What Makes Mid-Tertiary Mathematics Multiple Choice Test Items Difficult: Task Numbers or Cognitive Levels?

P.M.C. Ogomaka

Faculty of Education
Imo State University, Owerri-Nigeria

A.C. Akukwe

Department Of Computer Education,
Alvan Ikoku Federal College of Education, Owerri-Nigeria

M.C. Nosike

Department of Educational Psychology
Imo State University, Owerri-Nigeria

Abstract

The researchers ascertained the linear relationship between any two Multiple Choice Test (MCT) items' facility indices (z), cognitive level (x) and "task numbers (y). A 50 itemed multiple choice test was administered on a random sample of 226 NCE students who offered a General Studies Education (GSE) mathematics course. The test has a reliability coefficient of 0.92. The sampled students' responses were scored. The estimates of correlation coefficients: $r_{zx} = - 0.33$, $r_{zy} = - 0.33$ and $r_{xy} = 0.87$ were obtained and found significant (p<0.05). From the results, cognitive level and task numbers could provide a bridge between Classical Test Theory (CTT) and Item Response Theory (IRT).

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

Introduction

The Longman Dictionary of contemporary English, 3rd edition sees test as "a set of questions, exercises or practical activities to measure some one's skill, ability or knowledge".

For Gronlund, 1985, "Test is an instrument or systematic procedure for measuring a sample behaviour". It answers the question "How well does the individual perform – either in comparison with others or in comparison with a domain of performance tasks?".

Nworgu (2003) stated that "Test connotes a structural situation comprising a set of questions to which an individual is expected to respond". He want further to say that the questions have preferred answers and based on an individual's response, his/her behaviour is quantified. Ogomaka (2011) has a modified view on the general opinion of many who see test as "a measurement instrument". For him, "a test incorporates a measurement instrument, which is the marking scheme that assigns marks but the part that elicits the attribute or behaviour is not a measurement instrument". He illustrated further by saying that "measuring is a systematic assignment of numbers to quantities or qualities or attributes. Test can go beyond measurement and get to the level of evaluation and decision making. In this case, it becomes an instrument for evaluation, and not a measurement instrument". Extending further this argument, test can equally be an instrument for

diagnosis, accountability, teaching, promotion or demotion, and placement of students into Science, Arts, vocational etc classes (Anastasi & Urbina, 1997).

For Gronlund, 1985, some terms used to describe test include informal versus standardized test, mastery versus survey tests, speed versus power test, supply versus selection test etc. In supply test, the examinees are required to supply the answers as is the case with essay tests, or fill in type test. The selection test requires the testees to select the correct response from a set of alternatives as is the case with multiple choice tests. This study worked with Multiple Choice Test (MCT) items.

According to Thorndike (1997), the MCT item is the most flexible test item in use and is made up of the stem which presents the problem and the list of possible options. The testee is requested to read the stem as well as the list of possible options and to select the correct option or best option which is regarded as the key. The rest of the options that are incorrect are called distracters since they function to distract those testees who are in doubt about the correct answer. Highlighting the versatility of the multiple choice test item, Gronlund (1985) stated that it can measure a variety of learning outcomes from simple to complex and it is adaptable to most types of subject-matter content. He also highlighted the six levels of Bloom's taxonomy of cognitive learning outcomes namely knowledge, comprehension, application, analysis, synthesis and evaluation. These represent the item cognitive levels.

Nwana (2007) demonstrated in clear terms how MCT items can be developed in each level of the bloom's taxonomy of cognitive learning. However, Ogomaka (2011) stated that most mathematics educators, considering the nature of mathematics, are of the view that there are four levels of cognitive learning in the area of mathematics namely knowledge comprehension, Application and Analysis. To be able to solve a test item correctly, the testee has to undertake distinct mental or cognitive operations or steps. Unfortunately, these cognitive processes or steps taken by the testee to arrive at the correct answer are not taken into consideration in scoring multiple choice test items. These cognitive steps taken to arrive at the correct answer of a MCT item is what we refer to as "task levels or" "task numbers".

Ogomaka (2011) opined that "task number of an item is obtained by counting the number of distinct mental/cognitive operations/processes/steps involved in answering/working out/solving the item correctly through the common/usual approach (not through short cuts or approaches more advanced than the group of testees". A few instances will suffice).

(i) Multiple 213 by 3. (ii) Multiply 213 by 4 (iii) Find the value of x if $4x - 6 = x+9$
(iv) find the value of a^o in the diagram below.
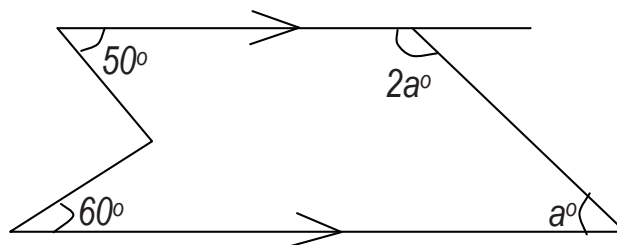
Solution:

i)	2 1 3	Having knowledge of multiplication	+1
	x 3	Multiplying correctly (knowledge of multiplication table)	+1
	6 3 9	Total task number	+2
		Cognitive level is knowledge	+1

ii) 2 1 3   Having knowledge of multiplication   +1
 <u>x   4</u>   Multiplying correctly (knowledge of multiplication table)   +1
 <u>8 5 2</u>   Knowledge of how to "carry over" and add   +2
              Total task number   4
              Cognitive level is comprehension   +2

iii) $4x - 6 = x + 9$

 <u>-x   -x</u>   Subtracting x from both sides   +1

  $3x - 6$ $=$ 9

  <u>+ 6   +6</u>   Adding 6 to both sides   +1

  $3x$ $=$ 15

 <u>÷3   = ÷3</u>   Dividing both sides by 3   +1

 <u>x   =   5</u>   Total task number   +3

              Cognitive level is comprehension   -+2

iv)
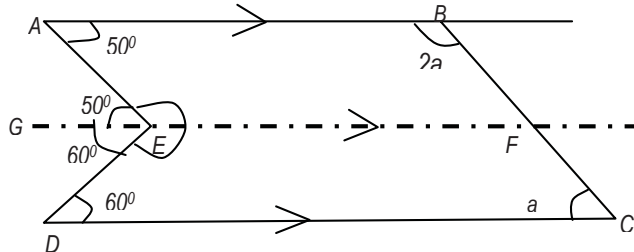
$\overline{EF}$ is parallel to $\overline{AB}$ and $\overline{DC}$ by construction   +1
AÊ G = EÂB = 50^0 Alt. angles   +1
GÊD = <EDC = 60^0 Alt. angles   +1
Sum of interior angles of a pentagon, ABCDE
is (2n - 4)rt angles   +1
That is 2 x 5 (90^0) = 4 (90^0) = 540   +1
AÊD = 360^0 - 110^0 = 250   +1
∴ 50^0 + 2a + a + 60 + 250   =   540   +1
  3a + 360   =   540
  3a   =   540 – 360   +1
  3a   =   180
  ∴a   = 60^0 <u>+1</u>
Task number is   9
Cognitive level is analysis   =   +4

Alternatively the above could be solved thus:

$\overline{EF}$ is parallel to $\overline{AB}$ and $\overline{DC}$ by construction	+1
GÊA = EÂB = 50^0 alternate angles	+1
so AÊF = 180 – 50 = 130^0	+1
<GFB = DĈF = a^0 corresponding angle	+1
In the quad, ABFE,	+1
Sum of inferior angles is 360^0	+1
That is 50 + 2a^0 + a^0 + 130^0 = 360^0	+1
3a + 180 = 360	
3a = 360 – 180 = 180	+1
a = 60^0	+1
Task number is 9, cognitive level is Analysis	+4

Evaluation experts who are mathematics educators validated the cognitive levels and task numbers of the test items used in this study. The reliability coefficient of the test items was established as 0.92 using Kuder Richardson formula 20. While highlighting cognitive levels and task numbers, the authors did not lose sight of the already existing two theories that support the development of multiples choice test items namely Classically Test Theory (CTT) and Item Response Theory (IRT). Classical Test Theory (CTT) is the basis for item analysis which expresses itself in terms of Facility Index (FI), Discrimination Index (DI) and distracter index. There is assumption underlying the CTT thus:

Given three testees x, y, z such that x is better than y and y is better than z, if y gets an item correct, x will get it correct. Any item z gets correct, x and y will certainly get it correct. But if x gets an item correct, y may get it correct or fail it. If y gets an item correct, z may get it correct or may fail it. Here, performance in a test item depends on the abilities of the testees. Hence those who know can be identified and separated from those who do not know. Consequently, one can divide the class into 3 – upper, middle and lower 1/3 (33%) and calculate from there, the facility index highlighted above thus:

The facility index or difficulty index or easiness index of an item is the proportion of testees who answered the item correctly. That is FI = $\dfrac{n_c}{N}$ where n_c is the number of testees who got the item correct while N is the number of testees who responded to the whole test. (Thorndike, 1997; Nworgu, 2003; Nwana, 2007 & Ogomaka, 2011). The above formula can be extended to include scores of group of testees in the upper and lower one third.

Thus:

$$FI = \dfrac{n_u + n_L}{2n}$$ where nu is the number of testees from the upper one third who got the item correct. n_L is the number of testees from the lower one third who got the item correct. n is the number that constitute one third of the testees. Nworgu (2003) pointed out that "an ideal item should have a facility index of 0.50 but realistically it could range from 0.30 to 0.70".

Item Response Theory (IRT) or the latent trait theory according to Thorndike, 1997 assumes the existence of a relatively unified underlying trait, or characteristic, that determines an individual's ability to succeed with some particular type of cognitive task. For Ogomaka, 2011 individuals differ in the composition of traits for responding to items and so one cannot say that individual x is better than individual y in any given item because it depends on the characteristics presented by the item to the testee. Every item is unique and the challenges it throws to the individuals differ. From the above, it is clear that for CTT, item Analysis and the indices there from depend on the group of testees in question. It cannot be constant for any set of items. If a weaker group is used, the indices (facility, discrimination and distracter) will change. IRT on the other

hand is independent of testees or group of testees. Emphasis is on the characteristics of the items independent of the individual testees or group of testees. This situation exposes the fact that there is no bridge between CTT and IRT.

Cognitive Level (CL) and Test Number (TN) as described above are independent of testees. If it is established that TN and CL determine difficulty levels in MCT items, then a bridge has been built to link up CTT and IRT! This study is therefore designed to build this bridge between CTT and IRT in terms of having item characteristics that belong to approaches to items analysis.

Purpose of the Study

This study aims at ascertaining if:
- Item difficulty is associated with cognitive level.
- Task numbers are associated with cognitive levels.
- The extent to which the variation in facility indices of items is accounted for by the variation in task numbers of the same items.
- Item difficulty is associated with task number.
- The extent to which variation in facility indices of item is accounted for by the variation in cognitive levels of the same items.

Research Questions

- To what extent is the variation among facility indices of items accounted for by the variation among the task numbers of the same item?
- To what extent is the variation among facility indices of items accounted for by the variation among the cognitive levels of the same item?
- What is the coefficient of correlation between cognitive levels and task numbers of the items?

Hypotheses:

H_{o1}: The coefficient of correlation between test items facility indices and cognitive levels of the items is not statistically significant ($p < 0.05$).

H_{o2}: The coefficient of correlation between test items facility indices and task numbers of the items is not statistically significant ($p < 0.05$).

Scope

The multiple choice test items involved, covered the entire course content of GSE 212-Basic General mathematics III for NCE year II students. The coefficients concerned in this study are linear.

Design and Procedure

The study is a correlational study. The two independent variables were cognitive levels and task numbers while item facility or item difficulty index is the dependent variable. The area of study is Alvan Ikoku Federal College of Education, Owerri, Imo State.

The population of this study was comprised of 518 NCE year two students from the five schools that make up Alvan Ikoku Federal College of Education, Owerri,. A random sample of 226 students was drawn. The sample of 226 was arrived at using the Yaro Yamen model.

Instrument for Data Collection and its Validation

The instrument used for data collection is a Researcher Made Mathematics Achievement Test (RMAT) of the multiple-choice test format. A test blueprint was developed to cover the entire course content for GSE 212. Following the test blueprint, a 50 itemed test was constructed and sent to five mathematics educators who are also experts in Evaluation for vetting. Some of the items were dropped and replaced while some were modified/reframed based on the suggestions of the experts and distracter analysis. The above action was to ensure face and content validity of the test items. In each case, task number or cognitive level, it is the median of the scores assigned to each item as task number or cognitive level by the five experts that was used for analyses. The instrument was trial tested on a sampled 30 students different from those used for the study. The analysis carried out on the scores of the 30 students using Kuder Richardson formula 20 gave a reliability index of 0.92 for the instrument.

Techniques of Data Analysis

Research question one is answered by stating the linear correlation coefficient between the facility indices and task numbers of the items and the coefficient of determination $100r^2$.
Research question two is answered by stating the linear correlation coefficient between the facility indices and cognitive levels of the items and the coefficient of determination $100r^2$..
Research question three is answered by stating the linear correlation coefficient between the cognitive levels and task numbers of the items. Hypothesis 1 and 2 were tested using the tabulated critical values of PPMCC.

Results:

Research Question 1

To what extent is the variation among Facility Indices (FI) of items accounted for by the variation among the task numbers (TN) of the same item?

Note FI is represented by z; Cognitive Level (CL) is represented by x and TN is represented by y.

$$\bar{z} = 0.41 \, ; \, \bar{y} = 3.18$$
$$\delta z = 0.15; \; \delta y = 1.30$$
$$r_{z.y} = -0.33$$

From the calculation the linear correlation coefficient between the facility indices and task numbers of the items is a negative correlation. There is a correlation. The amount of variation in facility indices accounted for the variation in task numbers of same items is $100 (0.33)^2 \% \simeq 11\%$

Research Question 2

To what extent is the variation among facility indices of items accounted for by the variations among the cognitive levels of the same item.

$$\bar{z} = 0.41 \, ; \, \bar{x} = 2.98$$
$$\delta_z = 0.15; \; \delta_x = 1.01$$
$$r_{z.x} = -0.33$$

Again, from the calculations, there is a correlation (negative correlation) between facility indices and cognitive levels of the items. As under research question one the coefficient of determination is about 11%.

Research Question 3

What is the coefficient of correlation between cognitive levels and task numbers of the item?

$$\bar{x} = 2.98 ; \bar{y} = 3.18$$
$$\delta_x = 1.01 \; \delta_y = 1.30$$
$$r_{z.y} = 0.87$$

Result shows a high positive correlation between cognitive levels and task numbers.

Hypothesis 1

H_{o1}: The coefficient of correlation between test items facility index and cognitive level of the items is not statistically significant. ($p < 0.05$)

Table 1: Linear correlation coefficient summary table for facility indices and cognitive levels.

Variable	Statistics							
	Sum	Sum of square	Sum of product	n	$r_{(cal)}$	∝- level	rcrit	Decision
Z	20.35	9.3986	58.17	50	-0.33	0.05	0.23	Reject H_o
X	149	495						

From table 1, the sum and sum of squares for the facility indices are 20.35 and 9.3986. Those for the cognitive levels are 149 and 495. The calculated coefficient of correlation between the two variables is - 0.33 which is greater than the critical value of Pearson r of 0.23 at 48 degrees of freedom and ∝level of 0.05. The null hypothesis is rejected. The implication is that r_{zx} is significant. That is, the higher the cognitive levels, the lower the facility indices.

Hypothesis 2

H_{o2}: The coefficient of correlation between test item facility index and task numbers of the items is not statistically significant ($p < 0.05$).

Table 2: Linear Correlation coefficient summary table for facility indices and task numbers.

Variable	Statistics							
	Sum	Sum of square	n	Sum of product	$r_{(cal)}$	df	∝- level	rcrit	Decision
Z	20.35	9.3986	50	61.54	- 0.33	48	0.05	0.23	Reject H_o
X	159	590						

From table 2, the sum and sum of squares for the facility indices are 20.35, and 9.3986. Those for the task numbers are 159 and 590. The calculated coefficient of correlation between the two variables is – 0.33 which is greater than the critical value of Pearson r of 0.23 at 48 degrees of freedom and $\propto$ level of 0.05. The null hypothesis is rejected. r_{zy} is significant.

From the above, it stands to reason that task numbers and cognitive levels determine difficulty levels in multiple choice test items. That is, the coefficient of correlation between facility indices and task numbers is not zero.

Recommendations

- ➢ The study should be carried out using mathematics students and core mathematics courses
- ➢ The study should be carried out in different courses in other disciplines
- ➢ Test developers should welcome the newest concept, "Task Numbers" and employ it to improve upon the quality of their MCT items.

Conclusion

If it is found that cognitive levels and task numbers are significantly associated with facility indices, it means new concepts - cognitive levels and task numbers will have to be introduced in testing. Again, for selection purposes, test, developers should include high level cognitive items as well as items with high task numbers. What is more, the use of task numbers and cognitive levels which themselves are independent of groups of testees will serve as a link between CTT and IRT. Since determining the difficulty of items through IRT is rigorous and involves large number of testees, task numbers and cognitive levels may be used instead for small groups.

References

Anastasi, A and Urbina, S. (1997). *Psychological Testing* (7th Ed). New Jersey. Pearson Education Inc.
Gronlund, N.E. (1985) *Measurement and Evaluation in Teaching.* New York: Macmillan Publishing company.
Kpolovie, P.J. (2010). *Advanced Research Methods* Owerri.Springfield publishers Ltd
Kpolovie, P.J. (2010). *Statistical Technique for Advanced Research.* Owerri Spring field Publishers Ltd.
Mehrens, W. A and Lehmann, I.J. (1978) . *Measurementand Evaluation in Education and Psychology.* London. Holt, Rinehart and Winston.
Nwana, O.C. (2007): Educational Measurement and Evaluation theory and practice. Nsukka. University Trust Publishers.
Nworgu, B.G. (2003). *Educational Measurement and Evaluation theory and practice.* Nsukka University Trust publishers.
Ogomaka, P.M.C (2011): *Inferential Statistics for Research in Education and social Sciences.* Owerri Peace Wise System and Prints.
Ogomaka, P.M.C (2011): Classroom lecture.
Thorndike, R.M.C. (1997). Measurement and Evaluation in psychology and Education (6th Ed). New Jersey. Practice Hall Inc.