ASSESSMENT OF RELIABILTY AND DISTRACTER EFFICIENCY OF THREE, FOUR AND FIVE OPTIONS ECONOMICS MULTIPLE CHOICE ITEMS USING CONFIDENCE SCORING PROCEDURE

[1]**JIMOH K. & [2]ADEDIWURA, A. A.**

[1&2]Department of Educational Foundations and Counselling,
Obafemi Awolowo University, Ile-Ife
jimoh.bukola@yahoo.com, yemtoy20002000@gmail.com

Abstract

The study ascertained the impact of number of options on distracter performance when confidence scoring was used and established the impact of numbers of options for multiple choice test items on the reliability of the test. The survey research design was adopted for the study and the population consisted of secondary school students in Osun State. The sample comprised 360 students randomly selected from 18 schools. The instrument used for the study was 2015 West Africa School Certificate Examination (WASCE) Economics test items. The Economics Achievement Test (EAT), which was the 4-options format adapted from the 2015 WASCE Economics items had been validated for used by the West African Examinations Council. The validity of the other types (3-options and 5-options formats) were determined and scrutinized by experts in Tests and Measurement and Economics teachers in the secondary school to judge its face and content validity as well as item arrangement. The corrections were incorporated into final version of the instruments. The EAT instrument of 3-options and 5-options were validated in a pilot study conducted using 40 senior secondary school II Economics students who were not part of the final sample size but in a different study area with similar characteristics. Given the responses of the respondents used in the pilot study, the 3-options and 5-options which consisted of 50 items were subjected to a measure of internal consistency using Kurder-Richardson 21 (K21) to ascertain the reliability of the instrument. The result of the K-21 for both 3-options and 5-options yielded coefficients of 0.79 and 0.83 respectively. Data collected were analysed using ANOVA, Kuder -Richardson Formula (KR-20) and Fisher's Z-Test with aid of FZT compotator. The results of the study showed that number of options had significant impact on distracters' performance when scored using confidence scoring ($F= 6.679$, $p <0.05$). The results also showed that for each pair wise comparison of 3/4-options ($z_{obt}= 0.640$), 3/5-options ($z_{obt}= 0.837$) and 4/5-options ($z_{obt}= 0.196$) at $p <0.05$ the difference in the reliability coefficients were not significant. The study therefore concluded that option length of multiple choice objective test item have impact on its reliability and distracter efficiency.

Keywords: Confidence scoring, Number right, Multiple choice items, Item discrimination, Item Difficulty, distracter performance

Introduction

Teaching activity may not be completed until the students taught are authentically assessed. One of the fundamental instruments for such assessment is test, which Omirin (1999) in Awodele (2013) called systematic method of gathering data for the purpose of making intra and inter-comparisons between individuals within class or in a school system. A

test may contain several items. Each item tends to confront the testee with a task to provide a means for observing its response to the task. Kolawole (2005) viewed a test as a systematic procedure for comparing the performance of an individual with designated standard of performance, thus as an instrument to elicit a sample of behavior or human traits or attributes. The trait being measured in testing may be achievement test, intelligent skill, personality.

Ugbamadu, Onwuegbu and Osunde (2001) defined a test as an instrument made up of questions or tasks designed and presented to individuals or testees to respond to independently and the results of which can be used to determine quantitative academic change in individuals and for the quantitative comparison of performance of different individuals or their level of achievement. Achievement test serves as a psychological instrument which the school teacher applies to find out the amount of knowledge that students have acquired in a specific course of study at a specific time. Achievement test consists of essay and objective test. The essay format requires the learner or testee to write a sentence, paragraph or long passage which demands subjective judgment regarding the quality of the written statement, while objective format is devoid of subjectivity because every expert arrives at precisely the same score. In primary, junior secondary school levels and entrance examinations to secondary and tertiary institutions objective test is the most commonly used test formats (Ajayi, 2012).

Objective tests have gained prominence in Nigeria, particularly the multiple-choice test because it is one of the most flexible, versatile and widely applicable test item for measuring different types of knowledge effectively; and it also measures different types of complex learning outcomes in the areas of application, analysis and synthesis. In addition, it can be scored quickly, accurately and with much ease by teachers and even clerks and students. Today, multiple-choice test is the most common and widely used assessment tool for the measurement of knowledge, ability and complex. The concept of multiple-choice questions is not and has been adopted in a variety of disciplines including Physical Education , Mathematics, Economics, Fine Arts and the physical sciences new (Bracey, 2000; Chan &Kennedy, 2002) . Ultimately, there is no one ideal assessment process and so too, multiple-choice questions testing has its advantages and disadvantage. Whatever purpose of a test or exam has a major factor in its success or failure as a good measuring instrument will be determined by the item types that it contains.

There are 2 types of the test items, direct and indirect. Indirect test items try to find out about students‟ language knowledge through more controlled items, such as multiple choice questions. The multiple choice test items is often quicker to design and crucially, easier to mark and produce greater scorer reliability. Harmer (2004) states that multiple choice tests were considered to be ideal test instruments for measuring students‟ knowledge. Because the multiple choice test is easy to mark and since the advent of computers the answer books for these tests can be read by machines not people or manual, thereby cutting out the possibility of scorer error. Grondlund (2003) states that the multiple choice item consists of a stem, which present a problem situation, and several alternatives (option or choices), which provide possible solutions if the problem. The stem may be a question or an incomplete statement. The alternatives include the correct answer and several plausible wrong answers called distracters.

According to Grondlund (2003) there are rules for writing multiple choice items: First, designing each item to measure an important leaning outcome, presenting a single clearly formulated problem in the stem of the item, stating the stem of the item in simple, clear

language, and put as much of the wording as possible in the stem of the item. Moreover, stating the stem of the item in positive form, wherever possible, emphasizing negative wording whenever it is used in the stem of an item, making certain that the intended answer is correct or clearly best, making all alternatives grammatically consistent with the stem of the item and parallel in form, avoiding verbal clues that might enable students to select the correct answer or to eliminate an incorrect alternatives and making the distracters plausible and attractive to the uninformed.

Second, varying the relative length of the correct answer to eliminate length as a clue, avoiding using the alternative "all of the above," and use "none of the above" with extreme caution, varying the position of the correct answer in a random manner and controlling the difficulty of the item either by varying the problem in the stem or by changing the alternatives. Third, making certain each item is independent of the other items in the test, after that using an efficient item format following the normal rules of grammar, and breaking or bend any of these rules if it will improve the effectiveness of the item.

A multiple choice item is designed for objective measurement and contains a stem and response options, one of which is the correct answer (Murayama, 2009). It is a kind of test item in which some options are given and the examinee is expected to pick the correct one out of those options provided. Multiple-choice items consist of a stem and a set of options. The stem is the beginning part of the item that presents the item as a problem to be solved, or a question asked of the examinee, or an incomplete statement to be completed, as well as any other relevant information. The options are the possible answers that the examinee can choose from, with the correct answer called the key and the incorrect answers called distracters. A multiple choice item is expected to be as long as necessary to ensure maximum validity and authenticity to the problem at hand. The stem ends with a lead-in question which describes what the examination taker must do. The stem is expected to be expressed clearly and concisely, avoiding poor grammar, complex syntax, ambiguity and double negatives. Negative statements are not characteristic of normal thought processes, and consequently may place the candidate who is attempting to decipher the item at a disadvantage (Emaikwu, 2012). If a negative question is used, it is expected to be emphasized with italics or underlining. The options are divided into two and they include the key and the distracters. The key is the correct option, while the distracters are the options which appear to the examinees to be the correct answers but are not correct in the actual term.

Typical multiple-choice items have three parts: a stem that presents a problem or task to be answered or solved; the correct or best answer; and several distracters (i.e. the wrong or less appropriate option). The incorrect responses are often called foils, distracters or distractors while the correct response is called the key. In multiple-choice test the examinees are prone to blind guessing, which enable testes to be credited with undeserved scores, where academically poor students score more marks than the knowledge he has in the subject therefore, making it difficult to discriminate between the bright students and poor student. To preserve the advantages of objectives test in general and that of multiple choice test in particular, a number of scoring procedures have been developed such as liberal making, logical-choice weight, eliminating testing, confidence scoring, probability testing, partial order, complete orderly, permutational multiple-choice test.

Confidence scoring, in which a student is asked not only about the correct answer of a question but also about how confident he or she feels about his or her answer, is one of the methods which improve scoring of different types of objective tests (Gardner-Medwin,

2006). It is also claimed to be a method for improving learning (Gardner-Medwin&Gahan, 2003). In other words, confidence-based assessment can be used both in formative and in summative assessment. In the former case, it contributes to improving learning. In the latter case, it is just a scoring method. Multiple choice tests incorporating confidence assessment require candidates to assign a confidence level to each of their selections to reflect their degree of certainty. One such scheme as used on medical students, according to Gardner-Medwin (1995), requires candidates to attach a confidence level of 1, 2 or 3 to their selected answer for each question; this is the mark awarded if their selection is correct, while 0,–2 or –6 are awarded (respectively) otherwise. Confidence scoringprocedure uses both the correctness of the answer as well as thestudent's selection of "confident" or "not confident" to determine the grade for each question. An important skill for students is to recognize how confident they are in their stated answers to questions. Even if students get the right answers, they may not be sure it is right and may get similar questions wrong. Confidence-based scoring aims to address the last aspect – providing the incentives and mechanism for students to assess and state their confidence in their answers (Gardner-Medwin 2006).

Confidence scoring is designed to combine their answer with their confidence so they are encouraged to further examine their answers and more seriously evaluate their confidence level. When the students recognize that they do not know the answer, they get some credit for correctly stating their lack of knowledge or understanding. When the students get the correct answer but are not sure, they do not get the full credit due to the uncertainty. Confidence scoring method encourages students to evaluate their abilities throughout a course and enables them to become more aware of important aspect of decision-making (Gardner-Medwin&Gahan, 2003). One of the earliest academic papers studying the use of confidence-based scoring hypothesized that the reliability of grading would be improved by incorporating confidence into the student's answer since a correct answer on a multiple-choice question always has the relatively high chance of being a guess. Over the years, researchers have continued to study various confidence-based assessment methods. The CBS methods studied tend to evaluate multiple levels of confidence.One of the repeatedly-studied methods for CBS includes three levels of confidence with a correct or wrong answer (Gardner-Medwin, 2013). There are different types of marking schemes in confidence-based assessment. The main sources of difference among marking schemes are the number of certainty levels that students are to choose among and the way different certainty levels are marked for correct and wrong answers. In most marking schemes, there are three certainty levels: high, mid, and low (C=1, C=2 and C=3). In fact, after answering each question, a student should choose one of these three certainty levels. In a marking scheme developed by Gardner-Medwin (1995), which was initially proposed for true/false questions, a correct answer with high, mid, and low certainty levels will receive 3, 2, and 1 point(s) respectively. A wrong answer, on the other hand, with high, mid, and low certainty levels will receive -6, -2, and 0 point(s) respectively.

The reason for choosing such marking is that it motivates students to report their real level of confidence. The negative scores for wrong answers with mid and high certainty levels guarantees that students will not report high levels of confidence when they are not that confident about their answer (Gardner-Medwin&Gahan, 2003). As Gardner-Medwin (2006) mentions "this is the motivating characteristic of the mark scheme, rewarding the student's ability to judge the reliability of an answer, not their self-confidence or diffidence. One might think that the negative points are too high and that it would be better to choose lower negative

points. As Gardner-Medwin (2006) argues, in true/false questions, the probability of answering a question by pure chance and getting the answer right is 50 percent. Therefore, even when we talk about low level of certainty, it is above 50 percent. As a result, the penalties should be great.

In evaluating confidence testing, it is necessary to show that the procedure adds more ability variation to the system than error variation and that any increase in the amount of information gained is, in fact, worth the effort. However, confidence assessment is rarely emphasized during typical assessment in Nigeria secondary schools' education. The confidence-based scoring method encourages students to both think about their answers in a different way and to evaluate their confidence in the answer. Each answer is scored based on whether the answer is right or wrong and whether the student is confident or not in that answer. The use of confidence scoring is not popular among classroom teachers in Nigeria and thus its ability to improve the psychometric quality of three-, four and five option multiple-choice tests have not been well established empirically; hence the need for this study to:

i. ascertain the impact of number of options on distracter' performance when confidence scoring is used.
ii. establish the impact of numbers of options for multiple choice test items on the reliability of the test

Research Hypotheses

i. Number of options have no significant impact on distracter performance when confidence scoring is used
ii. There is no significant difference in the multiple-choice test items reliability base on numbers of option when scored using confidence scoring.

Methodology

The survey research design was adopted for the study. The population of the study consisted of secondary school students in Osun State. The study sample consisted of 360 students selected using multistage sampling techniques. In each of the three senatorial districts of the State, two Local Government Areas (LGAs) were selected randomly and from each LGA, three schools were also selected randomly to make a total of 18 schools. A total of 20 senior secondary school two (SSII) students were selected from each school using purposive sampling technique. The best 20 students in a pre-test in each school were selected. The instrument used for the study was 2015 West Africa School Certificate Examination (WASCE) Economics test items. The Economics Achievement Test (EAT), which was the 4-options format adapted from the 2015 WASCE Economics items had been validated for used by the West African Examinations Council. The validity of the other types (3-options and 5-options formats) were determined and scrutinized by experts in Tests and Measurement and Economics teachers in the secondary school to judge its face and content validity as well as item arrangement. The corrections were incorporated into final version of the instruments. The EAT instrument of 3-options and 5-options were validated in a pilot study conducted using 40 senior secondary school II Economics students who were not part of the final sample size but in a different study area with similar characteristics. Given the responses of the respondents used in the pilot study, the 3-options and 5-options which consisted of 50 items were subjected to a measure of internal consistency using Kurder-Richardson 21 (K21) to ascertain the reliability of the instrument. The result of the K-21 for both 3-options and 5-

options yielded coefficients of 0.79 and 0.83 respectively. This indicated that the 3-options and 5-options EAT still remains internally reliable despite the adaptations made by the researcher in the course of the study. Data collected were analysed using ANOVA, Kuder - Richardson Formula (KR-20) and Fisher's Z-Test with aid of FZT compotator.

Results

Hypothesis One: *Number of option formats have no significant impact on distracter performance when confidence scoring is used*

To test the hypothesis, the distracters performance for each of the 50 items of three, four and five options multiple choice tests were estimated using Microsoft Excel Package. The results item distracter analysis were further analyzed using one way analysis of variance (ANOVA) for the determination of the difference in the distracters performance when three, four and five option multiple choice test items were scored using confidence scoring. The result is as presented in Table 1

Table 1: The difference in the number of options impact on distracter performance

Source	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1965.444	2	982.722	6.679	.002
Within Groups	21629.000	147	147.136		
Total	23594.444	149			

$p < 0.05$ significant

The results in Table1 showed F-ratio value ($F(2,147 = 6.679, p < 0.05)$) implies that multiple choice test items number of options had a significant impact on the performance of the distracters. A post hoc analysis was then carried out using Scheffe multiple comparison to determine between which pairs of the options lies the difference noticed in the distracter performance. The results were as presented in Table 1

Table 2: Multiple comparison of three, four and five-options multiple choice test item

(I) Multiple Choice Test Item Options	(J) Multiple Choice Test Item Options	Mean Difference (I-J)	Std. Error	Sig.
3-Options	4-Options	-5.53333	2.42599	.078
	5-Options	-8.76667*	2.42599	.002
4-Options	3-Options	5.53333	2.42599	.078
	5-Options	-3.23333	2.42599	.414
5-Options	3-Options	8.76667*	2.42599	.002
	4-Options	3.23333	2.42599	.414

*The mean differences were not significant at the 0.05 level.

The results as presented in Table 2 showed that the impact of number of option on distracter performance for a pair of three and four option as well as pair of four and five-options were not significant. However, the significant impact of number of options on distracter performance as noticed in the ANOVA results lies between three and five-options multiple choice test items.

Hypothesis Two: The difference in the multiple-choice test items reliability based on number of options when scored using confidence scoring is not significant

To test this hypothesis, the Kuder and Richardson Formula 20 reliability index, an internal consistency of measurements with dichotomous choices (i.e. correct versus incorrect) with the formula;

$$rhoKR20 = \frac{K}{k-1}\left(1 - \frac{\sum jk=pjqj}{\sigma 2}\right)$$ where

- K = number of questions
- ∑jk=pjqj= sum of all pq's (p= probability of correct answer, q = probability of wrong answer)
- σ2 = variance of the total scores of all the testees

In order to estimate the reliability of each of three, four and five-options multiple choice test items when scored using confidence scoring, the p, q, sum of all pq's and σ2 and finally the reliability estimates were determined using Microsoft Excel Package. The differences in the estimated reliabilities were determined using Fisher's Z-Test with aid of FZT compotator. The result is as presented in Table 3

Table 2: Difference in the reliability of each of three, four and five options multiple choice test items when scored using confidence scoring

	3- Options	**4-Options**	**3-Options**	**5-Options**	**4-Options**	**5-Options**
rhoKR20	0.623	0.691	0.623	0.600	0.691	0.600
Z	0.640		0.196		0.837	
P	p > 0.05		p >0.05		p > 0.05	

The Results as presented in Table2 showed that for each pairwise comparison there were no significant difference in their reliabilities because zobt= 0.640, 0.196 and 0.837 for 3/4-options, 3/5-options and 4/5-options respectively were less than z0.025 = 1.96. (That is for two tailed z-critical to be significant at 0.05 levels of significant it must ≥ 1.96). Thus, the hypothesis was accepted, that difference in the multiple-choice test items reliability based on number of options when scored using confidence scoring was not significant.

Discussion of findings

Prior to the result of the hypotheses one that number of option had no significant impact on distracter performance when confidence scoring was used, the finding indicated that the number of options had significant impact on the performance of the distracters and number of option on distracter performance for a pair of three and four option as well as pair of four and five options were not significant. However, this finding was contrary to the finding of Deepak, Al-Umran, Al-Sheikh, Adkoli and Al-Rubaish (2015) their study revealed that the non-functionality of distracters inversely affected the test reliability and quality of items in a predictable manner. The non-functioning distracters made the items easier and lowered the discrimination index significantly. Three non-functional distracters in 5-option multiple choice questions significantly affected all psychometric properties.

The result of the analysis of hypothesis two of the present study also indicated that the difference in the multiple-choice test items reliability based on number of options when scored using confidence scoring was not significant. This result was in consistent with that of Afolabi (2000) who discovered that in the 3-MC, the internal consistency of items tends to increase as random guessing decreases whereas the highest reliabilities were obtained in the 4-MC and 5-MC. It was also revealed, in agreement with Salehi and Bagheri (2015)who found that confidence-based scores had better predictive validity than conventional scores.

Conclusion/ Recommendations

The study concluded that confidence Scoring Method is more adequate to capture student's cognitive status in multiple-choice tests and also increase the competencies of the multiple-choice item exams to ensure greater fairness of assessment, effective examination, authentic testing and precise estimation and higher construct validity and reliability than the number right.

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

1. The confidence scoring procedure should be encouraged and used in schools as it has been found to be effective in reducing the contribution of random guessing to testees' total score and in rewarding the partial knowledge of testees' on multiple choice tests.
2. Confidence scoring procedure considerably reduces the 'craze' for a do or die affair to pass examination at all cost, hence should be used in all schools.
3. Public Examination Council such as West African Examination Council, the National Examination Council (NECO), Joint Admission and Matriculation Board (JAMB) should apply Confidence Scoring Methods in setting and scoring their multiple-choice tests and Nigeria Teacher Institute (NTI).
4. Classroom teachers should be encouraged to develop skills on how to conduct and score Economics Multiple-choice tests using Confidence Scoring Methods. This can be achieved through organizing seminars for them on Testing Techniques.
5. Four option items especially in multiple choice Economics tests should be encouraged but if five options items should be used more attention should be given to psychometric properties of the tests.

References

Afolabi E.R.I.2000). The reliability and validity of the three confidence scoring methods in multiple choice test unpublished manuscript. Faculty of Education, Obafemi Awolowo University, Ile-Ife.

Ajayi B. K (2012). The Effect of Logical Choice Weight and Corrected Scoring Methods on Multiple Choice Agricultural Science Test Scores. Asian Economic and Social Society ISSN (P): 2304-1455, ISSN (E): 2224-4433 Volume 2 No. 4 December 2012.

Awodele B. A. (2013) Comparative Effectiveness of logical-choice weight and confidence scoring methods on reliability and validity of multiple-choice test items in Nigerian Secondary Schools. Journal of Educational and Social Research Vol.3 pp. 387-39.

Bracey, G.W., (2001), Thinking about tests: A short primer in assessment literacy, TheAmerican Youth Forum, Washington D.C.

Chan, N., & Kennedy, P. E. (2002). Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple- Choice and "Equivalent" Constructed-Response Exam Questions. Southern Economic Journal, 68(4), 957-971.

Deepak ..K. Al-Umran. U. Mona H. Al-Sheikh B. Adkoli .V. &Abdullah A. (2015). Psychometrics of Multiple Choice Questions with Non-Functioning Distracters: Implications to Medical Education. Indian J Physiol Pharmacol, 59(4), 829-835

Emaikwu, S O. (2012). Fundamentals of test, measurement and evaluation with psychometric theories. Makurdi: Selfer Academic Press.

Gardner-Medwin, A. R. (2006). Confidence-based marking: Towards deeper learning and better exams. In C. Bryan & K. Clegg (Eds.). Innovative assessment in higher education (pp. 141-149). London: Routledge.

Gardner-Medwin, A. R. &Gahan, M. (2003). Formative and summative confidence-based assessment. Proceedings of the 7th International Computer-Aided Assessment Conference, Loughborough, pp. 147-155.

Gardner-Medwin, A.R. (1995). Confidence assessment in the teaching of basic science. Journal of Association for Learning Technology, 3, 80-85.

Gronlund, N. E. (2003). Assessment of Student Achievement. Seventh Edition. Boston: Pearson Education Inc

Harmer, J. (2004). How to Teach English, New Edition. Edinburgh Gate: Pearson Education Ltd.

Kolawole, E.B.(2007). A Comparative Analysis of Psychometric Properties of Two Nigerian Examining Bodies for Senior Secondary Schools Mathematics. Research Journal of Applied Sciences,2(8): 913-915.

Kolawole, E. B. (2005). Test and Measurement. Lagos: Bolabay Publications.

Murayama, K (2009). Improving your test questions. Center for Innovation in Teaching and Learning, University of Illinois.

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

Omirin M.S. (1999). Construction and validation of a science – Oriented Attitudinal Scale for Nigerian Schools. An unpublished PhD thesis. Ekiti State University, Ado Ekiti

Salahi, M, Bagheri, M. S. (2015). Comparing confidence-based and conventional scoring methods. The case of an English grammar class. Journal of Teaching Language Skills, 33(4), 123-152