**Journal of Modern Education Review, ISSN 2155-7993, USA**
**January 2014, Volume 4, No. 1, pp. 57–73**
**Doi: 10.15341/jmer(2155-7993)/01.04.2014/008**

**http://www.academicstar.us**

# Item Types and Upper Basic Education Students' Performance in Mathematics in the Southern Senatorial District of Cross River State, Nigeria

*Joy Dianabasi Eduwem[1], Imo Edet Umoinyang[2]*

*(1. School of Health Information Management, University of Calabar Teaching Hospital, Calabar, Nigeria;*

*2. Institute of Education, University of Calabar, Calabar, Nigeria)*

**Abstract:** This study was designed to establish the influence of item type on Upper Basic Education students' academic performance in mathematics. Using an ex-post facto design with a population comprising all Upper Basic Education students in both public and private secondary schools within the Southern Senatorial District of Cross River State, Nigeria, a sample of 886 Upper Basic Education (UBE 3) students in the Senatorial District supplied the needed data. This sample was drawn using the stratified random sampling technique. The same Mathematics achievement test constructed by the researchers was in three different forms, namely: the multiple choice test, the essay test and the completion test. The data were collated and analyzed using the Analysis of Variance repeated measures. The study revealed that there is a significant influence of item type on UBE students' performance in mathematics. Students performed significantly better in mathematics test presented in multiple choice form than in the completion and the essay form. The study concluded that students' performance in mathematics at the Upper Basic classes could be influenced by the item type administered during testing. Based on this empirical finding, recommendations were made to the authorities concerned.

**Key words:** item types, mathematics performance, multiple choice test, essay test, completion test

## 1. Introduction

For any society to grow and be enlightened, its members must acquire good education. Education is the vehicle through which men and women acquire knowledge and skills that make them useful to themselves and the society. The Nigerian National Policy on Education (FGN, 2004 revised) states that education is the pivot to national development. Education trains the mind, increases the individual's understanding of his environment and broadens his capability to deal with complex situations. It enables the individual to function more actively and maximally for personal, or societal development (Sandy, Baum & Ma, 2007; Olatunde, 2010).

To achieve national development, society strives to offer well-rounded education to her members, and thus, organize education into bodies of knowledge and skills which encapsulate every aspect of human endeavor. The series of experiences offered in the schools are often referred to as subjects and are broadly categorized into two paradigms; the sciences and the arts, although developmental priorities of societies often make them to focus on either one or both of these paradigms.

Joy Dianabasi Eduwem, Ph.D., School of Health Information Management, University of Calabar Teaching Hospital; research areas: measurement and evaluation. E-mail: eduwemjoy2@yahoo.com.

The Commission of Science and Technology for Development Network (CSTDN) [2011], states that Science and Technology have become the major paradigm for development in many modern societies. This is where the importance of mathematics as a school subject is evident. Mathematics is by its nature concerned with objective and quantitative measurement of ideas and concepts in the environment. According to Cangiano (2008), mathematics "is the queen of Science and the language of nature". It is about pattern and structure and about logical analysis, deduction and calculations within these patterns and structures. When patterns are found, often widely different areas of science and technology, the mathematics of these patterns can be used to explain and control natural happenings and situations (Brown & Porter, 2011).

In nations the world over, efforts have been geared towards effective scientific and technological development. In making such efforts, the utility value of mathematics has remained an indispensable factor. To Gowers (2008), mathematics provides the basis upon which basic concepts and ideas are built and organized for societal advancement. It is an essential tool for human progress. Therefore, for the assurance of meaningful growth of the Nigerian society, training the youths in mathematical skills is imperative. Nigeria is now embarking on "vision 20:20:20", which is seen as a strategy for catapulting Nigeria into being one of the top twenty industrialized countries of the world by the year 2020. Since the realization of this dream depends on the acquisition and utilization of scientific knowledge, the educational system at various levels has been adjusted to give priority attention to the teaching and learning of science subjects including mathematics. Although that is the case, the teaching and learning of the sciences and particularly mathematics in the secondary schools in Nigeria, leave much to be desired.

Evidence (e.g., Georgewill, 1990; Bergeson, Fitton & Blysina, 2000; Korau, 2006; Schoenfeld, 2007) has shown that there are problems in the teaching and learning of mathematics in the Nigerian secondary schools. Their positions are informed by low performance recorded every year in the subject by students.

The relatively low achievement in mathematics, a core subject, recorded by upper basic students the Southern Senatorial District of Cross River State appears a threat to Nigeria's dream of effective industrialization. This is so because without adequate knowledge of mathematics and its application, having scientific and technological advancement is a mirage.

In order to solve these problems, the Federal Government of Nigeria has made meaningful effort at improving on teaching and learning generally in schools and particularly in mathematics. Moreover, Government has improved its efforts in the training of mathematics teachers through scholarship awards to individuals who upon graduation would come back to the school system to effectively teach mathematics. That apart, the problem of lack of infrastructural facilities in these schools has also been largely addressed. These notwithstanding, performances in mathematics over the years in the schools and in external examinations have not shown any remarkable improvement.

Aside from efforts made by Government towards improving mathematics performances, various schools have committed much effort towards solving this problem. For instance, extra classes in mathematics are usually arranged for students after normal school hours. Furthermore, mathematics quiz competitions (internal and external) are often organized by schools in a bid to develop the intellect of students. Also, best mathematics students, at the end of each term, are usually encouraged by being awarded prizes. This practice, apart from encouraging the recipients to do even better, serve to motivate others to also put in their best in studying mathematics. Schools have also employed the services of counselors to give useful counsel to students on why they should put in extra efforts in studying and achieving good results in mathematics. Most schools that are of

private ownership nowadays employ qualified graduates of mathematics to teach their students. Such teachers are usually very well paid and such schools usually charge huge amount of tuition fees in order to have enough money to adequately pay for services provided by staff, of which most parents usually gladly pay to ensure that their wards are given the very best academically.

Yet, despite all these efforts, students' performances in mathematics over the years have not yielded the expected results. This had given one the impetus to research on how to aid students' improvement in their mathematics performance. In this study, effort will be directed at finding out if item type is relevant in explaining upper basic education students' performances in mathematics.

Item type is a basic consideration in item construction involving test of knowledge and skills in any subject in schools. Under test construction, emphasis is placed among others, on the students' ability to understand the test items and the potential levels of students' performance in such a test. Item types are either selective or constructive. Selective response items are questions which require the test-taker to choose the correct answer from a number of options such as the multiple choice questions, true/false, matching items etc. Constructive response items involve tasks which require the test-taker to provide constructive solutions and ideas such as the essay-type questions, short answer and completion items in an examination.

### 1.1 Objective of the Study

The objective of the study was to determine if item types do influence upper basic education students' performance in mathematics.

### 1.2 Research Question

Does item type influence upper basic education students' performance in mathematics?

### 1.3 Hypothesis

Item type does not significantly influence upper basic education students' performance in mathematics.

## 2. Literature Review

### 2.1 Item type and Students' performance in Mathematics

Bridgeman (1992) compared quantitative questions in open-ended and multiple-choice item types. In an experiment designed to determine how closely scores on multiple-choice tests correlated with scores on an open-response test, Bridgeman (1992) found that there was a statistically significant correlation between the scores. The main question he addressed was the extent to which the open-ended versions of the items paralleled the multiple-choice versions in terms of difficulty, discrimination and correlational structure. For his study, Bridgeman (1992) developed opened-ended items from the quantitative (GRE-Q) and verbal (GRE-V) section of the Graduate Record Examination.

Bridgeman (1992) developed a scannable answer sheet that allows almost any question from the current GRE-Quantitative test (GRE-Q) to be answered in an opened-ended response format. The candidates had to grid-in number or a simple formula rather than selecting answers A-E. The answer sheet accommodates decimals, fractions, negative numbers and equations with one variable. Also the researcher developed seven experimental test forms with items adapted from the discrete quantitative and data interpretation sections of two GRE-Q forms (83-1 and 83-3). A symbol for Pi was not included in the answer sheet but candidates were to use a value of 3.1 for problems that demanded Pi. Also problems that demanded the interpretations of graph were at times modified

to allow approximate answers, e.g., "to the nearest $100". Two items out of the 30 of them could not be modified easily for the free-response type. The quantitative comparisons items from the two old forms were retained in their original multiple-choice format on the experimental test forms.

The experimental form contained 10 four choice quantitative comparison items and 14 grid-in items in 30 minute (i.e., six fewer items than usual). There were seven forms of the test that were administered. Forms 1 and 2 were spiraled and administered in test centers in one set of states Forms 3–7 were spiraled and administered at centers in a different set of states. Forms 6 and 7 provide a replication for Form 3 and 4 to make certain that findings are not unique to that one set of items. To address reliability concerns, parallel forms of the tests were administered at other testing centers. The elements of the seven test treatments were as follows: (1) Form 1 contained 30 multiple-choice questions from section 4 of Form 83-1 of a regular GRE examination, with no indication that the test was experimental; (2) Form 2 was identical to Form 1 but contained an indication that the test was experimental; (3) Form 3 was identical to Form 1 but used a special experimental answer sheet; (4) Form 4 contained 10 multiple-choice items that were identical to the 10 items on Form 1 and 14 open-ended questions that corresponded to multiple-choice questions in Form 1, 2, and 3; (5) Form 5 was the same as Form 4 but contained only five multiple-choice items; (6) Form 6 contained 30 multiple-choice questions from section 3 of Form 83-3 of a regular GRE examination; and (7) Form 7 contained 10 multiple-choice questions from Form 6 and 14 open ended questions that corresponded to multiple-choice questions in Form 6.

Participants were randomly assigned to one of the test treatments, and were informed that their test results would not affect their regular GRE scores. This action was taken to eliminate potential effects of test stakes on participants' performance. Data collected during the study included students' test scores and responses to survey instruments that measured student's attitudes about item types (Bridgeman, 1992). No reliability data was provided for the instruments used in the study. An examination of the test scores based on item difficulty indicated that a higher percentage of students correctly answered the easiest multiple-choice items than the easiest open-ended items. Bridgeman attributed this finding to the possibility of students guessing the correct answer or using other questions for corrective feedback. This conclusion may have been partly based on the finding that 88% of the students responded that they had at one time or another worked backwards on a multiple-choice test to obtain an answer. Differences in scores on multiple-choice versus open-ended items varied depending on the level of difficulty of the question. Bridgeman concluded that items that were relatively easy in the multiple-choice test were relatively difficult in the open-ended test. However, it was unclear whether difficulty was defined as an item characteristic based on the distracters used in the question, or was determined by the number of students who answered the question correctly.

Bridgeman (1992) suggested that open-ended items may be more indicative of the specific skills that students possess, due to the possible effects of guessing. Another implication was that the nature of distracters used in multiple-choice questions could contribute to the variations in item difficulty and in item scores. Comparisons of total test scores, though, indicated that scores attained on the multiple-choice items were comparable to scores attained on the open-ended items.

The 14 items for computer delivery and scoring were identical to the items in the scannable portion of the study. The candidates used the computer keyboard to enter a numerical answer for all the responses. The questions were designed for delivery on networked PCs with one file server for five to nine PCs. The scannable items (or their multiple-choice counterparts) were administered at all of the test centers in six states as the last section of a regular GRE General Test administration. This last section required a special answer sheet and contained

experimental items. Students receiving the special answer sheet were informed at the beginning of the last section that their scores on that section would not count toward their regular GRE scores. Forms were spiraled within the center to assure random assignment to multiple-choice or grid-in versions of the test.

The computer items were administered to 364 students out of 3,277 candidates that had taken October 1989 GRE General Test, who had completed the Biographical information Questionnaire (BIQ) and who lived near one of the four Educational Testing service (ETS) where the computer test was to be administered. The data for multiple-choice and Grid-in items was screened by excluding only the most obviously unmotivated cases.

A maximum of 2.1% of the cases was deleted from any form. Means and standard deviations on the experimental items for males and females on Forms 1–3 were presented. The means were based only on the 14 items used in subsequent analyses of test type (i.e., multiple-choice vs. grid-in) effects. In spite of the slightly lower quantitative ability of the Form 3 students (as was indicated by regular GRE scores), means scores for the unmotivated Form 3 students were less than one question lower than the mean scores for the Form 1 students. After the adjustment using the operational GRE-Q and GRE-V scores as covariates, the intercepts differed by less than 0.2 of a question. An analysis of covariance (with GRE-Q and GRE-V as covariates) was conducted comparing the Form 1 scores on the 14 common items with the Form 3 scores on the same items for the 3,810 students who provided sex and ethnicity information and who indicated that English was their best language.

Also four ethnic categories were used, Asian American, Black, Hispanic and White. Each main effect and interaction was corrected for every other term in the model. It was noted that despite the large sample size, not a single main effect or interaction was significant at the 0.05 level. For critical interactions of form with gender and ethnicity the p-values were greater than 0.5. In Bridgeman (1992) study, the effect of knowing that scores do not count appeared to be minimal in terms of across and within gender and ethnic groups.

It was observed that 71% of the examinees answered the easiest items correctly in the grid-in item type. That number was still not nearly as high as the 92% who got it correct in the multiple-choice test. The higher percentage of examinees answering the items correctly in the multiple-choice test was explained to be caused not only by the opportunity to guess but also by the implicit corrective feedback that is part of the multiple choice test. The 14 questions that were common across the three item types (computer, grid-in and multiple-choice) were presented. The percent-correct scores were always highest for the multiple-choice items which were explained to be due to that impact of guessing. It could be deduced that the actual percent correct sometimes differed remarkably from the expected value.

For the grid-in and multiple-choice item types, students were randomly assigned but the score for the computer format came from a separate sample of paid volunteers. The mean GRE-Q score in this group was 577 (SD = 129) compared to the GRE-Q means of 543 (SD = 131) in the form 4 sample. However, with a few exceptions, the percent correct of the grid-in and computer groups were similar. The largest difference between the computer and grid-in groups in terms of percent correct was 8 points; for the grid-in and multiple-choice samples, a difference of at least 8 points between the actual and expected percent correct was found for 6 of the 14 questions. Thus for individual question-level analyses the difference between free-response and multiple choice tests appeared to be much more important than the difference between computer and grid-in administration modes, even after adjusting for the impact of random guessing.

This study found that "if the intent of the test is to describe specific skills that student possess, the open-ended test seems to be clearly superior" (Bridgeman, 1992, p. 269). Bridgeman's study further found that examinees scored higher on multiple-choice tests than open-ended tests; also, he found that total test scores in the

open-ended and multiple choice test types appeared to be comparable. He went on to say that both "ranked the relative abilities of students in the same order, gender, and ethnic differences were neither lessened nor exaggerated; correlations with other test scores and college grades were about the same" (Bridgeman, 1992, p. 269).

Bridgeman (1992) used three-parameter item characteristic curves for graphically comparing the performance of items in the multiple-choice (Form 3) and grid-in (Form 4). Some of the curves were quite similar for the multiple-choice and grid-in groups while others differed remarkably. It was clear from the curve that the change in item type does not produce a uniform impact on all items or at all ability levels within the same item. He found that data from both form comparisons suggest that item statistics derived from the multiple-choice test administration may provide only a crude estimate of how the item will perform in an open ended test. His analysis of test type effects on total scores revealed that gender difference, in standard deviation units was nearly identical in the multiple-choice Form 3 (0.44) and the grid-in Form 4 (0.39).

For the Form 6 (multiple-choice) versus Form 7 (grid-in) comparisons, the gender difference was again about the same for both forms (0.42 and 0.45 for Forms 6 and 7 respectively). He carried out covariance analysis comparing Form 3 (multiple-choice) to Form 4 (grid-in). This indicated the expected form effect ($F [1, 4044] = 213.5$, $P < 0.0001$) and no significant two-or three-way interactions (all ps. > 0.14). The replication comparison of Form 6 to Form 7 also indicated a form effect ($F [1, 3862] = 24.9$, $P < 0.0001$) and to no significant two-way interactions (Ps. > 0.6). However the three-way form X gender X ethnicity interaction was statistically significant at the .05 level in this large sample ($F[3,3862] = 2.74$, $P = 0.04$). This was of little practical significance, contributing only 0.0007 to the squared multiple correlation (i.e., accounting for 0.07% of the total score variance).

From the above analysis he felt that it would appear that differential test type effects by gender or ethnicity are non-existent or trivial. His speededness analysis reveals that the additional time in Form 5 was of minimal benefit, with average scores improving by less than 1 point. He carried a correlation analysis with the assumption that if quantitative scores from the grid-in and multiple-choice tests are measuring the same underlying construct, they should have similar correlations with other variables.

From this presentation, the correlation of the 14 common items from the experimental forms have about the same correlation with regular GRE scores and undergraduate grade point average (UGPA as reported on the BIQ) regardless of question formats. The coefficient alpha reliability for the multiple-choice Form 3 was 0.77 and for the grid-in Forms 3 and 5 the reliabilities were 0.78 and 0.81 respectively. The alpha reliability for the multiple-choice Form 6 was 0.64 and for Form 7 it was 0.73. He concluded that in spite of substantial test type differences noted for individual items, total scores for the multiple-choice and open-ended tests demonstrated remarkably similar correlational patterns. Also that there was no significant interactions of test type with either gender or ethnicity.

Pajares and Miller (1997) found in a study of eighth grade pre-algebra and algebra students that students taking a multiple-choice test out scored those students who took the same test in an open-ended format. Students that took the multiple-choice test answered on average 3.67 more questions correctly than those that took the open-ended test. Although this study was more on how self-efficacy affects mathematical performance the results support Bridgeman's study in that students can score higher on multiple-choice tests than on open-ended tests.

In another study, Martinez (1991) looked into the comparison of multiple-choice and constructed figural response items. The latter includes illustrations and graphs as the response medium. In his investigation, Martinez

(1991) contrasted multiple-choice and constructed response items. Two parallel forms of a test were constructed. One using multiple-choice questions, the other using constructed-response questions that are stem equivalent. This stem-equivalence means that item stems (questions or instructions) are identical, but the response options provided in the multiple-choice format are eliminated in the alternative format.

His subjects were from Grades 4 (N = 347, 8 (N = 365) and 12 (N = 322) form a national sample of students representing a broad range of characteristics such as racial/ethnic group, socioeconomic status, and national region of residence. He also gathered data in conjunction with field testing for the 1990 National Assessment of Educational progress (NAEP). The items he used consisted of twenty-five constructed response items written in three content areas: life sciences, physical sciences, and earth/space sciences. The questions were developed in accordance with NAEP content and process specifications. Each question was classified by topic (e.g., ecology), nature of science, e.g., designing an experiment), and thinking skill (e.g., solving problems). These items according to the researcher were reviewed by an outside panel of scientists. He matched each figural item with a multiple-choice counterpart, some which were already part of NAEP item pool. Twenty-three of the items had four response options each; while the remaining two offered five response options.

Aside from the stem differences which was needed to clarify the intended response (e.g., "draw an X" or "draw arrow heads") wording was parallel across items. These items were part of a larger field test of science items. Subjects were given three blocks of science items and were allowed 15 minutes per block. The figural response items composed an entire block whereas the multiple-choice items which typically were answered more quickly, were accompanied by ancillary multiple-choice science items not connected with the study. The assignment of subjects to condition (Figural response or multiple choices) was random. He observed that for across grade levels, constructed-response questions were, in general, more difficult than their multiple choice counterparts. He also pointed out that relative difficulty between item types interacted with the difficulty of the question in its constructed-response form. For such questions that were relatively difficult (P less than or equal to 0.5), constructed-response questions were almost uniformly more difficult: 29 out of 33 being more difficult as constructed response items (sign test, $P < 0.001$). But it was noted that items with $P > 0.5$, 9 of 11 were more difficult in the multiple-choice form (sign test, $p < 0.05$).

For these calculations items that were administered at more than one grade level were counted separately at each grade level. He also made a scatter plot of item difficulties in the two item types. The plot showed that data points fall above the diagonal illustrating that in general multiple choice items were easier than constructed-response items. It also revealed that only a few of the items were very difficult ($p < 0.2$) in their multiple-choice forms, whereas a number of items were very difficult in their constructed-response forms.

Results of this study therefore indicates that there was a significant difference in performance of students in the multiple choice and the constructed response tests as students' scores were significantly higher in the multiple choice items than the constructed response items. He observed also that item/total score (r-biserial) correlations were generally higher for constructed response items than for their multiple-choice counterparts.

The discrimination offered by constructed-response items was moderated by whether they were easier or more difficult than their multiple-choice counterparts. Where the constructed-response versions were easier, format differences indiscrimination were small. The mean discrimination values for those items were recorded as follows: at Grade 4, 0.62 for constructed-response and 0.65 for multiple-choice. At Grade 8, 0.56 for constructed-response and 0.51 for multiple-choice. At Grade 12, 0.53 for both constructed response and multiple-choice. Where constructed response items were more difficult, he observed that advantages in

discrimination for constructed-response test type were more evident.

Mean discrimination values were: Grade 4, 0.62 for constructed-response and 0.43 for multiple-choice. At Grade 12, 0.60 for constructed response and 0.42 for multiple-choice. He noted that the reliabilities for total scores were marginally better in the constructed-response test at Grades 8 and 12. The standard errors were seen to be uniformly and remarkably lower for the constructed-response items across grades. The researcher noticed that when the problem was given to the 12 grade students, 43% of the constructed responses corresponded to the four multiple-choice options including 9.7% that matched the correct option. Across all items and grades, 63% of the constructed-responses corresponded to the multiple-choice options, 39.8% matched the correct option. The researcher concluded that figural-response items were generally more difficult especially for questions that were difficult ($P < 0.5$) in their constructed-response forms. Results showed that students' scores in the multiple-choice questions were significantly higher than in the figural response items. Also, that figural-response questions were slightly more discriminating and reliable than their multiple choice counter parts although they had higher omit rates.

In another study, Shohamy (1994) examined the effect of item types (multiple choice and open ended items) and language of assessment (first language = L1 vs. second language = L2) on performance in second language (L2) reading tests. The main conclusion drawn from her study was that item type can affect test scores: multiple choice items were found to be generally easier to answer than open-ended items. She attributed this to different language processes required to do the tasks.

Wolf (1993) examined the effects of different assessment tasks, languages of assessment and L2 language competence on L2 reading comprehension test performance. She compared effects of tests with multiple-choice items, open-answer items and cloze-tests. In cloze-tests, test developers delete every nth word in as passage and test takers have to fill in each blank. Wolf found that the item type used to assess learners' reading comprehension affects their test results: test takers' performance on the multiple choice items was significantly better than that on the open-ended and cloze-test tasks.

At the University of Pretoria, web-based courses in Mathematics, using WebCT, have been running for several years. Students are assigned groups to participate in the on-line quizzing courses. Groups are large with up to two hundred students. The structure and nature of the courses are discussed in detail in Engelbrecht and Harding (2001(1) and (2)). Online quizzing forms an integral part of the web-based courses. Firstly, students do a weekly online quiz, used as a formative tool. Secondly the two semester tests and the final examination taken per semester each comprises of a combination of a paper section and an online section, carrying equal weight. The investigation into performance differences in constructive response and multiple choice questions exposed in this paper, is based on calculus courses taught in the first two semesters of study.

Comparing performance in online multiple choice questions and online constructive response questions, the question addressed in this section is whether there is any intrinsic difference between the performance level of students in online multiple choice questions and in online constructive response questions in a first year calculus course. For the investigation the following experiment was conducted: In the online section of one of the semester tests in a first semester calculus course taken by 83 students, the same concept is assessed twice in almost identical questions, firstly formulated as a constructive response question and two questions later in multiple choice format. The question in constructive response form requires a single, numerical answer and tests whether the student knows and can apply the chain rule for differentiation. The question in multiple choice form (no negative marking) tests the same concept but now in multiple choice form. The two questions and average

performance of the group of students are as follows:

Comparison of an online constructive response question with an online multiple choice question.

Online constructive response question:

If f(x) = cos3 (3x+1), give a value for f '(2).

Give your answer to two decimals and work in radians.

Online multiple choice question:

If f(x) = sin5 (2x-6) then f '(x) is given by

• 5 sin4 (2x-6)
• 10 cos4 (2x-6)
• 10 sin4 (2x-6) cos (2x-6)
• 10 cos4 (2x-6) sin (2x-6)
• None of the above

Average mark: 57.2% Average mark: 94.4%

There is a remarkably large difference in scores and judging from these two questions alone it appears that students experienced a substantial difference in the degree of difficulty between the two question types. The fact that most students had the multiple-choice formulation of the question right is an indication that the students recognized the right application of the chain rule amongst the distracters. Yet they did not observe that the two questions are in essence the same, otherwise students would surely have gone back to the constructive response question to correct a possibly wrong answer.

It is true that the constructive response formulation has the added complication of having to substitute a value to obtain the correct answer but that alone can hardly account for the large difference in scores. In another semester with a different group, the experiment was repeated and the two questions and average marks were as follows:

Comparison of an online constructive response question with an online multiple choice question

Online constructive response question:

The function f(x) = e3x is approximated by a Taylor-polynomial of degree 1 around x = 0. What is the value of the Taylor polynomial approximation at x = 0.1? One decimal.

Online multiple choice question:

The Taylor polynomial of degree 1 of y = Ln x around the point x = 1 is given by:

• x – 1
• x
• 1 – x
• 1 – 1/x
• None of the above

Average mark: 35.4% Average mark: 44.2%

Again students performed better on average in the multiple choice question but the difference was definitely not as large and both averages were low. It was not surprising that students showed poorer performance in finding Taylor polynomials than in a simple differentiation manipulation.

Ozuru, Best, Bell, Witherspoon, and Namara (2007) conducted experiments on the influence of test formats on reading comprehension performance. They varied question formats, i.e., multiple choice vs. open-ended items and passage availability, i.e., allowing the test taker to access the text while answering comprehension questions

(with-text condition) or taking the text away (without-text condition). While the authors found high and significant correlations between the test takers' performance when answering multiple choice items and open-ended items in the without-text condition, there were only very low and non-significant correlations in the with-text condition. Ozuru et al. (2007) deduced that "the processes underlying open-ended and multiple choice test items answering in the with-text condition are likely to share less similarity [than those in the without-text condition]" (p. 426). The studies outlined generally found that multiple choice items in reading assessment seem to be easier than open-ended items.

Shohamy (1994), Pearson et al. (1999) and Ozuru et al. (2007) further suggested differences in underlying reading processes, but only Pearson et al. (1999) further investigated the nature of these differences. One way to approach differences in answering multiple choice and open-ended items and underlying reading processes is to look for differential correlations to reading precursor skills like general cognitive abilities, vocabulary knowledge, orthography knowledge and reading fluency.

According to Chan and Kennedy (2002), however, a high correlation between scores on multiple-choice and constructed-response items could be misleading. For example, scores could be strongly correlated, but one score could still be significantly higher than the other score. Upon hypothesizing that students' scores on a multiple-choice test would be higher than scores on a constructed-response test, Chan and Kennedy (2002) conducted a study to determine the extent of the correlation between test scores. The types of constructed-response items used in the study were questions that required an answer in the form of a brief sentence or phrase. The basis of this decision was that this type of constructed response item was most prevalently used in the literature.

For the purposes of the experiment, equivalence of multiple-choice and constructed response items was established by using the same answer stem in both item types. The experiment also controlled for guessing on multiple-choice items by ensuring that questions could not provide implicit feedback for answering other questions. Two tests, A and B, were administered as the students' final examinations in an economics course (Chan & Kennedy, 2002). There were 196 students who were randomly assigned to either test type. The composition of each test was 36 multiple-choice items, 12 constructed-response items, and 7 other questions. The first 12 multiple-choice items on Test A were equivalent to the constructed-response items in Test B. Similarly, the first multiple-choice items on Test B were equivalent to the first 12 constructed-response items on Test A.

Data collected during the study consisted of students' test scores (Chan & Kennedy, 2002). The multiple-choice questions were graded objectively, with a point given for each correct answer. The constructed-response questions were graded subjectively by one person. It was not clear, though, if the grader knew any of the participants or knew which tests belonged to which participants. It was also unclear whether a rubric or set of criteria was used in grading responses to the constructed-response items. However, the report mentioned that no partial credit was given for incomplete answers. Chan and Kennedy (2002) used a statistical procedure to correct scores on the multiple-choice items for the effect of guessing. Although there was a discussion of the method that was used, no reliability data was provided for the statistical procedure. There was, however, a very detailed explanation and justification of the statistical methods that were used to analyze the test scores.

Recognizing that students could also guess on some of the constructed-response questions, Chan and Kennedy conducted two sets of analyses on scores. Unadjusted test scores were used in one set of analyses, and scores that had been adjusted to correct for the effects of guessing were used in the other set of analyses. Chan and

Kennedy had defined expected difference questions as items where students could guess on the multiple-choice version of the item, but could not guess on the corresponding constructed-response item. There were also items of no expected-difference in which the students could guess on both the multiple-choice and the constructed-response items. There were eight expected-difference questions and seven no-expected difference questions.

Analysis of the test scores indicated that students scored higher on the multiple-choice questions when there was an expected difference between the multiple-choice and constructed-response items. Students scored comparably on items where there was no expected-difference. Although students scored higher on the multiple-choice items than on the constructed response items for the expected-difference questions, the results were not statistically significant at an alpha level of 0.05.

One explanation for the differences in test scores was the use of certain distracters on the multiple choice tests. Chan and Kennedy (2002) suggested that these distracters may have prevented students from engaging in the same thought processes that would have been used in free recall. Analysis of the no-expected-difference questions indicated that there was no significant difference in the scores on adjusted multiple-choice questions compared to scores on the equivalent multiple-choice questions.

The implications were that multiple-choice items could facilitate the deduction of an answer whereas constructed-response items did not provide similar cues, and that constructed-response items could simulate multiple-choice questions when the answers were overtly obvious. These conclusions, which were similar to conclusions made by Bridgeman (1992), underscored the need to carefully consider the distracters that are used in multiple-choice test items.

Prior studies that have attempted to evaluate the equivalence of multiple-choice and open-response items on the basis of test scores have produced inconclusive results, according to Hancock (1994) because the studies lacked clear definitions of the cognitive skills that each item type measured. Consequently, interpretation of the results of such studies to demonstrate equivalence of test types has limited validity.

In a study designed to determine the "degree to which multiple-choice and constructed-response tests measure the same cognitive skills" (p. 144), Hancock controlled for cognitive complexity. The theoretical basis used to define the cognitive requirements of each test item was Bloom's taxonomy. Reasons that Hancock provided for selecting this theoretical basis were wide acceptance of the taxonomy and lack of empirical evidence that contradicted the taxonomy. Accordingly, cognitive complexity was measured in terms of four taxonomic levels: knowledge, comprehension, application, and analysis.

The participants in Hancock's (1994) study were 90 students who were enrolled in either an introductory educational measurement course or a research statistics course at a Pacific Northwest university. Forty-six of the students were enrolled in the educational measurement course and forty-four students were enrolled in the research statistics course. Students in the educational measurement course were administered two midterm examinations and one final examination, while students in the research statistics course were administered one midterm examination, and one final examination. It was unclear why one class received two midterms. While some of the test items had been used in prior semesters, most of the instruments were created for the purposes of the experiment. Steps taken to ensure alignment of the test items with content presented in class included a review by Hancock. The midterm examinations consisted on 40 test items, with 5 items designated for each taxonomic level; the final examinations consisted of 48 test items, with 6 items designated for each taxonomic level. The tests were composed by analyzing each item in the test bank and classifying the item based on Bloom's taxonomy.

To address reliability concerns, an additional qualified rater reviewed the categorization that Hancock had performed. Only items that had been assigned to the same taxonomic category by each rater were selected for the tests. Although Hancock noted that the use of identical stems on the multiple-choice and constructed-response items would have improved reliability, this method was intentionally not used. The reason given for not using identical stems was to prevent one test item from influencing the results of another item. The design of the multiple-choice items was a stem and four possible answer choices. No empirical basis was presented for the determination of the appropriate number of answer options. Answers for the constructed-response items ranged from a word or number to an explanation. Two graders scored all of the examinations, and there were no scoring inconsistencies between graders. For the purposes of analyzing the test scores, Hancock (1994) used a summed score of the responses in each taxonomic grouping for each participant.

To improve reliability, Hancock excluded the item with the lowest correlation between the item score and the total score for each taxonomic level. The results of the study indicated that the mean scores attained for the multiple-choice tests were not uniformly higher than the mean scores attained on the constructed-response tests for each class. Although this finding was based on the assumption that the use of the summed scores was valid, Hancock concluded that the lack of consistently higher scores indicated that guessing did not influence the scores.

While acknowledging that interpretation of the results was predicated on the assumption that the multiple-choice and constructed-response items provided reliable correlations, Hancock (1994) indicated that there were high levels of association between the formats across each taxonomic level. Correlations were lowest, though, for the taxonomic level of comprehension in the educational measurement class and the analysis level in the research statistics class. Hancock attributed this difference across the classes to the nature of the subject matter. Given the limitation of reliability assumptions, Hancock concluded that scores on the multiple-choice tests were highly related to the scores on comparable constructed-response tests that measured the same taxonomic level. Explanations provided for the findings were that the cognitive processes involved in recall tasks were also used in the selection of answers on multiple choice tests and that the processes of recall and recognition were similar processes.

## 3. Methodology

The research design adopted in this study was the ex-post facto design. The study was carried out in the Southern Senatorial District of Cross River State. The district comprises seven local government areas namely; Akamkpa, Akpabuyo, Bakassi, Biase, Calabar Municipality, Calabar South, and Odukpani.

The population of this study comprised all upper basic education students in both public and private secondary schools within the Southern Senatorial District of Cross River State of Nigeria. A total population of about 50,747 students were enrolled into the upper basic education within this district in the 2011/2012 academic session. The stratified random sampling technique and the simple random sampling technique were adopted in this study. The Southern Senatorial District was stratified based on the seven local government areas namely; Akamkpa, Akpabuyo, Bakassi, Biase, Calabar municipality, Calabar South and Odukpani. In each local government area, 10% of public secondary and 10% of private secondary schools were selected for the study. This gave a total of 16 study schools. The simple random sampling procedure was used in selecting the 16 schools that participated in the study. The sample for this study comprised 886 upper basic education students from both public and private secondary schools.

The mathematics achievement test was constructed by the researcher in three different forms, namely: the multiple choice test, the essay test and the completion test. All three forms of the mathematics test were one and the same test, and measured the same thing, though presented in different forms.

When the instrument was subjected to the Cronbach Coefficient Alpha reliability test, reliability coefficients of 0.81, 0.78 and 0.83 was obtained respectively for multiple choice, essay and completion items respectively. The collected data was analyzed using repeated measures- a form of general linear model. The test was carried out at .05 significance levels.

## 4. Findings

There is a significant influence of item type on upper basic education students' performance in mathematics. Results indicated that students performed significantly better in the Mathematics test presented in multiple choice form than in essay and completion items. Also, Students performed significantly higher in completion than essay items. Table 1 and Table 2 presents the results.

**Table 1 Summary Data and Analysis of Variance Results of within Subject Test of Item Types**

| S/NO | Item Type | N | $\overline{X}$ | SD |
|---|---|---|---|---|
| 1 | Multiple choice | 886 | 7.74 | 4.15 |
| 2 | Completion item | 886 | 6.63 | 3.89 |
| 3 | Essay item | 886 | 5.76 | 3.81 |

| Source | Item Type | SS | DF | MS | F-ratio | value | Partial Eta Squared |
|---|---|---|---|---|---|---|---|
| Item type | Multiple choice vs. Essay | 3472.290 | 1 | 3472.290 | 363.631* | .000 | .292 |
| | Essay vs. completion | 665.485 | 1 | 665.485 | 139.269* | .000 | .136 |
| | Mc vs completion | 1088.644 | 1 | 1088.644 | 131.209* | .000 | .129 |
| Error | Multiple choice vs. Essay | 8431.710 | 885 | 9.527 | | | |
| | Essay vs completion | 4219.515 | 885 | 4.768 | | | |
| | Mc vs completion | 7326.356 | 883 | 8.297 | | | |

From Table 1, the mean students' performance in Mathematics was highest for the multiple choice item-type ($\overline{X}$ = 7.74; SD = 4.15) while the least mean performance in Mathematics ($\overline{X}$ = 5.76, SD = 3.81) was by students in essay item type. Therefore, the students studied showed their maximum performance in Mathematics when given multiple choice item tests and minimum Mathematics performance in essay type tests.

The result in Table 1 also compares the students' performance in Mathematics on the basis of the different item types. That is, variation within the students' scores based on item types. The ANOVA results showed a significant F-value of 363.631; (P = .000) of variation within the scores obtained by students in both multiple choice and essay type items. This means that significant differences exist within the students' performance in both multiple choice and essay items, when compared.

Also, the results from Table 1 showed that the calculated F-value of 139.269; (P = .000) was statistically significant at 0.05 level compared within values of performance in mathematics between essay and completion item tests. This implies that students' performance in Essay and completion vary significantly within their individual scores and based on the item types. Also, the results from Table 1 showed that the calculated F-value of (131.209; P = .000) was statistically significant at 0.05 level compared within values of performance in mathematics between multiple choice and completion item tests. This implies that students' performance in

multiple choice and completion vary significantly.

Furthermore, the variation in the students' performance in Mathematics based on the comparison of their performances in multiple choice and essay items resulted in a partial eta square of .292. This shows that variation in students' performance based on multiple choice items and essay questions reasonably explains up to 29.2 percent of their performance in Mathematics. However, the explained variance in the students' mathematics performance due to essay and completion test items is 13.6 percent. This implies that partial differences of up to 29.2 percent variation in Mathematics performance by students is explained and attributed to the use of multiple choice items over essay type items. That is, students are likely to perform up to 29.2 percent better in multiple choice items when compared with essay items. Also, the use of completion test accounts for 13.6 percent variation in the students' Mathematics performance over essay type items. Also, the variation in the students' performance in Mathematics based on the comparison of their performances in multiple choice and completion items resulted in a partial eta square of .129. This shows that variation in students' performance based on multiple choice items and completion questions reasonably explains up to 12.9 percent of their performance in Mathematics.

Moreover, Table 2 presents the results of the analysis of variance testing the variation in the students' Mathematics performance based on their average scores in the three types of tests used.

**Table 2 Test of between Subjects Effects of the Influence of Item Types on Students' Performance in Mathematics**

| Source | SS | Df | MS | F | P-value | Partial ETA Squared |
|---|---|---|---|---|---|---|
| Intercept | 39667.575 | 1 | 39667.575 | 3027.503* | .000 | .774 |
| Error | 11569.425 | 885 | 13.073 | | | |

Transformed variable: Average

*Significant at .05 level

From Table 2, the calculated F-value was 3027.503; P = .000. This value was significant at .05 probability level as indicated by the P-value. Also, the partial eta squared was 0.774.

The implication of this result is that significant differences exist in the students' mean scores based on the item types. Therefore the hypothesis which states that item types do not significantly influence students' performance in Mathematics is rejected. Indeed, item type significantly influences students' performances in Mathematics.

Also from the result of Table 2 the partial eta square value of 0.774 implies that item types accounts for up to 77.4 percent of students' performance in Mathematics in relation to the essay type, multiple choice and completion items of the Mathematics Achievement Test.

Finally to examine the points of significant mean differences in the students' mean performance in Mathematics based on item types, the Fisher's least significant Differences (LSD) was used as the Post Hoc test analysis. The result is presented on Table 3.

**Table 3 Fisher's LSD Test of Points of Significant Differences of the Influence of Item Types on Students' Performance in Mathematics**

| S/NO | Item Type(i) | Item Type(j) | Mean Difference(i-j) | S.E | P-value |
|---|---|---|---|---|---|
| 1 | Multiple choice | Essay | 1.982* | .104 | .000 |
| | | Completion | 1.114* | .097 | .000 |
| 2 | Essay | Multiple choice | -1.982* | .104 | .000 |
| | | Completion | -.868* | .074 | .000 |
| 3 | Completion item | Multiple choice | -1.114* | .097 | .000 |
| | | Essay | .868* | .074 | .000 |

*Mean difference is significant at .05 level.

The result in Table 3 shows that significant marginal mean differences in students' performance in mathematics exist based on the difference in item types. For instance, students significantly perform better in Multiple choice items than in Essay items, and significantly better in Multiple choice than in Completion items. They perform significantly lower in Essay items than in Completion items.

## 5. Discussion

The findings of this study showed that item types significantly influence students' academic performance in mathematics. It is not surprising that students showed better performances in multiple choice items than in essay and completion item types. One possible explanation for the magnitude of difference between scores on multiple choice and completion or essay test is that multiple choice items are fundamentally recognition task which only demands identification of the correct response. They require only a recall of declarative knowledge, while the essay items require higher levels of cognitive processing. In addition, factors such as guessing could contribute to the high scores obtained in multiple choice items. This result is in line with Bridgemann (1992) findings, which used three-parameter item characteristic curves for graphically comparing the performance of items in the multiple-choice (Form 3) and grid-in (Form 4). Findings revealed that 71% of the examinees answered the easiest items correctly in the grid-in item type while 92% got it correct in the multiple-choice test.

The findings of this study is also in line with results obtained by Pajares and Miller (1997) who found in a study of eighth grade pre-algebra and algebra students that students taking a multiple-choice test out scored those students who took the same test in an open-ended format. Students that took the multiple-choice test answered on average 3.67 more questions correctly than those that took the open-ended test. Although this study was more on how self-efficacy affects mathematical performance the results support Bridgeman's study in that students can score higher on multiple-choice tests than on open-ended tests.

Results obtained by Martinez (1991) are also in line with findings of this study. Martinez (1991) looked into the comparison of multiple-choice and constructed figural response items. Results indicates that there was a significant difference in performance of students in the multiple choice and the constructed figural response tests, as students' scores were significantly higher in the multiple choice items than the constructed figural response items.

The findings of this study are also in line with the findings obtained in the research obtained at the University of Pretoria where web-based courses in Mathematics, using WebCT, were used for the study. Comparing performance in online multiple choice questions and online constructive response questions, the results of the two experiments together with the difference in averages indicate a difference in performance between online multiple choice and online constructive response questions. The average for the multiple choice questions was 63.08% and the average for the online constructive response questions was 53.05%.

Findings from this study is also in line with results obtained by Shohamy (1994) who examined the effect of item types (multiple choice and open ended items) and language of assessment (first language = L1 vs. second language = L2) on performance in second language (L2) reading tests. The main conclusion drawn from her study was that item type can affect test scores: multiple choice items were found to be generally easier to answer than open-ended items.

Also, findings obtained in this study is in line with results obtained by Wolf (1993) who examined the effects of different assessment tasks, languages of assessment and L2 language competence on L2 reading comprehension test performance. Wolf (1993) compared effects of tests with multiple-choice items, open-answer items and

cloze-tests. In cloze-tests, test developers delete every nth word in as passage and test takers have to fill in each blank. Wolf found that the item type used to assess learners' reading comprehension affects their test results: test takers' performance on the multiple choice items was significantly better than that on the open-ended and cloze-test tasks.

Also in line with the findings of this study are results obtained by Ozuru, Best, Bell, Witherspoon, and Namara (2007) who conducted experiments on the influence of test formats on reading comprehension performance. They varied question formats, i.e. multiple choice vs. open-ended items and passage availability, i.e. allowing the test taker to access the text while answering comprehension questions (with-text condition) or taking the text away (without-text condition). While the authors found high and significant correlations between the test takers' performance when answering multiple choice items and open-ended items in the without-text condition, there were only very low and non-significant correlations in the with-text condition. Ozuru et al. (2007) deduced that "the processes underlying open-ended and multiple choice test items answering in the with-text condition are likely to share less similarity [than those in the without-text condition]" (p. 426). The studies outlined generally found that multiple choice items in reading assessment seem to be easier than open-ended items.

The results obtained in this study is also in line with results obtained by Chan and Kennedy (2002) who conducted a study to determine the extent of the correlation between multiple choice and constructive response test scores. Analysis of the test scores indicated that students scored higher on the multiple-choice questions when there was an expected difference between the multiple-choice and constructed-response items. Students scored comparably on items where there was no expected-difference.

However, some prior studies had attempted to evaluate the equivalence of multiple-choice and open-response items on the basis of test scores and had produced inconclusive results. For instance, findings obtained by Hancock (1994) differed slightly from the findings of this study. The results of the study indicated that the mean scores attained for the multiple-choice tests were not uniformly higher than the mean scores attained on the constructed-response tests for each class. Although this finding was based on the assumption that the use of the summed scores was valid, Hancock concluded that the lack of consistently higher scores indicated that guessing did not influence the scores.

## 6. Conclusion

Based on the findings of the study, the conclusion drawn was that Performance in mathematics in the upper basic education could be influenced by the type of item administered during testing.

## 7. Recommendations

Parents should use the result of this study to advise and encourage their children to put in more effort in their studies, improve on their study skills and prepare adequately for their examination in whatever form it may take to enhance better performance.

Examination bodies should take note of the findings of this study and also note that the various item types used as tools for assessments have limitations when used individually, as such, both the multiple choice and essay item types should be used in carrying out assessment of mathematics achievement at the JSS level, for more valid, reliable and comprehensive but should ensure that a greater percentage of examination questions are administered using the multiple choice formats for improved performances.

**References**

Bergeson T., Fitton R. and Blysma P. (2000). *Teaching and Learning Mathematics*, New York: McGraw hill.

Bridgeman, B. (1992). "A comparison of quantitative questions in open-ended and multiple-choice formats", *Journal of Educational Measurement*, Vol. 29, No. 3, pp. 253–271.

Brown R. and Porter T. (2011). "Why study mathematics", available online at: http://www.popmath.org.uk/centre/pagescpu/lmahob95.html.

Cangiano A. (2008). "The importance of mathematics", available online at: http://mathslog.com/2008/03/31-the-importance-of-mathematics.

Chan N. and Kennedy P. E. (2002). "Are multiple-choice exams easier for economics students? A comparison of multiple choice and equivalent constructed response examination questions", *Southern Economic Journal*, Vol. 68, No. 4, pp. 957–971.

Engelbrecht J. and Harding A. (2001). "Mathematics at the University of Pretoria: The trial run", *South African Journal of Science*, Vol. 97, No. (9/10), pp. 368–370.

Federal Government of Nigeria (Revised) (2004). *The National Policy on Education*, Abuja: NERD.

Georgewill J. W. (1990). "Causes of poor achievement in West African School Certificate Mathematics examinations in Rivers state secondary schools, Nigeria", *International Journal of Mathematics Education, Science & Technology*, Vol. 24, No. 3, pp. 379–385.

Hancock G. R. (1994). "Cognitive complexity and the comparability of multiple-choice and constructed-response test formats", *Journal of Experimental Education*, Vol. 62, pp. 143–157.

Korau Y. (2006). "The problems of mathematics science teaching", available online at: http://www.abu.edu.ng/staff_details.Php?staffed=2563.

Martinez M. (1991). "A comparison of multiple choice and constructed figural response items", *Journal of Educational Measurement*, Vol. 28, No. 2, pp. 131–145.

Ozuru Y., Best R., Bell C., Witherspoon A. and McNamara D. S. (2007). "Influence of question format and text availability on the assessment of expository text comprehension", *Cognition and Instruction*, Vol. 25, pp. 399–438.

Pajares F. and Miller M. D. (1997). "Mathematics self-efficacy and Mathematical problem-solving: Implications of using different forms of assessment", *Journal of Experimental Education*, Vol. 65, pp. 213–228.

Pearson P. D., Garavaglia D., Lycke K., Roberts E., Danridge J. and Hamm D. (1999). *The Impact of Item Format on the Depth of Students' Cognitive Engagement*, *Washington*, DC:Technical Report, American Institute for Research.

Sandy I., Baum R. and Ma J. (2007). "Education pays: The benefits of higher education for individuals and society", *Journal of Social Sciences*, Vol. 5, No. 2, pp. 201–205.

Schoenfeld A. H. (2007). "Mathematics teaching and learning", available online at: http://gse.berkeley.edu/faculty/ahschoenfeldmathteachingandlearing.pdf.

Shohamy E. (1994). "The validity of direct versus semi-direct oral tests", *Annual Review of Applied Linguistics*, pp. 188–211.

The Commission on Science and Technology for Development Network (CSTDN) (2011). "Global technological development", available online at: http://www.unctad.info/en/science-and-technology-for- developmentstDev/.

Wolf D. F. (1993). "A comparison of assessment tasks used to measure FL reading comprehension", *The Modern Language Journal*, Vol. 77, pp. 473–489.