Task Number and Cognitive Complexity as Determinants of Difficulty Levels In Multiple – Choice Test Items.

OGOMAKA, P. M. C[1], Nosike, M.C[2] and Akukwe, A.C.[3]

[1] Director ICEP, Imo State University, Owerri pmcog@hotmail.com
[2] Dept. of Ed. Psychology (Research & Evaluation), Imo State University, Owerri. nikkynosike@yahoo.com +2348033214387
[3]Dept. of Computer Education, AIFCE, Owerri, akuchidiss@yahoo.com +2348035476676

Abstract

In this study, the researchers sought to ascertain the simple and multiple linear relationships between any two and among three multiple-choice Mathematics test- item characteristics: facility indices (Z), cognitive levels (Y) and task numbers (X). The 70-itemed tests were administered on a random sample of 200 students drawn from 1200 SSSII students in four schools. The reliability coefficient of the test was found using the Kuder-Richardson formula 20 is 0.76. The scores of the sample of students to the items and the three sets of characteristics were found. The estimates of correlation coefficients obtained are $r_{ZY} = -0.60$, $r_{ZX} = -0.49$, $r_{XY} = 0.71$ *and* $R_{Z.XY} = 0.89$. *The three correlation coefficients is each significant* ($p < 0.05$). *Test constructors and users should include item task number and cognitive level in item analysis. These are testees or group of testees-independent and could be a bridge between CTT and IRT.*

Keywords: *CTT, IRT, MCT, task number, cognitive level, test item analysis.*

1.0 Introduction

Making fair and systematic evaluation of others performance can be a challenging task. Judgement cannot be made solely on the basis of intuition, haphazard guessing or custom [9]. Tests are tools that are frequently used to facilitate the evaluation process. A test can have varying definitions, but for the purpose of this study, a test is viewed as a device or mechanism for eliciting responses, such that from the responses or reactions, the quality or quantity of the features, attributes, or characteristics in question, which an object possesses or will be able to exhibit could be

ascertained [8].This implies that, these responses may be elicited without actual questions being asked. In testing for honesty, for instance, it could be an arranged situation which a person is exposed to.

Some tests are non-written or verbal, while some others are written. Indeed, the variety of items that make up psychological tests is immense, so it defies easy categorization. Test items vary in many ways, in terms of content, format, mode of administration, scoring rubrics and in the kind of response they call for in testees. All test items can be classified in two broad categories, namely,

(a) Selected – response items such as true – false, matching, ranking, multiple-choice test items; and

(b) Constructed – response test items such as essay, fill-in-the-gap, and short answers.

When a teacher announces that there will be a test, one of the first questions asked is 'what type of test? Will it be true-false test, a multiple-choice test, or fill-in-the-blanks test? This will be classifying test with regards to its format.

Test items are the units that constitute a test and are the means through which samples of testees behaviours or responses are gathered [10]. Multiple-choice Test Items (MCTs) are pliable to various levels of learning outcomes, from simple to complex levels, amenable to item analysis and are sample-dependent. The theoretical framework that supports item analysis in MCT items is based on the Classical Test theory (CTT), which is aimed at explaining the total final result, that is, the sum of responses provided to a series of items, expressed by the total score (S).

The focus of the analysis is on the total test score; frequency of correct responses (to indicate question difficulty); frequency of responses (to examine distracters); reliability of the test and item-total correlation (to evaluate discrimination at the item level) [4]. CTT utilizes measures of item difficulty and item discrimination, the values of which are dependent upon the distribution of examinee proficiency within a sample-that is-sample dependent. Item analysis practices were reported in details and gave the indices as item difficulty, item discrimination, and item distractor – the psychometric properties of MCT items [2], [3], and [9].

In a study using CTT method to analyse the psychometric properties of MCT items in JSCE Mathematics examinations for two consecutive years. The analysis was done to examine various categories of MCT items, the levels of difficulty, the content validity and the positions of the distracters. The findings revealed that, a relatively low content validity, low internal consistency and about 50% of the items fell within the good item category [5].

In a study comparing the psychometric properties of two Nigerian examining bodies' MCT items for SSCE mathematics, the findings revealed that, there were no significant differences between their difficulty level indices, discriminating powers, distractor and validity indices, they were adjudged as being equivalent [6].

In a study aimed at modelling the achievement of medical students that there is a negative correlation between students' achievement on a 150-item MCT and the level of postgraduate qualification achieved by the students [1].

The levels of MCT items are now being advanced, not just in the cognitive complexity-low, moderate, and high- or in terms of the item difficulty but with regard to the task numbers. For the purpose of this study, item analysis will go beyond; the item difficulty, item discrimination, and item distractor of a MCT item to considering the cognitive complexity and task number, as they associate with the difficulty levels of MCT items.

The facilities index (F.I) of an item ranges from 0 to 1, indicates the extent of ease with which the item is gotten correct by a given set of testees [7].

$$F.I = \frac{n_u + n_l}{2n} \quad (1)$$

Where,
n_u = number of testees from the upper one third who got the item correct,
n_l = number of testees from the lower one third who got the item correct, and
n = the number that constitute one third (or 30%) of the testees.

Cognitive complexity of the test items are usually obtained from evaluation experts' classifications or ratings of such items. It refers to the cognitive demand associated with an item. The rationale for classifying items by their level of complexity is to focus on the expectation of the item, not the ability of the student. The demands on thinking that an item makes-what the item requires the student to recall, understand, analyse or do-are made with the assumption that the student is familiar with the basic concepts of the operation. The categories-low complexity, moderate complexity, and high complexity form an ordered description of the demands an item may make on a student. For example, low complexity (knowledge-based items) may require a student to solve a one-step problem. Moderate complexity items may require multiple steps. While a high complexity items may require a student to analyse and synthesize information.

Task number of an item is obtained by counting the number of distinct mental / cognitive operations / processes / steps involved in an answer / working out / solving the item correctly through the common / usual approach (not through shortcuts or approaches more advanced than the group of testees [8]. He posits that, an item task number is the number of distinct operations or steps or skills, which a testee works through to successfully solve an exercise or to successfully accomplish a task, to arrive at a solution. In a MCT item, each testee may decide to approach an exercise (task) from more than one route, working through steps of cognitive processes, depending on their level of preparedness or their level of understanding of the concept(s) being evaluated. When there is a repeated operation in tackling an item, it is counted once. For instance, repeated addition is counted once. Task numbers take on numerical values and increase monotonically, depending on the cognitive complexity of the test item. Test items that require more steps in their solutions are more difficult than those that require fewer steps, given that the steps are similar. Though, in scoring the MCT items, all correct items are scored one, irrespective of the task number and incorrect items scored zero.

The concept will be illustrated with examples below

(1) Multiply 132 by 3
(2) Multiply 132 by 5
(3) Find x if 3x – 5 = x + 7
(4) What does the symbol ∈ stand for?
(5) Calculate the area of an equilateral triangle of side 8cm are got by answering the test items showing all details in each case.

(1) 132x3 = 396
(i) knowledge of multiplication and
(ii) knowledge of multiplication table

Task number = 2

(2) 132x5 = 660
(i) know what x stands for,
(ii) knows the multiplication table,
(iii) know how to carry over and
(iv) add.

Task number = 4

(3) 3x – 5 = x + 7

-x -x (1)
2x – 5 = 7
+5 + 5 (2)
2x = 12
/2 /2 (3)
x = 6

Task number = 3

(4) What does the symbol ∈ stand for?
(a) is an empty set
(b) is the universal set
(c) is a member of
(d) is an infinite set. Task number
=1→ **recall of a symbol**

(6) Calculate the area of an equilateral triangle of side 8cm.

(a) $4\sqrt{3}$ cm^2
(b) $8\sqrt{3}$ cm^2
(c) 16cm^2 (d) 8cm^2 (e) $16\sqrt{3}cm^2$

Steps:

(i) recalling formula for area of a triangle
(ii) recall of Pythagoras' theorem to find h
(iii) correct substitution of values
(iv) squaring
(v) Substituting values correctly
(vi) Subtraction (vii) Finding square root
(viii) multiplying out

Task number = 8

Academics managing very large classes find MCT items an attractive option due to the ease of marking. The academics may check the overall associated mark distribution for a test, only a few check the psychometric properties of the item within the test. The psychometric qualities of an item can be described in part by item analysis. CTT says that, if there are two persons A & B studying a course and A outperforms B. Then, for every item B gets correct A must get correct, but, there are items that A gets correct that B may not get. On the basis of this we talk about those who know and those who do not know. This leads to the upper and lower 33% of the group, implying that some MCT items are difficult and some are not.

One cannot correctly conclude that person A is better than person B in any item because individuals differ in their composition of traits to responding to test items, since there is a variation in the characteristics of items. This brings to the front burner the concept of IRT. Cognitive complexity and task number as described earlier are sample-independent [8].

So what actually makes an item difficult? Does CTT provide what makes it difficult? Is it that it was not well taught? Is it that the students have not learnt well? Could it be some other factors? Could it be as a result of the cognitive complexity involved? Could it be the higher the more difficult? Or could it be the number of tasks involved?

Purpose of the Study

The aim of this study is to determine:

- the extent to which the variation in item difficulty is accounted for by the variation in cognitive complexity;
- the extent to which the variation in item difficulty is accounted for by in the variation in task number;
- the multiple correlation coefficient between the cognitive complexity and task number, taken together and the item difficulty;
- the extent to which the cognitive complexity and the task number, collectively, account for the variation in the item difficulty.

Research Questions

- What is the coefficient of correlation between the item difficulty and the cognitive levels of the items?
- What is the coefficient of correlation between the item difficulty and task numbers of the items?
- What is the coefficient of multiple correlations between the cognitive level and task number, taken together and the item difficulty?
- To what extent do the cognitive level and task number, taken together, account for the variance in the item difficulty of the item?

Hypotheses

- The coefficient of correlation between the item difficulty and the cognitive level is not statistically significant ($p < 0.05$),
- The coefficient of correlation between the item difficulty and task number is not statistically significant ($p < 0.05$),
- The multiple correlation coefficient of the cognitive level and task number, taken together, and the item difficulty indices do not statistically differ from zero ($p < 0.01$)

Significance of the Study

Suppose the result of the study shows that the task number is significantly associated with the difficulty index of the MCT item, and then evaluators will have to pay significant attention to the predictor variable task number in the area of test development. For instance, the time and effort put in doing item analyses are saved. The concept of task number can be applied in setting test items as estimates of difficulty levels of the items. In some achievement testing circumstances, there is a need to spread candidates over a wide range of marks, this calls for the use of test items with a wide range of difficulty levels. This implies that, in setting MCT items for selection purpose, the evaluator may need to vary the difficulty of the items by including more high cognitive complexity items and higher task numbers, thereby making the difficulty index low. Evaluators will have to put it at the fore-front as a factor in the area of testing. The use of item task number and/or item cognitive complexity has an advantage since the two are independent of groups of testees and will serve as a link between CTT and IRT.

Design and Procedure

This is a correlational study involving the use of two independent variables as predictors, namely, the cognitive complexity and the task number and a dependent variable as the item difficulty - the criterion. The population comprises of 1200 senior secondary school year II students in four selected schools. From four randomly selected intact groups in four schools, a random sample of 200 students was obtained.

Instruments for Data Collection

The instrument used for data collection is a Researcher's Made Mathematics Achievement Test of the multiple-choice test format. The test comprises 70 items. The researcher constructed 100 items based on a test blueprint for the MCT items. The draft was presented to five mathematics educators who are also experts in research and evaluation for scrutiny. The items were reduced, based on the suggestions from experts and distractor analyses. The above actions ensured the face and content validity of the test.

To ensure the reliability of the items, the researcher conducted a trial testing and used scores of the testee in calculating the internal consistency reliability coefficient employing Kuder-Richardson formula 20.

The reliability coefficient so obtained is 0.76. The task number or cognitive level for each item is the median of the task numbers or cognitive levels assigned to the item by the experts.

Techniques of Data Analyses

Using the PPMCC, the coefficient of linear and multiple correlations were calculated and the test statistics were used to test the hypotheses.

- Research questions one is answered by stating the linear correlation coefficient between the difficulty indices and cognitive levels of the items,
- Research questions two is answered by stating the linear correlation coefficient between the difficulty indices and task numbers of the items,
- Research questions three and four are answered by stating the coefficient of multiple correlation between the cognitive level and task number, taken together, and difficulty indices, and the coefficient of correlation between items' task numbers and cognitive level,
- The hypothesis one to three were tested using tabulated critical values of PPMCC while hypothesis four was tested using the F-test statistic.

Results and Interpretation of Data Analyses

The results of the study are summarized in the tables and subheadings below:

Research question 1

What is the coefficient of correlation between the facility indices and the cognitive levels of the items?

Table 1: Linear correlation summary via raw score method for facility indices and cognitive levels.

Variable	Σ	Σ^2	N	ΣZY	R	df	r_{crit}	A	Decision
Z	32.15	16.66	70	78.36	-0.49	68	0.232	0.05	Reject H_O p<0.05
Y	181	519							

Table 1 shows that, the sum and sum of squares for the facility indices are 32.15 and 16.66, while that for the task number are 181 and 519 respectively. Using the PPMCC approach, the calculated coefficient of correlation between the two variables is -0.49, (a negative correlation) which is greater than the critical values of Pearson r (0.232) at 68 degree of freedom at α-level of 0.05; the research question is answered. Therefore the null hypothesis is rejected; this implies that, r_{ZY} is significant, the higher the cognitive levels the lower the facility indices.

Research question 2:

What is the coefficient of correlation between the facility indices and task numbers of the items?

Table 2: Linear correlation summary via raw score method for facility indices and task numbers.

Variable	Σ	Σ^2	N	ΣZX	R	df	r_{crit}	A	Decision
Z X	32.15 228	16.66 787	70	95.96	-0.95	68	0.232	0.05	Reject H_O p<0.05

Table 2 shows that, the sum and sum of squares for the difficulty indices are 32.15 and 16.66, while that for the task number are 228 and 787 respectively. Using the PPMCC approach, the calculated coefficient of correlation between the two variables is -0.95, (a negative correlation) which is greater than the critical values of Pearson r (0.232) at 68 degree of freedom at α-level of 0.05. Therefore the null hypothesis is rejected; r_{ZX} is significant.

Research question 3 & 4:

What is the coefficient of multiple correlations between the facility indices and the cognitive levels and task numbers, taken together?

To what extent do the cognitive levels and task numbers, collectively account for the variation in the facility indices?

Table 3: Hypothesis testing with multiple correlations of cognitive levels and task numbers (taken together) and facility indices

Variable definition	Variables	R	$R_{Z.XY}$	R^2	F	df	α	F_{crit}	Decision
F.I=Z	Z,X	-0.49							
CL=X	Z,Y	-0.60							
TN=Y	X,Y	+0.71					0.05	3.13	Reject H_o
	Z.XY	-	0.60	0.36	18.84	2,67	0.01	4.92	Reject H_o

$$R_{Z.XY} = \sqrt{\frac{r_{XZ}^2 + r_{YZ}^2 + 2r_{XZ}r_{YZ}r_{XY}}{1 - r_{XY}^2}}$$

Research question 3 is answered by $R_{Z.XY} = 0.6$, and for research question 4, the obtained R is squared and multiplied by 100 to get the coefficient of multiple determinations of 36%. For the hypothesis testing proper with multiple correlation, further statistical analysis is required in which the obtained R^2 is transformed into F-ratio, so that the critical value of F distribution is used in testing the significance of the R or R^2. Since F calculated is greater than critical value of F, the null hypothesis is rejected at both 0.05 and 0.01 levels of significance; this implies that $R_{Z.XY}$ is significant. We then conclude that the cognitive level and task number

both predict the item facility indices significantly.

Conclusion

Making fair and systematic evaluation of others performance can be a challenging task and judgement cannot be made solely on the basis of intuition, haphazard guessing or custom. The MCT items are widely used to estimate what students know and can do in specific subject areas. In its extended use by examination bodies, they make visible to teachers, parents and policy makers some of the outcome of students' learning. The versatility and effectiveness of the MCT items is limited only by the ingenuity and talent of the test constructor. Many test constructors are highly experienced in developing questions and judging difficulty. However, the tacit nature of their knowledge prevents its wider use and transfer. A shared understanding of difficulty would give novice question setters' guidelines and make public the notion of difficulty and thus, improve construct validity of test items. There is an urgent need for all examiners, test constructors, and educators to be competent in the use of task number as a factor in item analysis; so that they can vary the difficulty levels of questions and apply the appropriate test, depending on the purpose of the test.

Recommendation

The researchers recommend that, examiners therefore consider a routine post-test analysis of their MCT items. Such an analysis need to include calculations of test reliability coefficient, item difficulty indices, item discrimination indices, and task numbers. Academics need to realize that task number and cognitive complexity considerations can improve the use of MCT items as an assessment instrument. The distinctions made in item cognitive complexity ensure that items will assess the depth of students' knowledge at each benchmark. These assessment practices can be improved primarily through greater teacher awareness

.

References

[1] Blackman, I & Darmawan, I.G. (2004). Graduate Entry Medical Students Variables That Predict Academic and Clinical Achievement. *International Education Journal*, 4(4), 30-41.

[2] Croker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. NY: Holt, Rinehart, & Winston.

[3] Grondlund, N. E. & Linn, R. L. (1990). *Measurement and Evaluation in Teaching* (6th ed.) NY: Macmillan.

[4] Impara, J. C. & Prake, B. S. (1998). Teachers ability to estimate item difficulty: A test of assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35, 69-81.

[5] Izard, J (2005). Trial Testing and Test Analysis in Test Construction. In Ross, K. N. (Ed.) *Quantitative Research Method in Educational Planning* (pp 1-84). Retrieved on 20-2-2012 from http://www.unesco.org/iiep

[6] Kolawole, E. B. (2007). A Comparative Analysis of Psychometric Properties of Two Nigerian Examining Bodies for SSS Mathematics. *Research Journal for AppliedSciences*, 2(8): 913-915.
[7] Nwana, O. C. (2007*). Educational Measurement and Evaluation*. Owerri: Bomaway Publishers.
[8] Ogomaka, P. M. C. (2011). Classroom lecture note.
[9] Sax, G. (1998). *Principles Of Educational & Psychological Measurement and Evaluation.* (3rd ed.). Belmont, CA: Wadworth.
[10] Urbina, S. (2004). *Essentials of psychological testing*. NY: John Wiley & Sons.