counts. *Language testing*, 17(1), 1-42.

- Ben-Simon, A., Budescu, D.V., and Nevo, B. (2009). A Comparative Study of Measures of Partial Knowledge in Multiple Choice tests. *Applied psychology Measurement Vol.21. P 105*
- Betts, L. R., Elder, T. J., Hartley, J., and Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment and Evaluation in Higher Education, 34(1), 1-15*
- Boyinbode, I. R. (1986). Effect of Confidence level on some properties of True-false test answers. *Nigeria Journal of Educational Psychology Vol.5(I), 51 – 56.*
- Davies, P. (2002). "There is no confidence in multiple-choice testing," Proceedings of 6th CAA Conference, Loughborough: Loughborough University, pp.119-130.
- Duchy, F., Kyndt, E., Baeten, M., Pottier, S., and Veestraeten, M. (2009). The effects of different standard setting methods and the composition of borderline groups: A study within a law curriculum. *Studies in Educational Evaluation, 35, 174-182*
- Espinosa, M. P., and Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology, 54(5), 415-425*
- Jennings, S., and Bush, M. (2006). "A comparison of conventional and liberal (free-choice) multiple-choice tests," *Practical Assessment, Research and Evaluation, 11, no. 8, pp. 1-5*
- Karandikar, R. L. (2010). *On multiple choice tests and negative marking.* Current Science, 9998), 1042-1045.
- Kurz, J. C. (2009). Exploring the Reliability and Validity of Research Instruments to Examine Secondary School Principals' Authentic Leadership Behaviour and Psychological Capital. The University of Arizona. 105-123
- Lee, H., and Winke, P. (2013). *The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test.* Language Testing, 30(1), 99–123.
- Odeyemi J.O. (2003). *Comparison of the Psychometric Properties of Three Multiple-Choice Test Using the Confidence Scoring Procedure,* Unpublished M.Ed. Thesis. Ekiti State University, Ado-Ekiti.
- Rahimi, M. (2007). L2 Reading Comprehension Test: Does the Language of Presenting Items Affects Testess's Test Performance? *Journal of Social Sciences and Humanities of Shiraz University, 26(4), 67-86.*
- Roja, K., and Nazli, S. (2012). *A Fairness Issue: Test Method Facet and the Validity of Grammar Sub-tests of High Admissions Tests,* 219 pages.

Shank, P. (2010). Create Better Multiple-Choice Questions. *The American Society for Training and Development*, 27, Issue 1009, 35-52.

Ugbamadu, K. A., Onwuegbu, O. C., and Osunde, A. U. (2001). *Measurement and Evaluation in Education*. Benin-City, World of books Publishers.

Van der Vleuten, C. P. M. (2010). *Setting and maintaining standards in multiple choice examinations: Guide supplement 37.1 – viewpoint*. *Medical Teacher*, 32, 174 – 176.