UNIVERSITY OF CAPE COAST

THE IMPACT OF ITEM POSITION IN MULTIPLE-CHOICE TEST ON STUDENT PERFORMANCE AT THE BASIC EDUCATION CERTIFICATE EXAMINATION (BECE) LEVEL

BY

SAMUEL NII NMAI OLLENNU

Dissertation submitted to the Department of Educational Foundations of the Faculty of Education, University of Cape Coast, in partial fulfilment of the requirements for award of Master of Arts Degree in Educational Measurement and Evaluation

JUNE 2011

DECLARATION

Candidate's Declaration

I hereby declare that this dissertation is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature... Date................................

Name: Samuel Nii Nmai Ollennu

Supervisor's Declaration

I hereby declare that the preparation and presentation of the dissertation were supervised in accordance with the guidelines on supervision of dissertation laid down by the University of Cape Coast.

Supervisor's Signature.. Date................................

Name: Prof. Kafui Etsey

ABSTRACT

The study investigated the impact of item position in multiple-choice test on student performance at the Basic Education Certificate Examination (BECE) level.

The sample consisted of 810 JSS 3 students selected from 12 different schools. A quasi-experimental design was used. Students for the project were drawn from schools representing public and private schools in urban and rural areas. The instrument for the project was a multiple-choice test consisting of 40 items in each of English Language, Mathematics and Science. The items were arranged using the difficulty order to obtain the three treatments i.e. Random (RDM), Easy-to-Hard (ETH) and Hard-to-Easy (HTE). The data collected were subjected to statistical analysis using ANOVA at $p \leq 0.05$.

The results of the analysis indicate that for English Language, Mathematics and Science at the BECE level, when item order was altered, the difference in performance was generally significant. However, there was no significant difference in performance between the ETH and HTE treatments of English Language. Similarly, there was no significant difference in performance between the RDM and ETH, and RDM and HTE of Science. The results for Mathematics were consistent, in that significant difference in performance was observed between all the three treatments for the subject.

The study led to the conclusion that the proposition of using re-ordering of format of an objective test to curb examination malpractice may not be the best after all especially in English Language, Mathematics and Science at the BECE level. It was therefore recommended that other methods should be investigated for the purpose.

ACKNOWLEDGEMENTS

My appreciation goes to my supervisor, Prof. Kafui Etsey for his patience and direction. I also wish to express my gratitude to the Management of the West African Examinations Council, Accra Office for sponsoring me in this study.

My thanks also go to Mr. Kofi Boakye and Mr. Felix Akuffo-Badoo for their immense support during the write up. I cannot forget my other colleagues who had to run errands for me during the project.

Last, but not least, I am grateful to my dear wife – Beatrix – who constantly reminded and encouraged me to complete the project.

DEDICATION

To my last son - Zephaniah - who I often refer to as my bonus baby

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

Background to the Study

Tests play an important role in determining achievement and certifying attainment. Tests are also used in providing incentives and goals for students, and providing answers for decision-making. This is because they offer facts and data that help in understanding people and offer some measure of their capabilities. Thus, in matters of selection for higher education and placement into jobs in both the private and public sectors, tests have been the obvious choice since they are devoid of most biases that may be termed systematic and provide equal chances to all candidates. The use of tests for selection and engagement is summative. In the formative use, diagnostic tests are administered to students to determine their weaknesses with the view to designing remedial programmes for their academic advancement. Tests can also be employed in defining the curriculum and structuring teaching and learning. In this vein, tests are regarded as yardstick to measure the effectiveness of educational policies. Thus, tests can be put to numerous uses. As Cronbach (1979) puts it: "Tests are neutral; they serve those who want to maintain society without change, and they are a weapon available to those who want to criticize present institutions and create a society based fully on merit and self-determination" (p.39).

There is no gainsaying the highly advantageous use of tests but tests have varied social implications and limitations. Consequently, the use of tests in obtaining facts and data about people has received its fair share of criticisms. The first of the social implications put forward by Anastasi (1976) is the issue of invasion of privacy of the examinee. She asserted that through a test an examinee may be compelled to reveal certain facts and details about himself which may embarrass him. Certainly, any intelligence, aptitude or achievement test may reveal limitations in skills and knowledge that an individual would rather not disclose. Another limitation put forward by critics is the fact that testing produces rigid grouping practices so far as the test result leads to inflexible classification, also referred to as categorization, labelling or grading. Worthen and Spandel (1991) argued that classification could be demeaning and insulting and harmful to students who are relentlessly trailed by low test scores. This obviously carries connotations which may cause more harm than any gain that could possibly come from such classifications. They further claimed that one of the most serious indictments on the use of test is the fact that most tests favour economically and socially advantaged students over their counterparts from lower socio-economic backgrounds. They argued that even well-intentioned uses of tests can disadvantage those unfamiliar with the concepts and language of the majority culture producing the test items.

Fear of the demeaning social implications of tests, generate the anxieties that accompany test taking and receiving of test scores. In extreme cases, anxious testees suffer from phobias that lead to biological disorders often referred to as examination fever, examination diarrhoea and temporary weakening of the bladder. For these reasons, Gronlund (1976) is reported to have regarded all tests to be having damaging effect on pupils because they can create anxiety and destroy their self-concept.

In an attempt to circumvent the negative effects imposed by the social implications of testing, which include impedance to academic progress, forfeiture of professional advancement and promotion, and the stigma of being labelled a non-achiever, many examinees resort to various kinds of malpractices during test-taking. The temptation to indulge in this vice is sometimes so strong that candidates who could be classified as well-behaved and would ordinarily not approve of wrong-doing fall prey to it. Yet, the phenomenon is a vice that should not be tolerated since it threatens the moral fibre of the society and can lead to the selecting of misfits into vital and sensitive positions. Unfortunately, in a report presented at the 52nd Annual Council Meeting of The West African Examinations Council (WAEC) in Freetown in March, 2004 it was clearly indicated that the phenomenon is on the increase. Tauber (1984) also reported that in university introductory courses which are usually heavily enrolled, one of the difficult challenges which the university authorities have to grapple with is that associated with cheating during examination administration. Pettijohn and Sacco (2001) also confirm cheating during university examinations and professors had to adopt all forms of methods to control the phenomenon.

By definition, an examination malpractice is any act that contravenes the rules and regulations which govern the conduct of the examination. The act could happen before, during or after the examination. Adeyegbe and Oke (1994) defined examination malpractice as ‘an impropriety, an improper conduct to one’s advantage during an examination’ (p.1).

Examination malpractice could take different forms. An outline of some of them is given below:

(i) Bringing foreign materials into the examination hall.

Materials which when brought into the examination hall constitutes an offence are notebooks, textbooks, notes on sheets of paper, blank pieces of paper or any other printed materials. In recent times, the list has been expanded to included mobile phones. Possession of these during an examination constitutes an offence when the rules of the examination prohibit it.

(ii) Irregular activities: These include

(a) stealing, converting or misappropriating other candidates' scripts;

(b) substituting or exchanging worked scripts during or after the examination;

(c) seeking or receiving help from non-candidates such as invigilators, teachers or other personalities during the examination.

(iii) Collusion: This is when a candidate passes on notes for help to other candidates or passes the notes on from other candidates, receiving or giving assistance to any candidate in anyway and talking to or with another candidate during the examination. Collusion also includes copying from the work of other candidates during the examination.

(iv) Other malpractices: These include impersonation, leakage and insult or assault on the supervisor or invigilator.

Examination malpractices are not new. For instance, the first recorded occurrence of examination malpractice in West Africa happened in Nigeria in 1914, when the Cambridge School Certificate Examination leaked. Since then there has been reported cases in all the different levels of the educational system: from the basic to the tertiary, from the civilian institutions to the disciplined forces. Adeyegbe and Oke (1994) reported what Adeyegbe observed:

> I witnessed an example of a show of shame by someone who was hidden in a place close to the examination hall announcing through a microphone the options to the multiple-choice items in one of our (WAEC) examinations. The voice was being heard, but the person responsible was not seen (p.6).

Hassan (2005) conjectures that while certain components, such as processing of registration data and scoring of multiple-choice objective tests, have been automated, examination personnel will continue to play a vital role. The credibility of any public examination therefore is dependent upon the personal and professional integrity of everyone involved in the system.

In recent times, examination malpractice has assumed a sophisticated technological dimension with the use of the cell phones as a means of transmitting answers to both multiple-choice and essay tests by both voice and text messages. This I state from my personal experience during inspection of some centres while examination was in progress.

It is sad to note that in the 15th October, 2005 issue of the *Daily Graphic,* a Ghanaian newspaper, Ransford Tetteh (a journalist working for the paper) reported that while on inspection at an examination centre in Accra during the November/ December 2005 Senior Secondary School Certificate Examination for private candidates, an officer of the West African Examinations Council discovered a number of candidates using cell phones to cheat. They were using the SMS feature to text answers to multiple-choice items from one candidate to another. Acting rightly, he confiscated the phones and instructed the candidates to collect them later after he had checked whether or not they contained information which may suggest they have been used in cheating. This did not go down well with the candidates and in the end the officer lost his life through a mob beating by the candidates. This demonstrates the extent to which perpetrators of examination malpractice could go.

One of their recommendations to examining bodies for curbing such practices, Adeyegbe and Oke (1994) said, is to “think of administering parallel tests to different students but having the same psychometric properties” (p. 11). According to Anastasi (1976) the use of several alternative forms of a test provides a means of reducing the possibility of cheating. Pettijohn and Sacco (2001) reported that to prevent cheating on examinations, many professors will mix up the order of multiple-choice test questions from examination to examination without thought of the consequences the change of order may have on student examination performance and perceptions. Text-book companies even provide randomization options for preparing examinations using electronic test banks to assist in this common practice. Carlson and Ostrosky (1992) stated that multiple forms of an examination are frequently used as a means of reducing likelihood of cheating in large classes. However, they noted that questions have been raised regarding whether the order of test items influences student performance. In the same vein, Bresnock, Graves and White (1989) claimed that objective testing in large sections of introductory economics classes is increasingly prevalent today. To eliminate cheating, several versions of tests are administered. They agreed that constructing several versions of tests poses several issues for those assigning grades. They asked “is it fair to give different versions of the same exam to different students?” (p. 239).

Statement of the Problem

Going by the recommendations of Adeyegbe and Oke (1994), Anastasi (1976), Pettijohn and Sacco (2001) if examining bodies are to minimize, if not to eliminate, the incidence of collusion in multiple-choice tests then they must use parallel tests or alternative forms of the multiple-choice test or mixed-up versions of the same test. However, Carlson and Ostrosky (1992) have raised questions regarding whether the order of test items influences student performance. There is therefore an indication that altering order of test items may have implications for the performance of the testees.

It is not known whether using different forms of a multiple-choice test at the Basic Education Certificate Examination level has any significant impact on the performance of the candidates. The focus of the study was, therefore, to find out what impact will change of item position to create different forms of a test have on performance of candidates.

Purpose of the Study

The study investigated the impact of item position in multiple-choice test on student performance at the Basic Education Certificate Examination (BECE) level. The purpose of the study is to obtain justification or otherwise for the development and use of different forms of a test to curb examination malpractice at that level of learning.

Research Questions

Questions raised were:

(i) What would be the effect of a change in item order on candidates' performance in English Language at the BECE level?

(ii) What would be the effect of a change in item order on candidates' performance in Mathematics at the BECE level?

(iii) What would be the effect of a change in item order on candidates' performance in Science at the BECE level?

Significance of the Study

The results of the study will be of interest to the West African Examinations Council (WAEC) since one of the major items on their agenda for various committees is how to deal with the menace of examination malpractice. Several task forces have been put together to come up with strategies in this regard. The findings of this study will in no doubt go a long way to assist in this connection.

Examination malpractice does not only confront examining bodies like

WAEC, but also other learning institutions. In contemporary times, owing to the large numbers of students in one class, many tutors have resorted more to the use of multiple-choice tests to assess their students in final examinations. In this connection, they are likely to be faced with the phenomenon of collusion. Thus the finding of this study will be of importance to learning institutions which employ the use of multiple-choice examinations.

Other stakeholders who use the results of examinations for selection and placement may also have interest in the findings of the study since its employment may improve the reliability of the test scores for their use.

During a recent visit to H.E., the President of the Republic of Ghana, Prof. J. E. Atta Mills by Council members of WAEC, he indicated his support for whatever effort the Council would put in to eliminate examination malpractice and urge them not to relent in any way. This is quite indicative that the Government may also be interested in the findings of this study.

The significance of this research is the discovery of a justification or otherwise for the use of different forms of a multiple-choice test in the Basic Education Certificate Examination to arrest the incidence of collusion among examinees, or at least, to discourage it. Should the difference in performance turn out to be not significant across the different forms, then a method of curbing one form of examination malpractice in public examinations has been found and would go a long in improving the credibility of the administration of multiple-choice tests as well as the reliability of the test scores obtained.

Delimitation of the Study

The study covered three subject areas which are compulsory for all candidates of the BECE. These are English Language, Mathematics and Science. These subjects were chosen because they are those with very high stakes among the lot. They are also critical in the selection exercise for the secondary level of education in Ghana.

Assumptions

In undertaking this research, the following assumptions were made:

(i) In the administration of the various forms of the tests, different groups of sample were used. It is assumed that the characteristics of the different groups are the same and this therefore would not affect the outcomes of the trial tests.

(ii) Because the test items used were crafted by professionals they are without flaws and capable of soliciting the intended responses from the testees;

(iii) There is no difference between the observed scores of the testees and their true scores.

(iv) The participants were not affected by test-wiseness.

Limitations

A quasi-experimental design was used. This design has an inherent limitation arising from the lack of random assignment which ultimately precipitates into low internal validity. Since this design does not require any random pre-selection process, the selection of the sample was mainly by convenience and this has limiting implications for generalization and lowers the external validity.

Some data were lost through computer virus infection in the course of processing the data collected. This resulted in low sample figures for some of the treatments. This could translate into a limitation with implications for generalisation.

Definition of Terms

Random order (RDM)

The test in which the items are randomly arranged in order of the syllabus or topics and not according to difficulty levels.

Easy-to-Hard order (ETH)

The test in which the items are arranged starting from the easiest ones and ending with the most difficult ones.

Hard-to-Easy order (HTE)

The test in which the items are arranged starting with the most difficult ones and ending with the easiest ones.

Organization of the Rest of the Dissertation

In Chapter 2, the results of a review of related literature in connection with the problem was given starting from the broad and finally narrowing to the problem at stake. On the main, it is an attempt to capture theories that are for and against the core issue of the research to provide a sound theoretical background for the research.

The methodology used is outlined in Chapter 3 indicating the research design, the sample and population. A description of the instrument used, the data collection procedure and analysis of the data collected are also given in the outline of the methodology. In the study, alternate papers of a multiple-choice objective test were developed. These had the same items but differently ordered. The order of the response options however, was not scrambled. This was to avoid the introduction of other factors which are not being investigated in this research.

Chapter 4 is a presentation and discussion of the findings of the research. Interpretation of the findings, general comments and recommendations including suggestions for future research work and concluding statements are put together in Chapter 5.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

Introduction

The study investigated the impact of item position in multiple-choice test on student performance at the Basic Education Certificate Examination (BECE) level. The purpose of the study was to obtain justification or otherwise for the development and use of different forms of a test to curb examination malpractice at that level of learning.

In this chapter, I present my findings from literature. The literature review has been grouped under the following subheadings:

1. Reasons for Assessment
2. Disadvantages of Use of Tests
3. Advantages and Usefulness of Tests
4. Fairness of Objective Tests
5. Effect of Item Order/Arrangement
6. Effect of Textbook Content Order/Arrangement
7. Impact of Use of Alternative Tests
8. Impact of Changing Response Order
9. State of the Art

Reasons for Assessment

Regarding assessment, Ohuche and Akeju (1976) noted that universally individuals differ in personalities and abilities. To appraise these differences, assessment or examinations of one kind or the other have been used from time immemorial. They added that each examination may consist of one or several tests. Gekoski (1964) stated that the differences in ability and personality are intangible because most of the characteristics involved are not physical entities and therefore are not amenable to physical manipulations. He listed the characteristics as follows: Intelligence, Interest, Personality, Special Abilities and Attitude. He added that the characteristics are abstractions of behaviour and though intangible they exist in different degrees from person to person. To be able to meaningfully compare the degrees of the behavioural abstractions in different persons, Gekoski stated that they need to be converted to measurement.

Ohuche and Akeju (1976) wrote that measurement in any field of human endeavour involves the assignment of numerical value to quality or attribute in a person or thing. There should be a tool to be used in assigning the values and a well-defined body of rules for assigning such values. Gekoski (1964) asserted that the tool to be used in assigning the values is psychological testing and stated that psychological testing, in the scientific frame of reference, describes human characteristics in fractionated, dimensionalized quantitative terms. He further remarked that in so doing it improves communication, enables reliable and objective description and analysis of human characteristics. It also facilitates the prediction of human behaviour.

Advantages and Usefulness of Tests

In spite of the above difficulties, Gekoski (1964) stated that educational assessment has much usefulness. In support, Ohuche and Akeju

(1976) wrote that tests serve many useful purposes in the educational system, in industry and in the world of work.

Ohuche and Akeju then listed some of them as follows:

1. Stimulus for Studying: Tests and test results provide impetus for learning and the stimulus which an average pupil or student needs for studying.
2. Administrative Decisions in Education: Test results are used in making administrative decisions about students, teachers and the curriculum.
3. Diagnosis: Test results are used in identifying weaknesses and strengths in a class of pupils. This aids in the desirable effort of giving pupils individual and remedial attention. One other diagnostic use of tests is for guidance and counselling to the most appropriate course to undertake.
4. Selection and Placement: Selection tests are very useful instruments for picking round pegs for round holes and square pegs for square holes. Selection picks among many people the best for a course, career or training and it is institution-centred. Placement chooses among many qualified career persons, the best for a position.
5. Certification: The most popular among the uses of examination is to determine who will receive what certificate. These certificates serve as passports to job, higher institution and instrument for social mobility.
6. Maintenance of Standards: Standards represent the minimum degree of excellence which society can accept. For example, professionals like lawyers and medical officers must meet certain standards before they

are allowed to practice their profession. Usually these standards are enforced through examinations which may be written, oral or practical.

7. Research: Most researchers in education depend somehow on tests. Tests are thus very useful tool in the hands of researchers, in curriculum work, teaching methods and learning theories.

Ohuche and Akeju (1976) assert that there two main forms of tests. These are the free response form – essay and short-answer items and the structured-response or objective form – like multiple-choice, true-or-false and completion items. Ohuche and Akeju wrote that the objective test derives its name from the fact that the marking is done with a standard key and is thereby objective.

Disadvantages of Use of Tests

Ohuche and Akeju (1976) stated that difficulties arise when it comes to educational measurement since the characteristics to be measured are not well defined and these difficulties translate into disadvantages. Their list is itemised as follows:

1. The complexity of the human nature: This has to do with the heredity of man, his environment and the effect of the interaction of these two on him. In addition one cannot ignore the general unpredictable changes which occur within him.

2. Use of Indirect Measurement Methods: Owing to the fact that most of the attributes to be measured are in-born and one cannot get inside a

person to measure these in-born traits, indirect methods are resorted to and we cannot be too sure of what we are measuring.

3. Effect of the Environment: The environment in which the individual human being is living is constantly changing. It is not always easy to narrow down this change and the corresponding effect on an attribute which is to be measured.
4. Measuring Instrument: There is the issue of the instrument to be used in educational measurements. Usually these instruments are tests of various kinds. It is nearly impossible to construct a representative test which does equal justice to all the testees and their various complexities at any time.
5. Measuring Scale: Unlike in physical science where zero on a scale means an absence of what is being measured, the zero score on a test does not necessarily indicate complete lack of the attribute being measured. This therefore lends the results of educational measurement to all sorts of interpretation, especially when they get into the hands of non-professionals.

Fairness of Objective Tests

The National Centre for Fair and Open Testing (2006) in the United Kingdom opined that test-makers often promote multiple-choice tests as "objective." This is because there is no human judgement in the scoring, which usually is done by machine. However, humans decide what questions to ask, how to phrase the questions, and what distracters to use. All these are subjective decisions that can be biased in ways that unfairly reward or harm some test-takers. Therefore, multiple-choice tests are not really objective. Multiple-choice items are best used for checking whether students have learned facts and routine procedures that have one clearly correct answer. However, an item may have two reasonable answer options. Therefore, test directions usually ask test-takers to select the "best" answer. If, on a reading test, a student selected a somewhat plausible answer, does it mean that she cannot read, or that she does not see things exactly the way the test maker does?

It is possible to get multiple-choice items correct without knowing much or doing any real thinking. According to the National Centre for Fair and Open Testing, because the answers are in front of the student, some call these tests "multiple-guess". Multiple-choice items can be easier than open-ended questions asking the same thing. This is because it is harder to recall an answer than to recognize it. Test-wise students know that it is sometimes easier to work backwards from the answer options, looking for the one that best fits. It also is possible to choose the "right" answer for the wrong reason or to simply make a lucky guess.

The Centre warns that relying on multiple-choice tests as a primary method of assessment is educationally dangerous and gave the following reasons:

(1) Because of cultural assumptions and biases, the tests may be inaccurate. (Of course, other kinds of assessments also can be biased.) Assuming the test is accurate because of its supposedly "objective" format, it may still lead to making bad decisions about how best to teach a student.

(2) Students may recognize or know facts or procedures well enough to score high on the test, but not be able to think about the subject or apply knowledge, even though being able to think and apply is essential to "knowing" any subject. Therefore, the conclusion or inference that a student "knows" history or science because she got a high score on a multiple-choice test may be false.

(3) What is easily measurable may not be as important as what is not measurable or is more difficult to measure. A major danger with high stakes multiple-choice and short-answer tests — tests that have a major impact on curriculum and instruction — is that only things that are easily measured are taught.

(4) Since the questions usually must be answered quickly and have only one correct answer, students learn that problems for which a single answer cannot be chosen quickly are not important.

(5) When schools view multiple-choice tests as important, they often narrow their curriculum to cover only what is on the examinations. For example, to prepare for multiple-choice tests, curriculum may focus on memorizing definitions and recognizing (naming) concepts. This will not lead students to understand important scientific principles, grasp how science is done, and think about how science affects their lives.

(6) When narrow tests define important learning, instruction often gets reduced to "drill and kill" - lots of practice on questions that look just like the test. In this case, students often get no chance to read real

> books, to ask their own questions, to have discussions, to challenge texts, to conduct experiments, to write extended papers, to explore new ideas - that is, to think about and really learn a subject.

The National Centre for Fair and Open Testing (2006) further stated that the decision to use multiple-choice tests or include multiple-choice items in a test should be based on what the purpose of the test is and the uses that will be made of its results. If the purpose is only to check on factual and procedural knowledge, if the test will not have a major effect on overall curriculum and instruction, and if conclusions about what students know in a subject will not be reduced to what the test measures, then a multiple-choice test might be somewhat helpful - provided it is unbiased, well-written, and related to the curriculum. If they substantially control curriculum or instruction, or are the basis of major conclusions that are reported to the public (e.g., how well students read or know mathematics), or are used to make important decisions about students, then multiple-choice tests are quite dangerous.

According to Cacko (1993) the assumption that candidates' poor performance in a test is often a reflection of lack of learning, need not be always correct. He argued that if a test does not assess the objectives of learning and/or is faulty in structure, clarity, complexity, level of order (i.e. arrangement of items), then the low score obtained may not be due to poor learning but to some other factors.

Effect of Item Order

In a study on arrangement of test items, Anastasi (1976) claimed that in a test that is timed, items appearing late would be answered correctly by a relatively small percentage of the total sample. According to her, the reason is that only a few candidates would have enough time to reach such items coming up at the end of a test.

Also Shepard (1994) asserted that tiny changes in test format (or arrangement) can make a large difference in student performance. For example, a high proportion of students may be able to add numbers when they are presented in vertical format, but many will be unable to do the same problems presented horizontally.

With regard to the impact of level of order or arrangement of items in a multiple-choice test on examinees' performance, MacNicol (1956) investigated the effects of changing an "easy-to-hard" arrangement to either;

(i) hard-to-easy; or

(ii) a random arrangement.

He found out that the hard-to-easy arrangement was significantly more difficult than the original easy-to-hard order while the random arrangement was not significantly different.

In a related study, Soyemi (1980) also found no significant differences between

(i) easy-to-hard and hard-to-easy arrangement;

(ii) easy-to-hard and random order; and

(iii) hard-to-easy and random order.

Recognizing the importance of appropriate arrangement of test items, Sax and

Cromack (1966) and Ahuman and Clock (1971) have advised that tests should be constructed in an easy-to-hard item-difficulty sequence.

On choice and arrangement of items for a multiple-choice test, Painter (1989) recommended the avoidance of using interrelated items. He asseted that if getting the correct answer to one item depends on correctly answering a previous item then undue weight is given to the first item. It is okay to refer to the same stimulus material, as long as getting the correct answer to one item does not depend on the correct answer to another item. On the difficulty level of an item in a multiple-choice test, Painter says that it will depend on two things:

1. The thought process called for by the stem. This can range from recalling factual information to evaluating new information;

2. Similarity of the options. Items with options with a high degree of conceptual similarity require greater understanding in order to identify a correct response.

The Research Division of WAEC, Lagos (1993) investigated the effect of item position on performance in multiple-choice objective tests at the level of the West African Senior School Certificate Examination (WASSCE). The subjects used were Agricultural Science, Biology, Economics, English Language and Mathematics. A forty-five multiple-choice objective test was developed in each of the selected subjects. The test consisted of three sub-tests indicated as sections in the question paper with each section having fifteen items. The items were arranged to reflect the following order:

(i) Easy-to-Hard;

(ii) Hard-to-Easy;

(iii) Random order

The research sample consisted of one thousand, one hundred and twelve (1,112) SSS.2 students randomly selected from 35 schools in eight states in Nigeria. The data collected were subjected to statistical analysis using ANOVA and t-test to determine if any significant differences existed in the performance of students in the three arrangements of items and the pattern of performance among the five subjects. As expected there were significant differences in the performance of students. Easy-to-Hard arrangement of items was the most effective in enhancing the performance of students in English Language and Biology. Surprisingly, the Hard-to-Easy arrangement was significantly the most effective in Mathematics and Economics while the Random Order was effective in Agricultural Science, among the three arrangements.

It was discovered from the mean score perspective that the Agricultural Science performance could be by chance. Analysis of the means of performance between pairs of the arrangement in each subject revealed that whereas the Hard-to-Easy enhanced the best performance in two subjects (Mathematics and Economics), the Easy-to-Hard favoured one subject (English Language), the Random order of test items did not significantly improve the students' performance in any of the subjects.

In a follow-up research, the Research Division of WAEC, Lagos (1995), investigated the effect of sex, ability group and school type on students' performance in the three formats of arranging multiple-choice test items were investigated. In the findings, it was reported that the general performance of students in the three arrangements was independent of sex except Hard-to-Easy in Mathematics and Random Order in English Language. This indicates that irrespective of the arrangement adopted in Economics, Biology and Agricultural Science, the pattern of performance, was not susceptible to differences in sex. However, the Hard-to-Easy arrangement in Mathematics disadvantaged the female candidates while the Random-Order in English Language seem to have reverse effect though it is assumed that the occurrence is by chance.

The research discovered that the performance of the different ability groups followed the same pattern irrespective of the arrangement adopted. This implies that the high ability group will still obtain the best performance in any arrangement while the low ability group will perform poorly notwithstanding the arrangement. The effect of the category of school was not significant. At the end, the researchers concluded that the way items are arranged in multiple-choice objective tests would have a significant effect on the performance of candidates. However, the effect is dependent on the subject involved.

On the contrary, empirical studies discussed by Gerow (1980) on the sequencing of questions for university students have all failed to indicate any difference between random ordering of questions and questions organized by the order in which it was taught. Gerow presented further empirical evidence to the effect that arranging the items in order of difficulty also has no effect provided that there is enough time for examinees to complete the test. A further study by Allison (1984) confirmed that even for sixth grade students there was no effect on performance by ordering the items according to difficulty, provided that there was enough time to complete the test.

On the notion that starting students off with easy question items to build confidence improves test scores, Skinner (1999) presented results suggesting that students may actually perform better if tests begin with difficult questions and students are given immediate feedback.

Effect of Textbook Content Order

Hopkins (1998) suggested that it is necessary to avoid arranging items in the order in which they were presented in the textbook in order to achieve the logical validity test. However, in an experimental investigation to determine the effect of randomization of questions and possible answers on the performance of students, McLeod (2003) arranged items in different orders including the order in which they were presented in the textbook, to develop four treatment combinations of a test. These include:

(i) completely randomized combination;

(ii) completely ordered combination;

(iii) partially randomized combination;

(iv) partially ordered combination.

The combinations were each identified by a three digit code and administered to 442 students in a university with neither the student nor instructor knowing which type of combination was used for a particular student. Analysis of the results using the mean and the standard deviation showed that there was no difference in scores between the treatment combinations. They therefore concluded in confirmation of the earlier discovery by Gerow (1980) and

Allison (1984), that there was no empirical evidence to suggest adverse effect on the performance of examinees when questions are randomized or ordered.

According to Balch (1989) students score higher on multiple-choice examinations when the questions are presented in the same order that the material was presented in lecture and text as opposed to when questions are randomly grouped by chapter or in completely random order. Providing advantage to one group of students who take the sequential versus a random test question order examination is problematic and unfair. Balch suggested that sequentially ordered examinations provide retrieval cues which may help with memory recalls, consistent with encoding specificity. The context of surrounding information used in encoding is utilized in information retrieval, and the sequential test question order provides a situation where context of encoding and retrieval are similar. In addition, Balch found that there was no significant difference in completion times between these versions of the examination.

Other researchers have challenged this rationale and these findings. Neely, Sprigston and McCann (1994) conducted a three study follow-up to Balch in which student performance on sequential and random order multiple-choice question examination in an introductory psychology class were compared and the influence of test anxiety was also considered. The results of the three studies showed no significant difference between the sequential and random order multiple-choice question tests. However, the researchers did report a significant interaction such that high-anxiety students performed "somewhat better" on the sequential question order test and low-anxiety students performed "substantially better" on the random question order test.

Similarly, Peters and Messier (1970) also found no differences in performance on sequential versus random question order multiple-choice tests in a class of graduate students studying research methods, and those students who reported high level of anxiety performed worse on the random question order test compared to the sequential question order test.

Perlini, Lind, and Mumbo (1998) further investigated the effects of items order and item difficulty on test performance of undergraduates in four studies. In Experiment 1, the investigators found no advantage in student performance on sequenced chapter-order multiple-choice question tests over random or reverse question order test. In Experiment 2, researchers varied chapter order and within chapter question order, but again found no performance difference between conditions. Item and chapter question arrangements were found to have little or no effect on test performance. Perlini et al (1998) also arranged test questions with respect to item difficulty in their fourth study: easy-to-hard, hard-to-easy or random. Again, there was no significant difference between difficulty arrangements.

In a similar investigation using first year university students, Laffittee (1984) created four versions of an introductory psychology multiple-choice test: easy to difficult by topic, easy to difficult across chapters, randomly within chapters and randomly across chapters. Laffittee reported that presentation order had no effect on achievement test scores or student perception of test difficulty.

Taub and Bell (1975) considered the positioning of test questions and concluded that a truly random arrangement of questions results in lower examination scores than does ordering the questions to follow lectures and textbook assignments in the university.

Impact of Use of Alternative Tests

In a research conducted in 1993 by the Research Division of the Lagos Office of WAEC Headquarters, the impact of using two alternate forms or parallel tests on candidates' performance was investigated. The investigation was conducted in English Language and Mathematics. Items used were selected from past West African Senior School Certificate Examination (WASSCE) papers obtained from standardized questions culled from Nigeria and The Gambia papers. The alternate forms of the English Language paper consisted of 50 multiple-choice items and examinees were expected to answer all in 30 minutes. The Mathematics consisted of 20 multiple-choice items each and examinees were to answer all in 45 minutes. 50 items were not used in the Mathematics so as to reduce fatigue on the part of examinees while responding to the items. The data generated were subjected to Pearson Product Moment Correlation and paired t-test statistical analysis. The analysis revealed that there was no significant difference between the pairs of students' scores in both English Language and Mathematics.

Subscribing to importance of difference in time required to complete tests, student perceptions of test difficulty, understanding of course material and anxiety levels on the outcome of alternative order versions of examinations, and their concern for fairness in examination performance drove Pettijohn and Sacco (2001) to design a research to investigate the effects of multiple-choice test question order on student performance and completion time. They initially predicted that there would be no significant differences in student performance and completion times between sequential (S), random (RA), and reverse (RE) question order examinations. To them, investigating the reverse (RE) test question order would provide the opportunity to determine the effect of testing the most recent information learned first and working backwards in a reverse sequential order. In the investigations, the researchers randomly assigned an equal number of participants to the sequential (S), random (RA) and reverse (RE) conditions. Once all the data were collected, Pettijohn and Sacco calculated mean examination scores for the S, RE and RA question order group for each examination. They conducted a repeated measure ANOVA (test question order conditions: S, RE, RA) for the dependent variable examination score. They discovered that participants performed similarly on the different versions of the examinations and individual comparisons revealed no significant differences between test question order conditions on test scores. Again they found out that test condition order and sex did not interact with test performance. Pettijohn and Sacco finally concluded that there were no statistically significant differences between the test performance across the sequential, reverse and random ordered multiple-choice tests suggesting that multiple-choice question order does not significantly influence test performance or test completion times.

Tauber (1984) suggested an alternative to Alternative Test Forms as a way of reducing cheating on multiple-choice examinations. He argued that literature on cheating suggests the use of alternative test forms of examinations as a way to curb cheating. Unfortunately, he continued, the literature does not highlight the problems associated with preparing and using alternative test forms. He asserted that these problems make their use impractical. Introducing his concept, Tauber wrote that with the traditional method of alternative test forms, two or more separate examinations must be prepared. They must be organized, typed, proofread and coded. He said, this is a waste of time, energy and secretarial resources.

In his alternative method, Tauber (1984) used the same version of the examination for every student but with a multiple numbering system. Each question on the examination has two (or more) numbers next to it in the left-hand margin where ordinarily there would be just one number. For example, the first question on a hypothetical 50 item multiple-choice examination is number “4/7”, the second is “6/2” and so on. Some students will use the first of each pair of numbers to determine where to place their responses on a separate answer sheet, while other students will use the second pair of numbers. The first question in the examination numbered “4/7”, would require some students to place their response at spot #4 on the answer sheet, while other students would record their response to this same question at spot #7. He argued that as long students place their answers at different spots on the same answer sheet, the net effect is the same as with the traditional, more time consuming, alternative test form examinations and it reduces the incidence of cheating. How are students told which number in the pair (trio, etc) to use? To answer this, Tauber said several techniques can be used, but all include a method of “keying off” some information which is normally known to an individual student. For instance, if the last digit in the student’s social security number is ODD, the student uses the number to the left. If the last digit is EVEN, he/she uses the number to the right. He said, apart from social security number, other information like birth date, student personal identification number, home telephone number, etc. could be used. As a little insurance, students in every other row (or males versus females) are instructed to start the examination from the back and work forward. Tauber warns that when numbering the questions care must be taken to use all the numbers; while at the same time, no duplication must occur.

To set the stage for acceptance of the multiple-numbered examinations, Tauber (1984) recommended that students must be briefed and told how they will place their answers at different spots on answer sheets. After the application of the method in an Introductory Economics and then Educational Psychology examination, feedback were collected from students on whether or not the multiple-numbering system caused any interference in their test taking. Out of the 150 and 40 students respectively, all resoundingly said it did not affect their test taking in any way. They however, indicated that it took a little extra time to locate the proper spot on the answer sheet to record their responses.

Impact of Changing Responses Order

Gohmann and Spector (1989) wrote that the sequence of the examination items does not have statistically significant effect on the mean level of performance of students. However, Carlson and Ostrosky (1992) are of the contrary view that an examination in which items are randomly ordered might be more confusing and, as such, might have adverse effects that could result in a reduced level of performance. They declared that this type of effect is undesirable and therefore conducted a study to investigate it. In the study two forms of an instrument were used. In Form A the individual examination items were content ordered, and the items corresponding to each topic on the examination were ordered to level of difficulty. In Form B, examination items were scrambled with respect to content and difficulty. The correct responses were distributed uniformly across the four possible choices. For the trial test, Forms A and B were distributed to students randomly. The resulting data were analyzed to determine whether item sequencing affects the average level of student performance, the variance of examination score and the distribution of scores on the examination, among others. Based on their analysis, Carlson and Ostrosky were able to reject the null hypothesis that the mean score of the content-ordered examination is less, or equal to, the mean score of the scrambled examination. They therefore concluded that considering the fact that the mean score of the content-ordered examination exceeded the mean score of the scrambled examination provides evidence that, in fact, the mean level of performance may be higher on the content-ordered examination than the scrambled examination. They claimed that sequencing would affect the performance of students. In particular, students taking the ordered form of an examination may benefit in the form of higher examination scores. They therefore recommended that if alternative forms of an examination must be used, it may be preferable to hold the order of questions constant and instead scramble the order of the responses to each question.

On this issue, Jessell and Sullins (1975) remarked that "with the exception of the occasional 'pattern sleuth' or 'pattern maker,' it would appear that examinees pay less heed to response patterning than might be supposed" (p.48).

Bresnock, Graves and White (1989) report that two types of approaches to combat cheating are commonly used: rearranged questions and rearranged responses. If questions are ordered to follow the textbook and lecture sequence on one test but are jumbled on the other, will the grade distribution for these tests be different? Do jumbled arrangements of the responses within the test questions lead to significant differences in performance across such tests? The trio set themselves the task of investigating these issues. In their literature survey, they discovered that as far back as 1945, efforts had been made to determine the impact of response position on performance. However, conclusions regarding randomization to remove personal position preferences were not fully satisfying. To them, it was not clear that an elaborate system of assuring random placement of correct answers significantly affects examination validity.

In their experimental setting, Bresnock et al (1989) used three examinations consisting of multiple-choice questions with four-lettered responses which were administered to 301, 295 and 305 students in an undergraduate Principles of Economics class. Two different tests were administered to students present. The first consisted of two jumbled versions of the same test, without manipulation of within-question response patterns. For both formats of the tests, the correct response distribution was typical: 18.9% at option A, 29.7% at option B, 32.4% at option C and 21.6% at option D. This test was set up to compare how well students perform when questions are jumbled and when they are presented in the order of classroom lectures. To test whether an abnormal response distribution would contribute to differences in test scores the trio developed a second exam with two formats. Format A contained a higher percentage of correct A and D responses i.e. 35% at option A, 18% option B, 13% at option C, and 33% at option D. The response distribution for Format B was 27% option A, 27% option B, 21% option C, and 24 % option D. The intention behind this distribution was to discover whether failing to hide the correct answer alters examination performance.

In a third examination, a test with an abnormally high percentage (36%) of correct A answers was compared with another test with an abnormally large percentage (35%) of correct D answers. Using the empirical results, a comparison of average test scores for each examination revealed that for the third examination only, there was a significant difference in scores on the A and B test formats.

Bresnock et al (1989) applied a chi-square test to the scores distribution to determine whether altering correct question or response format produces significantly different patterns of test results. From this frequency information, they discovered that test format was not significant in explaining differences in performance on the first and second tests. They reported that their findings from the first test in which questions were scrambled, agreed with the Monk and Stalling (1970) study that non-systematic arrangements of both test formats (when questions are jumbled and when they are presented in the order of classroom lectures) did not generate significant differences in performance. Thus, they concluded that tests of equivalent difficulty may be constructed by jumbling test formats. They reported that examining the results of the second test in which responses were scrambled, suggests that such scrambling does not affect examination performance. The disproportionately large number of non-hidden correct responses did not appear to be discriminatory across the groups. They explained that in an arrangement of this kind, students may disproportionately select a large number of correct A responses because they appear first and may systematically under select correct D because they appear last. One reason for this occurrence is that a large number of correct A responses saves time for those with that test format.

Finally, Bresnock et al (1989) concluded that the results of their third examination support the assertion that certain test-response formats will generate significant difference in performance. Again, changing the response patterns appears to alter significantly the apparent degree of difficulty under certain types of response alteration. However, changing item position does not alter performance significantly. They remarked that constructing equivalent tests that examine a student's command of the subject depends on careful attention to test design as well as to test content.

State of the Art

In the literature survey, researchers are not unanimous in their findings as to whether or not altering item position in a multiple-choice test would affect performance adversely. Some like Anastasi (1976) argued that different arrangement of items will affect performance. This view is supported by Cacko (1993). Researchers in the Research Division, WAEC, Lagos (1993) discovered that different arrangements of items could affect performance adversely or positively depending on the subject in question. Others like Gerow (1980) and Allison (1984) found no difference in performance when items were arranged according to a certain order of difficulty or randomly.

However, those who found no significant change in performance when item position was altered are slightly more than their counterparts who assert that performance will be affected by difference in item order.

This gives some idea of what to expect during the study which is directed at obtaining justification or otherwise for the development and use of different forms of a test to curb examination malpractice at the BECE level. The fact that there is no unanimity among researchers from the literature review indicate that there is a problem and provides motivation for the study.

CHAPTER THREE

METHODOLOGY

Introduction

The study set out to investigate the impact of item position in multiple-choice test on student performance at the Basic Education Certificate Examination (BECE) level.

In this chapter, the following sub-topics are discussed:

1. Research design adopted,
2. Population targeted,
3. The sample used,
4. The instrument employed,
5. Data collection method,
6. Data analysis.

Research Design

A quasi-experimental design was adopted. The term quasi-experimental design was first introduced by Campbell and Stanley (1963). It is a design which looks like an experimental design but lacks the random assignment. Quasi-experimental design involves selecting groups, upon which a variable is tested, without any random pre-selection processes. After this selection, the experiment proceeds in a very similar way like any other experiment, with a variable being compared between different groups.

Regarding the advantages of the quasi-experimental design, the Wikipedia – Free Encyclopedia online (2011) states that since quasi-experimental designs are used without randomization, they are typically easier to set up than true experimental designs, which require random assignment of subjects. Because randomization is absent, some knowledge about the data can be approximated. Additionally, utilizing quasi-experimental designs minimizes threats to external validity as natural environments do not suffer the same problems of artificiality as compared to a well-controlled laboratory setting. Since quasi-experiments are natural experiments, findings in one may be applied to other subjects and settings, allowing for some generalizations to be made about populations. Also, this experimentation method is efficient in researches that involve longer time periods which can be followed up in different environments. The method can be very useful in generating results for general trends, says, Shuttleworth and Martyn (2008).

The Wikipedia Encyclopedia (2011) warns about some limitations. It wrote that the control allowed through the manipulation of the quasi-independent variable can lead to unnatural circumstances, although the dangers of artificiality are considerably less relative to true experiments. Also the lack of random assignment in the quasi-experimental design method may pose many challenges for the investigator in terms of internal validity. This deficiency in randomization makes it harder to rule out confounding variables and introduces new threats to internal validity. Because randomization is absent, conclusions of causal relationships are difficult to determine due to a variety of extraneous and confounding variables that exist in a social environment. Moreover, even if these threats to internal validity are assessed, causation still cannot be fully established because the experimenter does not have total control over extraneous variables.

The method was adopted because it was impossible to randomly select testees from all over Ghana for the study. It was also chosen because it affords the opportunity to generalize over the population of BECE candidates specified below.

The variable in this design is the performance of the testees. The control group are those who took the Random option and the treatment or programme groups are those who took the Easy-to-Hard and Hard-to-easy options. The treatment is the re-arrangement of the Random order into the Easy-to-Hard and Hard-to-Easy alternatives.

Population

The target population for the study comprised all JSS 3 students who took the April 2006 Basic Education Certificate Examination (BECE). The size of this population was 308,325. Subjects available for the candidates included English Language, Mathematics and Science which were among the seven (7) compulsory subjects offered in the examination. The examination was taken in 1,121 centres spread all over the country, both rural and urban. It was available to a total of 8,079 Junior Secondary Schools consisting of schools from both private and public sectors.

Sample and Sampling Procedure

The sample consisted of 810 JSS 3 students who were selected from 12 different schools. Initially it was intended to have a sample size of 1200 i.e.

100 students from each school. This sample size was chosen to provide amply for all the treatments which were to be offered and provide sufficiently for extrapolation or generalization. But in some of the selected schools, enrolment was low and therefore could not supply the required number of students for the tests. The composition of the sample according to paper is as shown in Table 1. The low figures for some of the treatments in the table were as a result of lose of data through computer virus infection in the course of processing the data.

Table 1: Composition of Sample

Subject	Number of Pupils
English Language (Random order)	75
English Language (Easy-to-Hard order)	158
English Language (Hard-to-Easy order)	39
Mathematics (Random order)	249
Mathematics (Easy-to-Hard order)	291
Mathematics (Hard-to-Easy order)	64
Science (Random order)	134
Science (Easy-to-Hard order)	35
Science (Hard-to-Easy order)	295
Total	**810**

The figures in the table were obtained by adding up the candidates who took the treatment indicated. The schools were selected to represent the following categories of institutions:

(i) public school in an urban area;

(ii) private school in an urban area;

(iii) public school in a rural area;

(iv) private school in a rural area.

The list captures all the categories of schools which took the BECE examination. The urban schools were purposively selected from Accra for convenience while the rural schools were selected in the same manner from Ga-rural which shares borders with the Eastern Region and from Kasoa and its environs in the Central Region. In all, six schools were selected to represent each of the urban and rural settings.

Instrument

A multiple-choice test consisting of forty items was developed in each of English Language, Mathematics and Science. The assistance of professional test developers was sought for the crafting of the items for all the subjects. The items were developed to cover all areas in the WAEC syllabus for the subjects involved. After construction, the items were submitted to experts in the various subject fields for a second look. This was to ensure that the items were standardized and could go for any WAEC examination at the BECE level. Furthermore, the engagement of the subject experts was done to safeguard the validity and internal reliability of the items.

The English Language papers consisted of five sections: A, B, C, D and E. Section A had two comprehension passages followed by six questions to test understanding. Section C tested synonyms while Section D was on antonyms. Section E tested their ability to complete a sentence with the right word chosen from an option of four. Each section had its own rubrics stated clearly at the beginning of the section. The Mathematics and Science papers both had one section each with questions from all areas of the syllabus. The structure of all the papers was in consonance with the format and item specification used by WAEC. In all the papers, examinees were expected to answer all 40 questions and they had 45 minutes to do so. Again, this was in accordance with the usual time allocated by WAEC for such tests.

The first type of papers to be crafted was the Random ordered ones. They were then piloted in a conveniently chosen public school in Adabraka, a suburb of Accra. The cohort that took the trial test consisted of 82 pupils who were drawn from two streams of JSS 3. The responses were captured on scannable objective answer sheets and machine scored. They were then subjected to item analysis. From the results of the analysis, items which needed modification were re-fixed, although these were very few. The difficulty levels of the items were also determined.

With the results from the item analysis, the items were re-arranged, one from Easy to Hard, and another from Hard to Easy, while keeping the Random order intact. Arranging the items in order of difficulty was straight forward for Mathematics and Science since they both had only one section each. However, for English Language the difficulty order arrangement was on sectional basis. This was because rubrics for one section could not be applicable to another.

Samples of the RDM arrangement of each of the subjects are included as Appendices 1 – 3. Samples of the HTE and ETH arrangements are not added since they contain the same items and their addition will constitute a mere repetition.

Data Collection Procedure

The test instruments were administered to JSS 3 candidates in twelve different schools, in the last week of March 2006, four weeks before the final examination was taken at the national level. The tests took two days to administer with the assistance of professional test administrators. The conduct of the tests was strictly according to standards of the West African Examinations Council. Each school took only one of the options of a subject but participated in all subjects. Whether a cohort took the Easy-to-Hard (ETH), Hard-to-Easy (HTE) or Random (RDM) arrangement of a particular subject, was determined by a time-table to ensure fair participation in all the treatments. The drawing of the time table followed the pattern in Table. 2 below:

Table 2: Time Table Pattern

School	English	Mathematics	Science
School 1	RDM	ETH	HTE
School 2	ETH	HTE	RDM
School 3	HTE	RDM	ETH

The responses to the items were captured by encircling the correct option on the question paper. The responses were then keyed into the computer in a format compatible with the Iteman software for scoring. Thereafter, the scores were transposed into the SPSS software for analysis.

Data Analysis

The scores were first standardized for the following reasons:

1. To make up for variances in the sample sizes from the different schools;
2. To cater for variance in conditions and facilities from one school to the other;
3. To make up for missing values for candidates who may not have been able to complete all the items in a test;
4. To provide for other unknown factors which may not have been noticed.

The data collected for English Language were subjected to one-way analysis of variance (ANOVA) at $p \leq 0.05$ to determine if any significant differences existed in the performance of students. The independent variable was Item Order having three levels which are the Random order, Easy-to-Hard order and Hard-to-Easy order. The dependent variable was the scores of the tests. A preliminary test for homogeneity of variance was performed to ascertain if population variances were equal. Post Hoc test was performed to determine the test order in which the candidates excelled in performance.

Similarly, the data collected for Mathematics were subjected to statistical analysis using ANOVA at $p \leq 0.05$ to determine if any significant differences existed in the performance of students using Item Order as the independent variable with three levels which are the Random order, Easy-to-Hard order and Hard-to-Easy order. The dependent variable was the scores of the tests. A preliminary test for homogeneity of variance was performed to ascertain if population variances were equal. Post Hoc test was performed to determine the test order in which the candidates excelled in performance.

Again, the data collected for Science were subjected to statistical analysis using ANOVA at $p \leq 0.05$ to determine if any significant differences existed in the performance of students using Item Order as the independent variable with three levels which are the Random order, Easy-to-Hard order and Hard-to-Easy order. The dependent variable was the scores of the tests. A preliminary test for homogeneity of variance was performed to ascertain if population variances were equal. Post Hoc test was performed to determine the test order in which the candidates excelled in performance.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

In this chapter, the results of the research are presented and discussed.

The study sought to find answers to the following questions:

1. What would be the effect of a change in item order on candidates' performance in English Language at the BECE level?
2. What would be the effect of a change in item order on candidates' performance in Mathematics at the BECE level?
3. What would be the effect of a change in item order on candidates' performance in Science at the BECE level?

Analysis and Findings

Research Question 1:

> What would be the effect of a change in item order on candidate's performance in English Language at the BECE level?

The data collected for English Language were subjected to one-way analysis of variance (ANOVA) at $p \leq 0.05$ to determine if any significant differences existed in the performance of students. The independent variable was Item Order having three levels which are the Random order (RDM), Easy-to-Hard order (ETH) and Hard-to-Easy order (HTE). The dependent variable was the scores of the tests. A preliminary test for homogeneity of variance was performed to ascertain if population variances were equal. Post Hoc test was performed using Dunnett C to determine the relationship between pairs of test orders as they affect candidates' performance.

The results for the analysis of scores for English Language are tabulated below. Table 3 gives information on the mean and standard deviation values for the three levels of the item order.

Table 3: Descriptive Statistics for Performance in English Language

Order	N	Mean	Std. Deviation
Random (RDM)	75	32.97	3.093
Easy-to-Hard (ETH)	158	18.70	6.420
Hard-to-Easy (HTE)	39	19.77	6.737
Total	**272**	**22.79**	**8.521**

Table 4 gives the results of the ANOVA for English Language.

Table 4: One-way ANOVA for Performance in English Language

	Sum of Squares	df	Mean Square	F	Sig
Between Groups	10773	2	5386.6	162.7	.000
Within groups	8904	269	33.1		
Total	**19677**	**271**			

The means and standard deviations of the test orders are shown in Table 3. The large difference between the mean of the RDM on one side, and those of the ETH and HTE was rather unexpected. The preliminary homogeneity test at $p < .05$ gave a significant value ($p< .001$) showing that population variances were not equal. In Table 4, the results of the one-way ANOVA for English Language are presented. From the table, at $p\leq .05$, an F value of 162.7 and significant value of $p< .001$ were realized indicating that there was significant difference in performance when item order is altered in English Language. The results of a Dunnett C multiple comparisons Post Hoc test indicated from the mean difference in performance at the .05 level, that for English Language:

(i) there was significant difference in performance between the Random and Easy-to-Hard treatments;

(ii) there was significant difference in performance between the Random and Hard-to-Easy treatments;

(iii) there was no significant difference in performance between the Easy-to-Hard and Hard-to-Easy treatments.

From the means, it could be deduced that the candidates performed best in the Random order of the English Language test.

Thus an answer to Research Question 1 will be that the effect of a change in item order on candidates' performance at the BECE level was significant with regard to English Language.

Research Question 2:

> What would be the effect of a change in item order on candidate's performance in Mathematics at the BECE level?

To answer this question, the data collected for Mathematics were subjected to statistical analysis using one-way ANOVA at $p \leq .05$ to determine if any significant differences existed in the performance of students using Item Order as the independent variable with three levels which are the Random order (RDM), Easy-to-Hard order (ETH) and Hard-to-Easy order (HTE). The dependent variable was the scores of the tests. A preliminary test for homogeneity of variance was performed to ascertain if population variances were equal. Post Hoc test was performed using Dunnett C to determine the relationship between pairs of test orders as they affect candidates' performance.

The results for the analysis of scores for Mathematics are tabulated below. Table 5 gives information on the mean and standard deviation values for the three levels of the item order.

Table 5: Descriptive Statistics for Performance in Mathematics

Order	N	Mean	Std. Deviation
Random	249	15.21	7.950
Easy-to-Hard	291	18.32	8.077
Hard-to-Easy	64	10.44	3.854
Total	**604**	**16.20**	**8.063**

Table 6 gives the results of the ANOVA for Mathematics.

Table 6: One-way ANOVA for Performance in Mathematics

	Sum of Squares	df	Mean Square	F	Sig
Between Groups	3677	2	1838.6	32.1	.000
Within groups	35528	601	59.1		
Total	**39205**	**603**			

The means and standard deviations of the various test orders are shown in Table 5. The preliminary homogeneity test at $p < .05$ gave a signifcant value ($p< .001$) indicating population variances are not equal. In Table 6, the results of the one-way ANOVA for Mathematics are presented. From the table, an F value of 32.1 and significant value of .000 at $p \leq .05$ was obtained indicating that the difference in performance in Mathematics when item order is altered was significant.

The results of a Dunnett C multiple comparisons Post Hoc test indicated from the mean difference in performance at the 0.05 level, that for Mathematics:

(i) there was a significant difference in performance between the Random and Easy-to-Hard treatments;

(ii) there was a significant difference in performance between the Random and Hard-to-Easy treatments;

(iii) there was a significant difference in performance between the Easy-to-Hard and Hard-to-Easy treatments.

From the means, it could be deduced that the candidates performed best in the Easy-to-Hard order of the Mathematics test.

Thus, an answer to Research Question 2 is that the effect of a change in item order on candidates' performance at the BECE level was significant with regard to Mathematics.

Research Question 3:

> What would be the effect of a change in item order on candidates' performance in Science at the BECE level?

To answer this question, the data collected for Science were subjected to statistical analysis using ANOVA at $p \leq .05$ to determine if any significant differences existed in the performance of students using Item Order as the independent variable with three levels which are the Random order (RDM), Easy-to-Hard order (ETH) and Hard-to-Easy order (HTE). The dependent variable was the scores of the tests. A preliminary test for homogeneity of variance was performed to ascertain if population variances are equal. Post Hoc test was performed using Dunnett C to determine the relationship between pairs of test orders as they affect candidates' performance. The results for the analysis of scores for Science are tabulated below. Table 7 gives information on the mean and standard deviation values for the three levels of the item order.

Table 7: Descriptive Statistics for Performance in Science

Order	N	Mean	Std. Deviation
Random	134	15.36	5.206
Easy-to-Hard	35	13.49	4.293
Hard-to-Easy	295	16.69	7.831
Total	**464**	**16.06**	**6.998**

Table 8 gives the results of the ANOVA for Science.

Table 8: One-way ANOVA for Performance in Science

	Sum of Squares	df	Mean Square	F	Sig
Between Groups	414	2	207.2	4.3	.014
Within groups	22263	461	48.3		
Total	**22677**	**463**			

In Table 7 the means and standard deviations of the various test orders are shown. The preliminary homogeneity test at $p < .05$ gave a significant value ($p < .001$) showing that population variances are not equal. In Table 8, the results of the One-Way ANOVA for Science are presented. From the table, an F value of 4.3 and significant value of 0.014 at $p \leq .05$ was obtained indicating that there was significant difference in performance in Science when item order is altered.

The results of a Dunnett C multiple comparisons Post Hoc test indicated from the mean difference in performance at the .05 level, that for Science:

(i) there was no significant difference in performance between the Random and Easy-to-Hard treatments;

(ii) there was no significant difference in performance between the Random and Hard-to-Easy treatments;

(iii) there was significant difference in performance between the Easy-to-Hard and Hard-to-Easy treatments.

From the means, it could be deduced that the candidates performed best in the Hard-to-Easy order of the Science test. However, generally the performance across the treatments in Science was quite close.

Thus an answer to Research Question 3 is that the effect of a change in item order on candidates' performance at the BECE level is significant with regard to Science.

Discussion

These results generally disagree with the findings of researchers like Gerow (1980) who performed empirical studies on sequencing of questions for university students and Allison (1984) who found no difference in performance when items were arranged according to a certain order of difficulty or randomly for sixth grade students. The levels of students used by these researchers were higher than that of the BECE students used for this research. However, the results are still worth comparing.

In another study using Senior Secondary School Certificate Examination (SSSCE) involving students of Economics, Soyemi (1980) also found no significant differences between

(i) easy-to-hard and hard-to-easy arrangement;

(ii) easy-to-hard and random order arrangement;

(iii) hard-to-easy and random order arrangement

and therefore is not supported by the findings of this research. His findings are however, largely supported by the findings of this study in respect of Science.

Perlini, Lind and Zumbo (1998) who arranged test questions with respect to item difficulty in their fourth study: easy-to-hard, hard-to-easy or random also found that there was no significant difference between difficulty arrangements among university undergraduates studying psychology. Laffittee (1984) who also performed his study using first year university students reported that presentation order had no effect on achievement test scores. These researchers will also disagree with the finding of this study.

The differences in the findings of these researchers and the findings of the present study may be due to the vast difference in the levels of education in which the researches were conducted. Whilst they used university and high school students, this study used junior high students.

However, the studies cited above are supported by the findings for the ETH and HTE for English Language; the RDM and ETH as well as the RDM and HTE for Science where no significant difference in performance was recorded.

The results agree with educational measurement experts like Shepard (1994) who assert that tiny changes in test format (or arrangement) can make a large difference in students' performance. The results also agree in part with the findings of MacNicol (1956) who performed a research using similar treatment and found that out that the hard-to-easy arrangement was significantly more difficult than the original easy-to-hard order while the random arrangement was not significantly different. Recognizing the importance of appropriate arrangement of test items, Sax and Cromack (1966) and Ahuman and Clock (1971) have advised that tests should be constructed in an easy-to-hard item-difficulty sequence. Based on the findings of this study this advice will augur well for Mathematics in which performance was best in the ETH order. But the advice may not hold for English Language and Science in which performance does not follow this pattern. In the same vein, Skinner (1999) who presented results suggesting that students may actually perform better if tests begin with difficult questions, may agree with the finding of this research especially with regard to Science.

In a study by the Research Division of WAEC, Lagos (1993) the effect of item order on performance in multiple-choice objective tests was investigated. The subjects used were Agricultural Science, Biology, Economics, English Language and Mathematics. As they expected, they found significant differences in the performance of students. Again this finding is in consonance with the findings of this study.

Since this study has shown that altering arrangement according to difficulty level does affect performance, then if performance is poor at any time the item order could be a contributory factor as well as some other factors as Cacko (1993) asserts. His argument that some of the factors could be that the test does not assess the objectives of learning and/or is faulty in structure, clarity and complexity, should be taken seriously in addition to a consideration of the item order.

Thus an answer to the research question “Would a change in item order affect candidates’ performance in anyway?” would be in the affirmative.

It is noteworthy that there was better performance on the Easy-to-Hard treatment of Mathematics. The observation agrees with the recommendations of Sax and Cromack (1966) and Ahuman and Clock (1971) who advised that tests should be constructed in an easy-to-hard item-difficulty sequence. This is because they noted that students perform better on this sequence. On the contrary, a research by the Research Division of WAEC, Lagos on the effect of item position on performance had an unexpected outcome. They wrote that surprisingly, the Hard-to-Easy arrangement was significantly the most effective in Mathematics. That is to say, the researchers were surprised to observe that students performed better on the Hard-to-Easy treatment for Mathematics. A follow up research revealed another interesting finding. It was observed that the Hard-to-Easy arrangement in Mathematics disadvantaged the female candidates.

Since researchers are not unanimous on this observance, the variation for the better performance in the Easy-to-Hard treatment of Mathematics cannot be firmly confirmed. Further studies will be required to confirm this variation.

An interesting question which one may ask in the light of the findings in performance is:

> Which order of arrangement could be adopted to effectively develop the different forms of the test in each of the three selected subject areas?

This question would have been very important in this study if the difference in performance across the treatments had consistently not been significant. However, researchers like MacNicol (1956), Soyemi (1980), Perlini et al (1998) and WAEC Research Division (1993) used the

(a) Random,

(b) Easy-to-Hard;

(c) Hard-to-Easy

arrangements and found it most suitable. McLeod et al (2003) arranged items in

(i) completely randomized combination;

(ii) completely ordered combination;

(iii) partially randomized combination;

(iv) partially ordered combination.

These arrangements cannot be recommended because it is subjective since the degree of partiality to be introduced into the arrangements to make them suitable was not determined.

Pettijohn and Sacco (2001) adopted the sequential (S), random (RA) and reverse (RE) orders. Their investigation focused on the effect of testing the most recent information learned first and working backwards in a reverse sequential order. Since within the same topic both hard and easy questions could be crafted, this method of arrangement may not be suitable for public examinations if the focus is on difficulty levels. The arrangement may be good for only formative tests.

In the generality, the Random, Easy-to-Hard and Hard-to-Easy arrangements are more frequently used than the others mentioned above. Since in this study those treatments were used and they posed no challenges, the same order of arrangement could be followed to develop the different versions of a test if the method were to be adopted for the conduct of a public examination in any of the three subjects under study. However, further studies into other factors like the level of education, may need to be conducted to make the choice of an arrangement more conclusive.

Considering the other side of the coin, the issue which comes to mind is:

> If the difference in performance is significant, what adjustment could be applied to the test scores to neutralize the effect?

The present study has shown that there is indeed statistically significant difference in performance when the positions of the items were altered. Thus some adjustment may have to be applied to neutralize the effect of the difference in performance if the arrangements were to be used in a standardized test.

In a classic example, continuous assessment marks contribute 30% of the total score for final grading in the BECE conducted by WAEC. However, these continuous assessment marks are supplied by classroom teachers with little or no expertise in standardized assessment. Initial perusal of the marks submitted showed that the teachers lump the high achievers and the low ones together and awarded them very high scores. Obviously, their concern, as unprofessional as it might be, was to see their student coming out with good grade by foul or fair means. Thus the marks submitted had little or no correlation with the marks of the testees in the standardized test. To make up for the variation in performances and other factors which may have influenced the classroom teachers when determining the marks, the continuous assessment scores are adjusted using a formula which incorporates the means and standard deviations of both the classroom scores and standardized test scores. One of the reasons for doing this is to introduce some correlation between the performances in the classroom and that in the standardized test. It is also to reduce the skewness which the classroom continuous assessment marks may introduce into the grading system in order to bring about justifiable discrimination between the performance of the high achievers and that of the low ones.

Therefore, depending on what use the results of the test are to be put, further study could be undertaken putting forth different scenarios and determining the best adjustment formula or standardization to apply to make up for the statistically significant differences observed in the performances across the treatments. But this will not be investigated in this study.

Since the study was conducted with examining bodies in mind, the question to ask is:

> If the difference in performance is not significant, what recommendations could be made for examination bodies with regard to item order?

As mentioned by experts like Tauber (1984) examination malpractice has been a challenge for some time now in university examinations. This has been collaborated by Pettijohn and Sacco (2001) who added that professors had to adopt all forms of methods to control the phenomenon. It is not the universities alone which have been affected by this menace. It has been a thorn in the flesh of examining bodies for ages now. As stated earlier, it was reported at the 52nd Annual Council Meeting of the West African Examinations Council in Freetown in 2004 that the phenomenon is on the increase. If permitted to strife it has the potential of lowering standards and diluting the selection and placement processes. It therefore has to be arrested.

The main thrust of this study was to investigate the effect of use of different arrangements of test items for the same cohort of students. Since the study has shown that altering item order would significantly affect performance, the method cannot be recommended in its entirety to be employed as a tool against examination malpractice.

In their recommendations, Adeyegbe and Oke (1994) said that examining bodies should think of using parallel tests to curb examination malpractice. Anastasi (1976) also mentioned that the use of several alternative forms of a test provides a means of reducing the possibility of cheating. Again, Carlson and Ostrosky (1992) state that multiple forms of an examination are a means of reducing the likelihood of cheating in large classes. Further, Bresnock, Graves and White (1989) claimed that to eliminate cheating in objective testing among large number of testees, several versions of tests may be administered. But coming from the results of this study, the re-ordering of formats may not be an appropriate approach.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Summary

The study investigated the impact of item position in multiple-choice test on student performance at the Basic Education Certificate Examination (BECE) level. The purpose of the study was to obtain justification or otherwise for the development and use of different forms of a test to curb examination malpractice at that level of learning.

Three research questions were formulated. They touched on whether or not changing item order in a multiple-choice test will significantly affect performance in English Language, Mathematics and Science at the BECE level.

A quasi-experimental design was adopted for the study. This design was adopted because it has the advantage of being used without randomization and easy to set up. The target population was all Junior Secondary School (JSS) students in Ghana who sat for the April 2006 BECE. The size of this population was 308,325. The sample consisted of 810 JSS students selected from 12 different schools representing urban and rural, and public and private schools.

The instrument for data collection was a multiple-choice test made of 40 items each for English Language, Mathematics and Science. The items were developed to cover all areas in the WAEC syllabus for the subjects involved. A Random order (RDM) of the items was first developed and then using the results of an item analysis conducted, the items were re-arranged to form the Easy-to-Hard (ETH) and Hard-to-Easy (HTE) orders for each of the subjects involved. The test instruments were administered to JSS 3 candidates in the 12 schools selected in the last week of March 2006 was four weeks before the BECE was taken at the national level.

The data collected were subjected to one-way analysis of variance (ANOVA) at $p \leq 0.05$ to determine if any significant differences existed in the performance of students. The independent variable was the Item Order while the dependent variable was the Scores of the tests. A preliminary test for homogeneity of variance was performed to ascertain if population variances were equal. Post Hoc tests were performed to determine the relationship between pairs of test orders as they affect candidates' performance. Answers were then sought for the research questions.

Research Question 1 was:

> What would be the effect of a change in item order on candidate's performance in English Language at the BECE level?

From the findings, the answer given to this question was that the effect of a change in item order on candidates' performance at the BECE level was significant with regard to English Language. It was further discovered that candidates performed better in the RDM order of English Language than the ETH and HTE orders.

Research Question 2 was:

> What would be the effect of a change in item order on candidate's performance in Mathematics at the BECE level?

Again from the findings, the answer given to this question was that the effect of a change in item order on candidates' performance at the BECE level was significant with regard to Mathematics. It was also observed that candidates performed better in the ETH order of Mathematics than in the RDM and HTE orders.

Research Question 3 was:

> What would be the effect of a change in item order on candidate's performance in Science at the BECE level?

From the findings, the answer given to this question was that the effect of a change in item order on candidates' performance at the BECE level was significant with regard to Science. It was also observed that candidates performed better in the HTE order of Science than in the RDM and ETH orders.

Conclusions

The main thrust of this study was to investigate the impact of change of item order in a multiple-choice test at the level of the Basic Education Certificate Examination. From the study, it was discovered that there were significant differences in performance when item order is changed from Random order to Easy-to-Hard or Hard-to-Easy order for English Language, Mathematics and Science. It must be stated that this finding was unexpected. Yet I consider it as a new discovery at that level of education.

The study has further added to knowledge since at the beginning of the study the problem was that it was not known whether using different forms of a multiple-choice test would affect performance in any way at that level of learning. Now we know that if the item order is altered performance will be affected in the subjects chosen for the investigation. The study has therefore facilitated a better understanding of the problem involved in using different forms of a test in an examination of that kind. The study has shown that the proposition of using re-ordering of format of a test to curb examination malpractice may not be the best after all especially in English Language, Mathematics and Science at the BECE level.

Therefore the purpose of the study, which initially was looking for justification for developing different forms of a multiple-choice test to curb examination malpractice at the BECE level, has been re-directed by the outcome. This is because the results do not lend credence to the use of the re-ordering method for the purpose of curbing examination malpractice. The outcome has thus identified a method which may not be employed to arrest examination malpractice because it would not support that course of action. The results also give reason for further research into finding another method which would effectively deal with the menace of examination malpractice at that level of assessment.

The conclusion therefore is that altering item order in a multiple-choice test would affect performance and must, as much as possible, be avoided at the basic level of education.

Recommendations for Policy and Practice

Following from the findings of the study and the above conclusion, the following recommendations are made:

1. Should an examining body decide to pursue the option of administering different versions of multiple-choice tests to curb the spate of collusion in examinations, it must do it with great caution. This is because in employing the method some innocent candidate may be seriously disadvantaged as the outcome of this research has shown. Other methods such as Parallel tests should be investigated for tackling the vice which is assuming alarming rates and sophistication with the introduction of mobile phones into the equation.

   Until further research is done on this method to find way to neutralise the difference in performance, examining bodies would have to step up vigilance during supervision and invigilation of such examinations

2. Learning institutions which depend heavily on multiple-choice tests for summative assessment should also take interest in the search for an appropriate method to curb examination collusion. They should support their staff to undertake research projects that would find solutions for the irregularity.
3. Research units of examining bodies should not give up on the use of the re-ordering method but find appropriate adjustments to neutralize the impact of the difference in performance to facilitate the use of re-ordering of multiple-choice tests to arrest examination malpractices and increase the integrity of the certificates they issue.

Suggestions for Further Research

1. In line with Recommendation 3 above, since the West African Examinations Council (WAEC) is a major examining body in Africa, it should champion further research into the use of different order of a multiple-choice test with the view to finding appropriate adjustments which could neutralize the impact of the difference in performance. This should be done before WAEC embarks on the use of this method. In the event that appropriate adjustments are discovered, requisite training must be given to subject officers in the application of the method. For security reasons, setters may submit their items in a randomized order and subject officers should have the sole responsibility of applying the treatment to develop the different forms.

2. The research involved English Language, Mathematics and Science. Further study should be conducted using other subjects to give more knowledge and understanding for generalization of the findings at this level of assessment.

3. WAEC should consider commissioning a similar study into subjects of the West African Senior School Certificate Examination (WASSCE) to give the findings a universal application since the vice of collusion in examination is quite rampant at that level of education. It will be beneficial if the study could be conducted in each member country of WAEC and the results brought together to present a wider picture and offer a much better understanding of the problem.

REFERENCES

Adetoro, R. A. (2001). An appraisal of frequency of testing and students' performance in junior secondary school social studies certificate examination. *Ife Psychologia, 9*(2), 35-47

Adeyegbe, S. O. A., & Oke, M. G. (1994). *The new and widening dimension of examination malpractices and the effect on the integrity of educational credentials in the West African sub-region.* Compilation of Papers Presented at the 12th Annual Conference of the Association for Educational Assessment in Africa (AEAA), Paper 21, Sept. 19-21, 1994.

Ahuman, S. W., & Clock, N. D. (1971). Item difficulty level and sequence effects in multiple-choice achievement tests. *Journal of Educational Measurement, 9* (Summer), 105-11.

Allison, D. E. (1984). Test anxiety, stress, and intelligence-test performance, *Measurement and Evaluation in Guidance*, *16*, 211 – 217.

Anastasi, A. (1976). *Psychological testing*, New York: Macmillan Press Ltd.

Balch, W. R. (1989). Item order affects performance on multiple-choice exams. *Teaching of Psychology*, *16* (2), 75 - 77.

Bresnock, A. E., Graves, P. E., & White, N. (1989). Multiple-choice testing: question and response position. *Journal of Economic Education*, *(*Summer *1989)*, 239 - 244.

Cacko, I (1993). *Preparation of good objective test items as a step toward obtaining valid assessment of students' achievement at the SSSCE.* Articles of WAEC Monthly Seminar, Accra, March 1993 ed., 87 – 92.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. *In N.L.Gage ed., Handbook of research on teaching*. Chicago: Rand McNally, 29

Carlson, J. L., & Ostrosky, A. L., (1992). Item sequence and student performance on multiple-choice exams: Further evidence. *Journal of Economic Education, 23* (3), 232 -235.

CHARM-Controlled Experiment (2011). *Design of quasi-experiments*. Wikipedia – Free Encyclopedia Online. Retrieved June 6, 2011 from http://www.otal.umd.edu/hci-rm/cntlexp.html

Cronbach, L. J. (1979). *Essentials of psychological testing*. New York: Harperd Row Publishers.

Gekoski, N. (1964). *Psychological testing: Theory, interpretation and practice.* Springfield, Illinios, Charles C. Thomas Publisher.

Gerow, J. R. (1980). Performance on achievement tests as a function of the order of item difficulty. *Teaching of Psychology*, *7*, 93 – 96.

Gohmann, S. F., & Spector, L. C. (1989). Test scrambling and student performance. *Journal of Economic Education, 20* (3), 235 - 238.

Gronlund, N. D., (1976). Item difficulty level and sequence effects in multiplechoice achievement tests. *Journal of Educational Measurement ,9*, 105-111.

Hassan, S. (2005). *Integrity in public examinations: A Malaysian experience.* Proceedings of the First Cambridge Assessment Conference, University of Cambridge, 38 – 43.

Hopkins, K. D. (1998). *Education and psychological measurement and evaluation*. Boston: Allyn and Bacon.

Jessel, J. C., & Sullins W. L. (1975). The effect of keyed response sequencing of multiple-choice items on performance and reliability. *Journal of Educational Measurement 12* (Spring), 45 – 48.

Laffittee, R. G. (1984). Effects of item order on achievement test scores and students' perceptions of test difficulty. *Teaching of Psychology, 11*(4), 212 - 214.

MacNicol, K. (1956). *Effects of varying order of item difficulty in an unspeeded verbal test.* Unpublished manuscript. Educational Testing Service, Princeton, New Jersey.

McLeod, I. (2003). Multiple-choice randomization. *Journal of Statistics Education*, *11* (1), 8 – 9.

Monk, J. J., & Stalling, W. M., (1970). Effects of item order on test scores. *Journal of Educational Research, 63* (July/August), 463 - 465.

Mosier, C., & Price, H. G. (1945). The arrangement of choices in multiple choice questions and a scheme for randomizing choices. *Education and Psychological Measurement, 5* (Winter), 379 - 382.

Neely, D. L., Springston, F. J., & McCann, S. J. H. (1994). Does item order affect performance on multiple-choice exams? *Teaching of Psychology*, *21* (1), 44 – 45.

Ohuche, R. O., & Akeju, S. A. (1976). *Testing and evaluation in education, African educational resources*. Lagos.

Painter, J. (1989). Arrangement of items for a multiple-choice test. *Journal of Educational Psychology*, *11* (Spring), 45 – 48.

Perlini, A. H., Lind, D. L., & Mumbo, B. D (1998). Context effects on examinations: The effects of time, item order and item difficulty, *Canadian Psychology*, *39* (4), 299 – 307.

Peters, D. L., & Messier, V. (1970). The effects of question sequence upon objective test performance. *Alberta Journal of Educational Research, 16* (4), 253 – 265.

Pettijohn II, T. F., & Sacco, M. F. (2001). Multiple-choice exam order influence on student performance, completion time and perceptions. *Journal of Instructional Psychology*, *34* (3), 142 –149.

Sax, G., & Cromack, T. A. (1966). The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement, 3* (Winter), 309-11.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin Company. Retrieved June 8, 2011. from Wikipedia – Free Encyclopedia Online.

Shepard, L. A. (1997). The challenges of assessing young children appropriately. In Katheleen M. Cauley (12th ed.). *Educational Psychology*. Shefield: Dubuque Inc.

Shuttleworth, M. (2008). *Quasi-experimental design*. Experiment resources article. Retrieved March 20, 2011, from www.experiment-resources.com

Skinner, B. F. (1999). When the going gets tough, the tough gets going: effects of item difficulty on multiple-choice test performance. *North American Journal of Psychology, 1* (1), 79 - 82.

Soyemi, M. O. (1980). *Effect of item position on performance on multiple-choice tests*. Unpublished M.Ed. dissertation, University of Jos.

Taub, A. J., & Bell, E. B., (1975). A bias in scores on multiple-form exams. *Journal of Economic Education, 7* (Fall), 58-59.

Tauber, R. T. (1984), *Multiple numbering: An alternative to alternative test forms as a way to reduce cheating on multiple-choice exams.* US Department of Education, National Institute of Education, Educational Resources Information Center, Report 141, 1984.

Tetteh, R., (2005, Oct.15). Exams officer killed. *Daily Graphic* (N0.16588), p.1.

WAEC (1993). *The effects of item position on performance in multiple choice tests*. Research Report, Research Division, WAEC, Lagos.

WAEC (1993). *Examinations syllabus*. Test Development Division. Accra: WAEC Publication.

WAEC (1995). *The effects of sex, ability group and school type on students' performance on the three format of arranging multiple choice test items*. Research Report, Research Division, WAEC, Lagos.

Worthen, B. R., & Spandel, V. (1991). *Putting the standardized test debate in perspective.* Association For Supervision And Curriculum Development, Educational Leadership (Feb. 1991 ed).