The Number of Options in a Multiple-Choice Test Item and the Psychometric Characteristics.

Peter Ikechuckwu Nwadinigwe[1] & Louisa Naibi[1]
[1]Department of Educational Foundations, Faculty of Education, University of Lagos, Nigeria.
* E-mail of the corresponding author: l.naibi@yahoo.com

Abstract
Traditionally, multiple-choice test items(MCIs) have been written with four or five response options, and measurement textbooks have recommended this. However, in the recent past, many studies have theoretically and empirically found that three options are just as effective, and may be the optimal number of options for MCIs. This study investigated and compared the effect of two number of options (noOPT) formats, five options versus three options, on test and item psychometric characteristics. A Mathematics Achievement Test (MAT) was administered twice, first with five options, and then with three options per item. The study used a sample of one hundred and fifty-nine (159)Primary Six pupils in state government-owned schools in Bayelsa State, South-South Nigeria. The study revealed that noOPT significantly affected mean test scores, mean test difficulty and test discrimination indices, but not internal consistency reliability coefficient. Generally the findings provided more evidence to support the use of three option.

Keywords: number of options (noOPT), psychometric characteristics, item difficulty index, item discrimination index, test reliability

1. Introduction

The multiple-choice test (MCT) is perhaps the most extensively-used format in the assessment of student knowledge today. Globally, it is used, in part or in whole, in classroom tests, as well as in standard admission, placement and certification tests. Individuals encounter it throughout their academic (from primary, secondary, to tertiary levels of education) and professional career.

For test results to be accurately interpreted and applied, the psychometric features of validity and reliability are necessary and sufficient requirements. Validity is the extent or degree to which the test can measure the qualities, skills, abilities, traits, or information that it was designed to measure; while reliability is the extent to which it can consistently make these measurements. MCTs are a commonly preferred format for large-scale testing because they have been found to give valid and reliable results. They can be effectively used to assess large number of test-takers, to cover a wide range of content area (Downing, 2002, cited in Tarrant, Ware & Mohammed, 2009), as well as a wide range of learning objectives (Okoli, 2005), leading to valid results. They can be objectively, easily and accurately scored (Linn & Gronlund, 2000), and they contain a relatively large number of test items, leading to a high degree of reliability.

Structurally, MC questions or items, have two parts: a stem - the question, problem, or task to be answered or solved; and a set of response options or alternatives - possible answers or solutions to the question. The options comprise of the correct answer, called the key; and one or more incorrect or less appropriate answers called the distracters or distractors (Onunkwo, 2002). While the stem is an important part of a multiple-choice item (MCI), the options are no less relevant. No matter how well-written a stem, a single flawed option can invalidate the item. As a rule, well-written options are critical for an MCI to be adjudged as valid**.**

2. Test item discrimination: item difficulty and distracter plausibility

To reliably [and validly] use MCTs to rank students on the basis of achievement, the items must have the ability to discriminate or detect small differences in achievement (Michaels & Karnes, 1950). Item discrimination is the ability of the item to differentiate between low ability test-takers (those who have not mastered the material taught or achieved the learning outcome being tested) and high ability test-takers (those who have achieved a command of the concept or principle involved in the item). The discriminative power of a test item depends on its difficulty level and largely on the degree of plausibility of its distracter (Onunkwo, 2002). Item difficulty reflects how difficult or easy an item is, and is numerically expressed as the proportion of students who answer an item correctly. An effectively discriminating item must have suitable levels of difficulty, and to ensure this, each of their distracters must possess the feature of plausibility. Plausibility means that the distracters must be undeniably wrong, yet be compelling, rational, and logical enough to appear as correct, to those who do not possess the particular knowledge being tested. Further, the distracter must be seen as incorrect and consequently be rejected by those who do possess the knowledge. Hence, distracters should attract a high proportion of low ability candidates and a low proportion of high ability students. Otherwise, they are technically flawed and non-functional.

By effectively discriminating between the different ability groups, a well-written distracter ensures accurate assessment, and gives a credible and objective view of knowledge state. It is the distracter plausibility that ensures appropriate difficulty levels, which in turn ensures that the item poses the appropriate amount of challenge to the student. Item discrimination is important because it is an estimate of item validity (Obe,1980); and thus for MCTs, it serves a linchpin function, and the other two features plays a key role in carrying out this function.

Haladyna (2004) and Haladyna and Downing (1993) have averred that one of the most problematic and difficult areas in developing MCIs is writing, not the key option, but plausible and functional distracters. Reviewing distracter functionality in 477 MCIs in four examinations, they found that only between 1.1% and 8.4% of the items had up to three functioning distractors. Similarly investigating teacher-developed tests, Tarrant at al. (2009) assessed a sample of seven tests, with a total of 2056 options, and found that only about 13.8% of the items had up to three functioning distracters**.** One hypothesized cause of implausible and therefore non-functional distractors is the number of options (noOPT), particularly the distractors, included in the item.

3. Number of options in an MCI

According to standard measurement theory, the more response options in an MCI, the more reliable it is (Thorndike & Thorndike-Christ, 2010; Hopkins, 1998; Mehrens & Lehman, 1991) and so the prevailing guideline is to develop as many response options as is effectively or feasibly possible (Haladyna, Downing & Rodriguez, 2002). As a result of this, five, or at least four options, have been traditionally endorsed. Consequently, the noOPT in a typical MCI usually vary between three (i.e. two distractors and a key) and five (four distractors and a key).

There are very significant arguments in favour of five options, and thus against 3 options. Woodford and Bancroft (2004), Abad, Olea and Ponsoda (2001), and Farhady and Shakery (2000) have asserted the following:

- that lower numbers of options, such as three, increase, to an unacceptably high degree, the chances of successful random guessing and the extent of guessing effects, such as over-estimation of student achievement or ability (with five options, the degree of chance success is 20%; with four options, it is 25%; and with three options, it is 33.3%);
- that this decreases the psychometric quality the test scores, making it less reliable and consequently less valid;
- and that this psychometric limitation can only be corrected by using five, or at least four options per item

However, backed by contemporary research evidence, counter-arguments have been put forward. Haladyna and Downing (1993) have stated that in a majority of cases, it is often difficult and time-consuming to write up to four or even three plausible functional distractors for an MCI, so that not more than two of the distracters are usually functional, and additional options after the third one are often always implausible**.** They add that the additional distracters are merely fill-ins which are not plausible enough to distract weak students, who immediately see through them and easily guess the correct answer. They advocate three options (i.e. two distracters and the correct option) as a "natural limit"(p. 1008) under most circumstances.

A second argument in favour of three options, over five and four, also firmly based on the research result, is that they give similar, and sometimes superior outcome on many testing criteria, and are thus preferable for some testing purposes. According to Costin (1970) they are less arduous and time-consuming to develop and administer, and have reduced completion time. The time and resource thus saved could be used to develop more items for the test, boosting reliability. Consequently, this time, cost, and energy saved, and used for other relevant activities, increase efficiency of assessment without compromising test quality and hence, can be equated with increase in content validity and test reliability. Proponents of five and/or four options have countered these claims with the assertion that if the extra options are functional, [if they are well-written and plausible], the overall benefit of reducing guessing outweighs any extra time that may be gained by using 3 options, and constructing more items to boost reliability (Woodford & Bancroft, 2004); and that since the use of more options reduces guessing effects, which increases reliability, it also increases validity (Thorndike & Thorndike-Christ, 2010; Mehrens & Lehman, 1991). Hopkins,19 (8) and Farr, Pritchard and Smitten (1990) have described these effects as 'overrated' (p.148) and 'negligible' (p. 224) respectively, arguing that most serious students who have adequate time to write a test, use partial knowledge and educated guessing, rather than the

psychometrically-deleterious random guessing. Haladyna et al. (2002) believe that "three options are sufficient in most instances and that the effort of developing the fourth option… is probably not worth it" (p. 318).
It is noteworthy that despite the claims by traditionalists, there is a dearth of contemporary research evidence to support the superiority and continued use of five or four options. Yet, most achievement and ability testing programs and examinations still use five or four options, while test and measurement textbooks typically recommend this, based on the belief that the greater the noOPT, the higher the reliability. In their landmark review research study, Haladyna et al. (2002) reviewed twenty-seven measurement textbooks and twenty-seven research studies and reviews. They found significant research support for three options, but revealed that most of the assessment textbooks surveyed were still divided on the issue, with about 70% advocating the prevailing guideline, and a few others advocating a 'middle-of-the-road' 4 option, as an industry standard.

More than a few studies have substantiated the efficacy of three options and their suitability for educational tests. Theoretically, using mathematical formulae-driven perspectives, it has been contended that three options optimize the discrimination ability of a test and the information that it could provide (Tversky, 1964), especially with moderate item difficulty (Lord, 1977); that reliability significantly increases when the noOPT is increased from two to three, but not by much when increased to four, and that any increase beyond three would only give marginal increase - in the range of 0.02 to 0.05 (Ebel,1969); that three-option would give higher reliability and discrimination than the others (Grier, 1975). From the empirical perspective, studies have found three-option items to have higher mean scores and to be less difficult and more discriminating (Costin, 1970; Landrum et al., 1993;Trevisan et al., 1994). However some did not find any significant difference in difficulty (Owen & Froman, 1987), nor in both difficulty and discrimination (Crehan, Haladyna, & Brewer, 1993; Sidick et al., 1994; Shizuka, Takeuchi, Yashima and Yoshizawa, 2006), though Rogers and Harley (1999) found increase in item difficulty. In terms of reliability, Stratton and Catts (1980) found similar reliability (and higher standard error of measurement) for three options; while Landrum et al., (1993), Trevisan et al., (1994), Sidick et al. (1994), Cizek, Robinson and O'Day (1998), Rogers and Harley (1999) and Rogausch, Hofer and Krebs (2010) all found non-significant or little increase in reliability.

4. Statement of the problem

Currently, the optimal number of options to use in a multiple-choice test item is a debatable issue, trailed by contradictions and controversies, and stirred by lack of a firm conclusion from empirical and theoretical findings. The point of contention is between support for five, or at least four options (endorsed mainly because it is said to minimize the guessing effects which lower noOPTs generate) and three options (with growing research-backed support). While five-option seems to be an industry standard, an increasing number of studies have postulated that three-option has similar and/or superior psychometric features, mitigates the existence of non-functional distracters, and is more efficient in terms of time to develop, administer, and complete. It is thus appropriate to more seriously consider this issue in light of the relevance, usefulness, pervasive and potential application of MCIs in almost all disciplines and professions. There are always ongoing attempts to refine and improve them, not during item analysis, but during item-writing, where the problem of writing non-functional distractors can be best eliminated. While statistical item analysis would eventually detect flawed distracters, this is often not feasible in classroom assessment. It has therefore become important to consider the optimal noOPT per multiple choice item; and the point at which the noOPTs can be fixed, or at least reduced, without impeding the psychometric quality of tests.

5. Purpose of the study

With an underlying purpose of testing the validity, and generalizability of previous empirical findings to local contexts (primary school pupils in Nigeria), the study investigated the differential effects of five and three number of options (noOPTs) on mean test score, Kuder-Richardson (K-R) 20 reliability coefficient, mean item difficulty (*p*-value) and discrimination indices (*d*-value). It involved the testing of four null hypotheses about these characteristics, in an achievement test when it had five and three options, respectively.

6. Methods and Materials

The study utilized intact classes (quasi-experimental), using a repeated measures two-group design, where all participants were exposed to both option formats. The ABBA counterbalancing design was built-into the main design to control for the inevitable practice and order effects. The study was carried out in Yenagoa City, in Bayelsa State, South-South Nigeria. From a target population of Primary Six pupils in government-owned primary schools in the city, all in the second term of that academic session, and preparing to move to secondary school, a simple random sample of three schools in Yenagoa City, and of classes in these schools, (select a school, select a class) was used. The study sample comprised of one hundred and fifty-nine (159) pupils, M = 78, F = 81. The classes were assumed to be normally distributed.

The main instrument was a Mathematics Achievement Test (MAT) developed and validated by the researcher according to a specified test blueprint, consisting of 40 items, covering all relevant topics in the final year of primary level, and selected from an existing test bank of past questions of similar examinations. The process of option removal (removal of non-functional or least functioning distractors based on item analysis data), was used to modify the five-option format (MAT Test Form A) to the three-option version (MAT Test Form B). Where two distractors had equal *p*-values or *d*-values, randomization was used to discard one.

Research assistants were mostly teachers from the selected schools, with a minimum qualification of National Certificate in Education (NCE), and were trained in practical aspects of the study, in two one-hourly sessions, a week and a day respectively, before the first administration. A neutral school (one not selected) was used as the testing center for the first administration, while the respective schools, were used for the second. Version A was disguised as a mock test to assess pupils preparedness for the First School Leaving Certificate Examination. Version B was administered two weeks later, and served as the mid-term test for that term. For the counterbalancing (AB and BA), the first administration involved Form A being administered to half the class at random, while Form B was given to the remaining half. This sequence was reversed during the second testing session: Form B was given to those who had taken Form A before, while Form A was given to those who previously done Form B. It was ensured as much as possible that no pupil got the same version as the pupil they were sitting next to. A time limit of 60 minutes ($1^{1}/_{2}$ min per item) was given. Pupils were encouraged/ asked to answer all items, even those they were not sure of. The setting was thus that of a standardized external examination, though with familiar teachers present. The 27% mark was used as a cut-off point for the ability groups (Ebel & Frisbie, 1991).

7. Results and Discussion

The hypotheses were tested with the correlated *t-test* at 0.05 alpha level.

Hypothesis One detected a significant difference in the mean test scores: Test Form A had a lower mean score than Test Form B, invalidating the stated null hypothesis. Results are shown on Table 1 below.

Table 1: Summary table showing t-test analysis of test scores obtained with Test Forms A and B

Test Forms	N	$\bar{X}$ p-values	SD	df	t-calc	t-crit	Results
A (5 options)	159	21.748	15.522	157	±7.28	± 1.96	P<.05 (Sig)
B (3 options)		23.465	13.886				

Hypothesis Two did not find any significant difference in the obtained K-R 20 *r* coefficients, so the hypothesis was not rejected. However, the SEM increased between the test formats. Results are shown in Table 2.

Table 2: Summary table showing t-test analysis of K-R 20 *r* values of Test Forms A and B

Test Forms	N	$\bar{X}$	S.D.	*r* values	SEM	Df	t-calc	t-crit	Results
A (5options)	40	21.748	15.522	0.882	2.08	37	± 1.07	± 2.04	P >.05 (NSig)
B (3 options)		23.465	13.886	0.872	2.32				

df = n - 3.

Hypothesis Three failed to support the null hypothesis, when a significant difference between the mean item difficulty indices was revealed, with Test Form A having lower values. Results are shown in Table 3a below.

Table 3a: Summary table showing t-test analysis of *p*-values obtained Test Forms A and B

Test Forms	N	$\bar{X}$ p-values	SD	df	t-calc	t-crit	Results
A (5 options)	40	0.47175	8.860	38	±6.359	± 2.024	P<.05 (Sig)
B (3 options)		0.54925	10.491				

Table 3b shows a basic classification and interpretation of a range of difficulty indices. It shows that based on the pupils' responses, a larger percentage of items in both tests were of moderate difficulty (classified as 'Easy' & 'Difficult'). Test A had 35 such items, while Test B had 38). While 2 items were perceived as Very Difficult in Test A, there were none in Test B.

Table 3b: Classification of the range and interpretation of *p*-value of Test Forms A and B

Classification range		Test form A		Test form B	
Very Easy	0.70 - 1.00	1	2 . 5%	2	5 . 0%
Easy	0.50 - 0.69	13	32. 5%	25	62. 5%
Difficult	0.30 - 0.49	24	60 . 0%	13	32 . 5%
Very Difficult	0.00 - 0.29	2	5. 0%	Nil	0. 0%
	TOTAL	40	100%	40	100%

Hypothesis Four discovered a significant difference, indicating a lack of support for the stated null hypothesis. Test Form A had a lower mean discrimination value than Test Form B. Results are shown in Table 4a below.

Table 4a: Summary table showing t-test analysis of *d*-values obtained with Test Forms A and B

Test Forms	N	Mean *d*-values	SD	Df	t-calc	t-crit	Results
A (5 options)	40	0.4587	20.39	38	±3.221	± 1.96	P<.05 (Sig)
B (3 options)		0.5305	20.52				

Table 4b shows a classification and interpretation of the range of discrimination coefficient, based on Furst (1958) taxonomy. Based on the pupils' responses, the test items discriminated well in both test formats, being mostly in the category of 'moderate' to 'high positive' discrimination (Tests A and B both had 32 of such items). The remaining 8 items still discriminated positively, though to a lesser degree.

Table 4b: Classification of range and interpretation of *d*-values obtained from Test Forms A and B [based on Furst (1958) classification]

Classification (+ve = positive discrimination)		Test form A		Test form B	
High +ve discrimination	0.50 and above	18	45. 0%	25	62 . 5%
Moderate +ve discrimination	0.30 - 0.49	14	35. 0%	7	17. 5%
Borderline +ve discrimination	0.20 - 0.29	3	7. 5%	4	10 . 0%
Low to Zero +ve discrimination	0.00 - 0.19	5	12. 5%	4	10. 0%
Negative discrimination	-0.10 and below	Nil		Nil	
	Total no of items	40	100.0%	40	100.0%

The study found significant differences between the two test forms in mean score, item difficulty and discrimination (Hypotheses One, Three and Four). Essentially, the participants had higher mean scores in the three-option version, they found it less difficult and it discriminated better, than the five option format. Since these characteristics are interrelated (the item indices stem from the scores), it is not surprising that a significance in one would very likely lead to the same in another. While this might have occurred due to memory spill-over from the first testing, most of the pupils did not know that Test B was the same test. Participants who noticed item similarity felt that it was merely a coincidence, because at the time of the study, all of them were in the final year in primary school, and in preparatory classes for their First School

Leaving Certificate examinations and various other selection examinations into secondary schools, where they often came across similarly-phrased items. Differences might also have been partly due to the increased chance for guessing, afforded with fewer options in format B, especially for the low ability pupils, though Haladyna and Downing (1993) have opposed this explanation, arguing that in many cases, the additional option is what appeals to low ability students and encourage them to do random guesswork. The higher mean score could also be due to the fact that with fewer options, they had less distracting additional options and so focused on the test items. The results from testing the first and third hypotheses line up with Landrum et al. (1993) and Trevisan at al. (1994) who both found improved scores (and item difficulty) with three options. On the other hand Sidick et al. (1994) found no significant difference in mean scores with both three and five optioned tests; Owen and Froman (1987), Crehan et al. (1993) and Shizuka et al. (2006) all found no significant difference in the mean item difficulty level. The results supported Lord's (1997) theoretical-based prediction that three options would be optimal if the difficulty level was moderate, (which it was for both these tests). The results can also be held up by Ebel's (1972) assertion that *p*-values of around 0.50 tend to give maximum discrimination power. The difference in discrimination was categorically and statistically significant. Furst's (1958) guide for interpreting discrimination indices classified the range of values '0.30 to 0.49' (in which the mean discrimination index of 0.46 for Test Form A falls) as "moderate positive" and '0.50 and above' (in which the mean discrimination index of 0.53 for Test Form B falls) as 'high positive discrimination'. Thus the mean discrimination indices for both test forms were in different categories. From a theoretical perspective, it has been argued that three-option items are more discriminating than all others (Grier, 1975), and tend to optimize discrimination ability and the information that a test could provide (Tversky, 1964). Costin (1970) and Landrum et al. (1993) had similar findings to this study: they found that lower number of options led to increased discrimination. However Crehan et al. (1993) and Shizuka et al (2006) found no change in item discrimination in similar circumstances. Cizek et al. (1998) found contradictory results: with lower noOPTs having significant reduced discrimination in some items and significant increase in others. Generally, results to do with discrimination have tended to be inconsistent.

The only characteristic that did not significantly differ was reliability (Hypothesis Two). The obtained K-R 20 coefficients had a very minimal difference of 0.01, somewhat agreeing with Ebel's (1969) theoretical prediction of a marginal, in the range of 0.02 and0.05. The standard error of measurement (SEM) is the standard deviation of the score across a series of administrations and it is used to interpret the precision of test scores in relation to true score (Onunkwo, 2002). The increased SEM between the formats (2.08 to 2.32) implied that Test B scores may be slightly less accurate and less indicative of the true score than Test A. Reasons for this may be sampling error or score variance. Ebel (1972) asserted that score variability influences the reliability coefficients. The observed change in their standard deviations (Test A=15.52; Test B=13.87) also implied a slight effect on the score variability. Stratton and Catts (1980) found similar differences in reliability and SEM, while Trevisan et al. (1994), Landrum et al.(1993), Rogers and Harley (1999), and Rogausch at al., (2010) all found no significant difference in the reliability coefficients . Sidick et al.(1994), using coefficient alpha, found non-significant differences. This may suggest that reliability has little practical significance. On the other hand, Grier (1975) predicted that three options would lead to a higher reliability, while Cizek, at al. (1998) found reliability increase with reduced number of options, especially when non-functional options were removed. Rodriguez (2005) however, found both reduced reliability coefficient (with reducing five to four options, five to two and four to two options) as well as increased reliability coefficients (with reducing from four to three options).

5. Conclusion, Implications and Recommendations

Statistical analysis indicated that the number of response options used in an MCI impacted mean test scores, mean item difficulty and discrimination, but not the reliability coefficient (K-R 20). Generally the findings provided more evidence to support the use of three-option items. Specifically it demonstrated that additional options over three do not make much difference, and that reducing the test to three options actually improved some of its psychometric features. The findings have significant implications for test construction, not only for classroom formative and summative assessment, but for large scale testing. Thus it can be argued that achieving an optimal balance between number of options and test efficiency can enhance the effectiveness of MCTs in serving their
purposes.

The results have shown that the effect is found even with primary school pupils (in Nigeria). The rule of five options may not be as inviolate as believed, implying that the choice a particular noOPT should be based on practicality. Many teachers and testers who have found that it is not easy to develop five options can now use the practical three options without fear of a psychometric violation. Three options may also be more practical in primary schools where the pupils have a limited vocabulary. It is thus recommended that teachers should be encouraged to use three options in both formative and summative classroom assessment.

Test experts should use and recommend the use of the different noOPTs based on awareness and understanding of the issues surrounding their use.

The information obtained should be disseminated through regularly organized seminars, workshops and in-service training program for staff of schools and examination agencies.

It is also recommended that a similar study be carried out with using core primary school subjects like an English (which, unlike Mathematics, requires longer reading content and no calculations,)and Social Studies and Sciences.

References

Abad, F. J., Olea, J., & Ponsoda, V. (2001). Analysis of the optimum number of alternatives from the Item Response Theory. *Psicothema*, 13*(001),* 152-158.

Cizek, G. J., Robinson, K. L., & O'Day, D. M. (1998). Non-functioning options: A closer look. *Educational and Psychological Measurement,* 58*(4),* 605-611.

Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence of a mathematical proof. *Educational and Psychological Measurement,* 30, 353-358

Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, 53, 241-247.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. *Educational and Psychological Measurement*, 29, 565-570

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement.*(2nd Ed). Englewood Cliff, NJ: Prentice-Hall.

Farhady, J. & Shakery, S. (2000). Number of options and economy of multiple-choice tests. *Roshd Foreign Language Teaching Journal*, 15*(57),*132-141.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*, 27*(3),* 209-226.

Furst, E. J. (1958). *Constructing evaluation instruments.* NY: Longman Green and Co.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement,* 12*(2),* 109-113

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed). Mawah, NJ: Lawrence Erhbaum.

Haladyna, T. M. & Downing, S. M. (1993). How many options is enough for a multiple-choice item? *Educational and Psychological Measurement*, 53, 999-1010.

Haladyna, T. M., Downing, S. M., & Rodriguez M. C. (2002). A review of multiple- choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15*(3),* 309-334.

Hopkins, K. D. (1998). *Educational and Psychological Measurement and Evaluation.* (8th Edition). Needham Heights, MA: Allyn and Bacon.

Landrum, R. E., Cashin, J. R., & Thies, K. S. (1993). More evidence in favour of three option multiple-choice tests. *Educational and Psychological Measurement*, 53, 771-778

Linn, R. L. & Gronlund, N. (2000). *Measurement and assessment in teaching* (8th ed). Columbus, OH: Merrill.

Lord, F. M. (1977). Optimum number of choices per item – a comparison of four approaches. *Journal of Educational Measurement*, 14, 33-38.

Mehrens, W. A. & Lehman, I. J. (1991). *Measurement and Evaluation in Education and Psychology* (4th Ed). Forthworth, TX: Harcourt Brace Jovanovich

Michaels, W. , & Karnes, M. R. (1950). *Measuring educational achievement.* NY: McGraw Hill.

Obe, E. O. (1980). *Educational testing in West Africa.* Lagos: Premier Press

Okoli, C. E. (2005). *Introduction to educational and psychological measurement*. Lagos: Behenu.

Onunkwo, G. I. N. (2002). *Fundamentals of educational measurement and evaluation.* Owerri: Cape Publishers.

Owen, S. V., & Froman, R. D. (1987). What's wrong with three-option multiple choice items? *Educational and Psychological Measurement*, 47, 513-522

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educational Measurement: issues & practice*, 24*(2),* 3-13.

Rogauch, A., Hofer, R., & Krebs, R. (2010). Rarely selected distractors in high stakes medical examinations and their recognition by item authors: a stimulation and survey. *BMC Medical Journal*, 10, 85-91.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, 59, 234-247.

Seinhorst, G. (2008). *Are three options better than four?* Unpublished Masters thesis submitted to Lancaster University.

Shizuka, T., Takeuchi, O., Yashima, T, & Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing,* 23*(1)*, 35-57

Sidick, J. T, Barret, G. V., & Doverspike, D. (1994). Three alternative multiple choice tests: An attractive option. *Personnel Psychology*, 47, 829-835.

Stratton, R. G. & Catts, R. M. (1980). A comparison of two, three and four choice item tests given a fixed number of choices. *Educational and Psychological Measurement*, 40, 357-365.

Tarrant, M., Ware, J., & Mohammed, A. H. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions a descriptive analysis. *BMC Medical Education*, 9 *(40)*. 1-8.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology,* 1,*(2),* 386-391

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and Evaluation in Psychology and Education* (8th Ed). Upper Saddle River, NJ: Pearson/Merril Prentice Hall.

Trevisan, M. S., Sax, G., & Michael, W. B. (1994). Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement*, 54*(1),* 86-91.

Woodford, K., & Bancroft, P. (2004). Using multiple choice questions effectively in Information Technology education. In R. Atkinson, C. McBeath, D. Jonas Dwyer & R. Philips (Eds), *Beyond the Comfort Zone: Proceedings at the 21st ASCILITE conferences* Perth, December 5-8, 948-955