

UNIVERSITY OF CAPE COAST

TESTING PRACTICES OF JUNIOR HIGH SCHOOL TEACHERS IN THE  
CAPE COAST METROPOLIS

ERNEST OFORI SASU

2017

UNIVERSITY OF CAPE COAST

TESTING PRACTICES OF JUNIOR HIGH SCHOOL TEACHERS  
IN THE CAPE COAST METROPOLIS

BY

ERNEST OFORI SASU

Thesis submitted to the Department of Education and Psychology of the  
Faculty of Educational Foundations, College of Education Studies, University  
of Cape Coast, in partial fulfilment of the requirement of the Award of Master  
of Philosophy Degree in Educational Measurement and Evaluation

JULY 2017

DECLARATION

**Candidate's Declaration**

I hereby declare that this thesis 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: .....

**Supervisors' Declaration**

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

Principal Supervisor's Signature:..... Date:.....

Name: .....

Co-Supervisor's Signature:..... Date: .....

Name: .....

## ABSTRACT

The study aimed at finding out whether teachers in Junior High Schools in the Cape Coast Metropolis follow the basic laid down principles in their testing practices, with respect to test construction, administration and scoring of classroom or teacher made tests. The study adopted the quantitative method and descriptive survey design was used for the analysis. A total of 50 public Junior High Schools comprising 300 teachers were used for the study. Simple random and purposive sampling procedures were used to select the subject teachers at the Junior High Schools. The main instrument used for the study was questionnaire comprising 55 items with overall reliability coefficient of .873. The data was analysed using means and standard deviations, frequency, and percentages, and One-Way Analysis of Variance (ANOVA). The level of significance for all tests were at 0.05. The result from the study showed that, most teachers at the Junior High Schools did not follow the basic principles of test construction, and scoring of essay type test to an appropriate level, 7 out of 20 construction principles were often practice and 5 out of 13 scoring principles were followed. In test administration, most of the principles were followed, 7 out of 12 principles were often practiced. A statistically significant differences were found among teachers and their subject areas of specification (Mathematics, English, Social Studies, Science and Religious and Moral Education). It was recommended that regular in-service training in testing practices be organized for teachers in Junior High Schools by the Educational Directorate in Cape Coast Metro.

## KEY WORDS

Test construction

Test administration

Test scoring

Validity evidence

Reliability of test

Teaching experience

## ACKNOWLEDGEMENTS

In a quest to undertake this study, I relied on the uninterrupted support and services of many people without whose assistance the work would not have been successful. I would like to express my sincere thanks to Dr Keneth Asamoah-Gyimah, who as my principal supervisor, read through the manuscript and offered very useful suggestions. My profound gratitude goes to Prof. Y. K. A. Etsey, who as my co-supervisor, painstakingly read through the original manuscript and offered valuable suggestions. My unfathomable appreciation goes to the leadership and the vetting committee of the Samuel and Emelia Brew-Butler Research Fund who deemed me worthy for a support from their research fund. I would also wish to show my appreciation to Mr Jonathan Osae Kwapong of the Institute of Education, in the University of Cape Coast for his guidance throughout my educational career.

I wish to thank the Head teachers, and teachers of the fifty selected Junior High Schools who took part in the study, for their acceptance and cooperation during the data gathering process. My appreciation to all authors whose work served as source of reference to my work. However, I admit to be responsible for any errors or shortfalls that emerge from the study.

## DEDICATION

To my parent,

Mr. Joseph Kwadwo Sasu and Madam Christiana Obenewaa.

## TABLE OF CONTENTS

	Page
DECLARATION	ii
ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
DEDICATION	v
LIST OF TABLES	xiii
LIST OF FIGURES	xv
CHAPTER ONE: INTRODUCTION	
Background to the Study	2
Statement of the Problem	8
Purpose of the Study	9
Research Questions	10
Research Hypothesis	10
Significance of the study	11
Delimitation	12
Limitations	13
Definition of Terms	14
Organization of the Study	15
CHAPTER TWO: LITERATURE REVIEW	
Introduction	16
Classical True Score Theory	18
Assumptions of the Classical True Score Theory	18
Validity of Test Items	19
Validity evidence	22



Reliability of Test Items	27
History of Testing and its Development	29
Importance of Classroom Achievement Testing	30
Types of Classroom Teacher-Made Tests	34
Objective-type tests	38
Essay-type tests	41
Principles of Constructing Classroom or Teacher-Made Tests	43
Administration of Classroom Achievement Tests	46
Guidelines in Administering Achievement Tests	50
Scoring of Classroom Achievement Essay Tests	50
Empirical Review	54
Test Construction Practices of Teachers in Ghana	54
Test Administration Practices of Teachers in Ghana	56
Test Scoring Practices of Teachers in Ghana	57
Testing Practices of Teachers in the United States of America	57
Testing Practices of Teachers in England	60
Chapter Summary	61
<b>CHAPTER THREE: RESEARCH METHODS</b>	
Introduction	62
Research Design	62
Study Area	64
Population	65
Sampling Procedure	65
Data Collection Instruments	67
Validity of the Instrument	69

Trial-testing of the Instrument	70
Reliability of the Instrument	71
Ethical Consideration	72
Data Collection Procedures	73
Data Processing and Analysis	74
Research question one	74
Research question two	75
Research question three	76
Research Hypothesis one	77
Research Hypothesis Two	77
Research Hypothesis Three	77
Research Hypothesis Four	78
Research Hypothesis Five	78
Research Hypothesis Six	79
Chapter Summary	79
<b>CHAPTER FOUR: RESULTS AND DISCUSSION</b>	
Introduction	80
Analysis of Demographic Variables	80
Research Question One	83
Research Question Two	90
Research Question Three	95
Research Hypothesis One	100
Research Hypothesis Two	103
Research Hypothesis Three	106
Research Hypothesis Four	109

Research Hypothesis Five	113
Research Hypothesis Six	118
Chapter Summary	122
CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	
Overview of Study	123
Summary of Key Findings	124
Research Question One	124
Research Question Two	125
Research Question Three	125
Research Hypothesis One	125
Research Hypothesis Two	126
Research Hypothesis Three	126
Research Hypothesis Four	126
Research Hypothesis Five	127
Research Hypothesis Six	127
Conclusions	128
Recommendations	128
Suggestions for Future Research	129
REFERENCES	131
APPENDICES	144
APPENDIX A	144
APPENDIX B	152
APPENDIX C	153
APPENDIX D	154
APPENDIX E	155



## LIST OF TABLES

Table	Page	
1	Frequency Distribution of Circuits, Public Schools, and number of Teachers in Cape Coast Metro	66
2	Summary of the Reliability Coefficient of the Items	71
3	Respondents' Qualification	81
4	Form(s) Taught by Respondents	81
5	Years of Teaching of Respondents	82
6	Subject(s) Taught by Respondents	83
7	Test Format used by Teachers	84
8	Test construction practices of teachers	86
9	One Sample t- Test results of Test Construction Practices of Teachers	89
10	Test Administration Practices of Teachers	91
11	One Sample t- Test results of Test Administration Practices of Teachers	95
12	Tests scoring practice of teacher	96
13	One Sample t- Test results of Test Scoring Practices of Teachers	100
14	Test of Homogeneity of Variances	101
15	Descriptive Statistics of Teachers Test Construction Practices	102
16	ANOVA Summary Table for Teachers Test Construction Practices	102
17	Test of Homogeneity of Variances	105

18	ANOVA Summary Table for Teachers Test Administration Practices	105
19	Descriptive Statistics for Teachers Test Administration Practices	106
20	Test of Homogeneity of Variances	107
21	ANOVA Summary Table for Teachers Test Scoring Practices	108
22	Descriptive Statistics for Teachers Test Scoring Practices	108
23	Descriptive Statistics for Teachers Test Construction Practices in Subject Taught	110
24	Test of Homogeneity of Variances	111
25	ANOVA Summary Table for Teachers Test Construction Practices in the Core Subjects	111
26	Multiple Comparisons (Post-hoc/Follow up)	112
27	Test of Homogeneity of Variances	114
28	Descriptive Statistics for Teachers Test Administration Practices in Subject Taught	115
29	ANOVA Summary Table for Teachers Test Administration Practices in the Core Subjects	115
30	Multiple Comparisons (Post-hoc/Follow up)	116
31	Test of Homogeneity of Variances	119
32	Descriptive Statistics for Teachers Test Scoring Practices in Essay-type Subjects	120
33	ANOVA Summary Table for Teachers Test Scoring Practices in Essay-type Subjects	120



## LIST OF FIGURES

Figure		Page
1	Normal Q-Q Plot of Test Construction	101
2	Normal Q-Q Plot of Test Administration	104
3	Normal Q-Q Plot of Essay Test Scoring	107
4	Normal Q-Q Plot for Test Construction	110
5	Normal Q-Q Plot for Test Administration	114
6	Normal Q-Q Plot for Essay Test Scoring	119



## ABBREVIATIONS/ACRONYMS

**AAAS-** American Association for the Advancement of Integrated Science.

**AERA-** American Educational Research Association

**APA-** American Psychological Association

**CGPA-** Cumulative Grade Point Average

**CSTEPP-** Centre for the Study of Testing, Evaluation and Educational Policy

**ERIC-** Educational Resources Information Centre

**GES-** Ghana Education Service

**GMAT-** Graduate Management Admission Test

**GPA-** Grade Point Average

**JCSEPT-** Joint Committee on Standards for Educational and Psychological  
Testing

**JHS-** Junior High School

**MOE-** Ministry of Education

**NCME-** National Council on Measurement in Education

**RME-** Religious and Moral Education

**SPSS-** Statistical Package for Social Sciences

**USA-** United State of America

## **CHAPTER ONE**

### **INTRODUCTION**

Testing has an enormous impact on the practice of education, and it looms large in the minds of countless families as they decide the next step of action after they know the performance of their wards (Koretz, 2008). Classroom teacher made testing seems reassuringly straightforward and common tool used to assess students in almost all levels of education in Ghana. Precisely, because of the importance given to test scores in our society, any mistake that may emerge from the test can have serious consequences in educational decision making.

Previous studies such as Amedahe (1989), Anhwere (2009) and Oduro (2008) have indicated that most of the teachers in the second cycle institutions and tutors in the Colleges of Education in Ghana lacked the basic test construction skills. This pertinent issue has not been examined empirically at the Junior High School level in the context of testing practices of teachers with respect to the basic principles teachers follow in their test construction, administration and scoring of essay-type test in the Cape Coast Metropolis. A practical approach to finding out how teachers in the Junior High Schools adhere to the testing practices would help stakeholders understand the complexities inherent in testing, and provide holistic interventions to avoid the common mistakes and be able to use tests productively.

## **Background to the Study**

Decision making is a daily task and for one to make a sound decision, he/she needs an accurate and relevant information. Constantly in our educational system, decisions have to be made by stakeholders about students, curricula, academic programmes, and educational policies. Similarly, teachers would want to assess the progress of their students, the value or relevance of the curriculum and the effectiveness of instruction in order to make active decisions. Parents would also want to be informed on how much and how well their children are learning (Mehrens & Lehmann, 1991). Finding a solution to these is to a large extent, the sole responsibility of teachers. Hence, teachers ought to give an accurate report that reflect student's achievement. The decisions that teachers make often involve classroom instruction, placing students into different types of programmes, assigning them to appropriate categories, guiding and counselling them, selecting them for educational opportunities and credentialing and certifying their competence (Nitko, 2001).

According to Anhwere (2009), there is the need for teachers to describe the nature and extent of learner's learning in terms of how far the aims and objectives of teaching have been achieved and what is left to be covered. This need calls for assessment of the learner (Tamakloe, Atta & Amedahe, 1996). According to Asamoah-Gyimah (2002), classroom or teacher-made tests are frequently used as a major evaluating device of students' progress in schools. Scarcely can one conceptualize an educational system where the student is not put under a classroom or teacher-made tests. Tests are indispensable tools in the educational enterprise which provide tangible clues for determining the

attainment of learning objective to aid in making effective educational decisions (Anamuah-Mensah & Quagrain, 1998).

According to the Standard for Educational and Psychological testing, National Council on Measurement in Education (AERA/APA/NCME, 2014), “a test is a device or procedure in which a sample of an examinees behaviour in a specified domain is obtained and subsequently evaluated and scored using a standardized process” (p. 2). However, it must be noted that the psychological attributes of an individual cannot be measured directly as can height or weight. The existence of such psychological construct can never be absolutely confirmed. The degree to which any attribute characterises an individual can only be inferred from observation of his or her behaviour (Crocker & Algina, 2008). It becomes more prudent if one can quantitatively relate the subjective judgments of individuals about the estimated amount of construct or trait that exists in a person by establishing standards for such measurement.

Test is an essential tool that helps to quantify such constructs which helps one to make a value judgment about the degree to which such constructs might probably exist in an individual. A large number of assessment techniques may be used to collect information about students. These include formal and informal observation of students, paper-and-pencil tests, a student’s performance on homework, laboratory work, projects and oral questioning and analysis of students’ records (Gyimah & Anane, 2013) .

Teachers in the educational setting would want to estimate the degree to which their students are characterized by the knowledge they have imparted to them within a given period. All the domains of such constructs might not be known by a single test. Nevertheless, a well-constructed test could sample to a

large extent a reasonable amount of the construct on which value judgments could be made from. Educators and teachers must also be aware that a test itself is subject to errors which adversely could affect its use in making decision about students. According to Koretz (2008), “a test score is just one indicator of what a student has learned—an exceptionally useful one in many ways, but nonetheless one that is unavoidably incomplete and somewhat error prone” (p. 10). Tom and Gary (2003) further asserted that:

First, tests are only tools, and tools can be appropriately used, unintentionally misused, and intentionally abused. Second, tests, like other tools, can be well designed or poorly designed. Third, both poorly designed tools and well-designed tools in the hands of ill-trained or inexperienced users can be dangerous. (p. 1).

They went further to state that test misuse and abuse can occur when users of test results are unaware of the factors that can influence the usefulness of the test scores. Among the major factors are the technical adequacy of a test and its validity and reliability. The technical inadequacies might emerge from factors such as, test appropriateness for the purpose of testing, the content validity evidence, the appropriateness of the reading level, language proficiency and cultural characteristics of students and teachers and pupils’ factors that may have affected administration procedure and scoring of the test, among others. It must also be noted that even when a test is technically adequate, misuse and abuse can occur because technical adequacy does not ensure that test scores are accurate or meaningful.

When students’ achievement levels are not properly measured and interpreted, the teachers and school administrators will not be able to provide

the right educational opportunities and support each individual student needs. Testing provides feedback on which educational decisions are made. These decisions may be the ones that require information about the success of learning programmes or about students who have reached particular levels of skill and knowledge (Izard, 2005). Accurate and valid information about student achievement is widely understood to be essential for effective instruction, as it enables teachers to give appropriate feedback and adapt their instruction to match student needs. However, there is much less agreement about the relative merit of different measurement methods used to obtain this information. Previous research has often found substantial positive correlations between teacher judgments of student achievement and the scores the students obtain on standardized tests. However, the strength of this association has been asserted to be varying considerably across subjects, grades, and teachers (Hoge & Coladarci 1989; Perry & Meisels, 1996).

Achievement tests are significant for measuring important aspects of a subject. It reflects the emphasis placed on important aspects of instruction and also measure appropriate levels of student's knowledge in a school subject (Frey, 2007). It is obvious to either use teacher made test to measure students' knowledge or to use a standardized test. According to Childs (1989), the most instructionally-relevant achievement tests, are classroom-based and if carefully constructed, provide teachers with accurate and useful information about the knowledge retained by their students in particular school subjects. Asamoah-Gyimah (2002), confirmed this assertion by stipulating that, classroom or teacher-made test are frequently used as a major evaluating device of students' progress in schools. The extent to which different teachers are able to accurately

assess student achievement is related in part with their educational, professional, and personal backgrounds. For example, variation may be associated with differences in teaching experience (Leinhardt, 1983) or with differences in content area or pedagogical professional development (Rodriguez, 2004).

A particularly interesting area for investigation involves the kinds of practices teachers use in the classroom in testing and monitoring student achievement and progress. According to Stiggins and Conklin (1992), teachers spend a considerable amount of time conducting a range of activities related to the assessment and evaluation of student achievement. Moreover, teachers generally believe that the information they gather through these assessment activities is important for improving student performance (Stiggins & Chappuis, 2005). According to Amedahe (1989), teachers from stage one as well as those in kindergarten to the university level in the Ghanaian educational system engage in some sort of assessment practices which include testing. He explained that, these are done in order to determine whether learning has taken place or not, or sometimes for selection to the next ladder of education. He further stated that teachers also construct tests to find out about problem areas of students in relation to specific topics treated. These and several reasons have necessitated the need for teachers in the Junior High Schools (JHS) to engage in testing practices in their prospective schools.

Quality classroom-based assessment will partly mean adherence to standard procedures for test construction. Every classroom teacher is expected to possess and apply requisite skills in construction of good items for class assessments. A good test must both be valid and reliable to a given extent. Teachers today, perhaps more than ever before, have a need to be

knowledgeable consumers of test information, constructors of assessments and protocols, and even to a large extent, teachers must yearn for skills required about testing (Rudner & Schafer, 2002). Teachers, therefore, need to apply some acceptable degree of test construction skill in order to be able to develop to a degree, valid and reliable tests that will yield accurate feedback of students' achievement. According to Silker (2003), skill in test construction enables a teacher to construct tests with precision, appropriateness of language-use, objectivity and good grading scales.

Teachers need not be experts in educational measurement and evaluation to construct valid and reliable tests, but there are basic test construction skills which every teacher ought to possess to construct quality tests. These skills help teachers to structure items well enough to elicit clear and concise answers from students, construct tests that will be appropriate for learners of different ages, abilities and gender and set tests so that students finish within time and do not grow scared of tests Ali (as cited in Agu, Anyichie & Onyekuba, 2013). If teachers are not equipped with the skills in test construction, it is most probable that they may produce assessment results which may lack a higher degree of reliability. Any characteristic of a test item which distracts the examinee from the major point or focus reduces the effectiveness of that item. Any item answered correctly or incorrectly because of extraneous factors in the item results in misleading feedback to both examinee, teachers and parents as well (Frey, 2007)

Amedahe (1989) stated that:

Although teachers may succeed in their teaching to some extent without following to the later the prescribed principles in testing their students,



more could be achieved if scientific principles and practice deemed useful in measurement are followed. This is very crucial in the Ghanaian educational system where we lack, to a large extent, the availability of standardized achievement and intelligent test found in the developed countries like the United States of America (U.S. A) and Britain. What is called standardized achievement test (which cannot strictly meet the definition of standardized tests) are the tests conducted by the West African Examinations Council (WAEC), which are taken nationally and internationally in the West Africa sub-regions at the terminal point of educational system (p. 4).

### **Statement of the Problem**

Testing at the basic schools in Ghana is based on the assumption that, most teachers have had a course or training in “testing” as part of the assessment process at their various colleges of education. Previous research has indicated that most of the teachers in the second cycle institutions in Ghana lack the basic test construction skills. This was justified by the findings that not all teachers in the Secondary Schools in Ghana have undergone professional training in testing techniques (Amedahe, 1989). In addition, a study by Quagrain (1992) revealed that most Ghanaian teachers had limited skills for constructing the objective and essay type tests, which are the most frequently used in our schools. According to Stiggins (as cited in Agu, et al, 2013), a number of possible reasons could be deduced for such deficiency. This may be due to the fact that; the teachers’ pre-service training did not prepare them adequately in test construction due to little emphasis on assessment practices during their professional development or that most of the teachers failed to acquire test construction skills needed for quality

test item generation while in training. Previous research indicated lack of skills and competency in tutors' knowledge in assessment practices in the Colleges of Education of education (Amedahe, 1989; Etsey, 2003). It therefore seems to be possible that teachers at the basic school level might adversely be affected by the inadequate acquisition of competencies of tutors who are supposed to give them training in testing practices at the Colleges of Education of education.

This study seeks to find out the extent to which teachers in the Junior High Schools in Ghana go about their testing practices based on the basic principles of testing in terms of development and construction, administering and scoring of teacher-made test in the classroom. The issue of the research is: What basic principles do teachers in the Junior High Schools in Ghana follow in the construction, administering and scoring of their classroom tests?

### **Purpose of the Study**

The study seeks to investigate the underlying specific sub-problems of testing in the Junior High Schools in Ghana. Specifically, the study dealt with the issue of testing with regards to:

1. The construction of classroom achievement test by teachers.
2. Test administration and scoring of teacher-made test in the Junior High Schools.
3. Differences in construction of test items among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above.
4. Differences in administering of teacher-made tests among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above.

5. Differences in scoring among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above.
6. Differences in test construction, among teachers in the four core subject areas (Mathematics, Science, English and Social Studies).
7. Differences in test administration among teachers in the four core subject areas (Mathematics, Science, English and Social Studies).
8. Differences in test scoring among teachers in the essay-type subject areas (English, Social Studies and Religious and Moral Education (RME)).

### **Research Questions**

1. What principles do Junior High School teachers follow in constructing test items?
2. What process do Junior High School teachers follow in administering test items?
3. What basic principles of test scoring do teachers in the Junior High Schools follow in scoring of essay-type test items?

### **Research Hypothesis**

- H<sub>01</sub>: There is no statistically significant differences among teachers who have taught for the following number of years in their test construction: 1-5, 6-10, 11-15 and 16 years and above.
- H<sub>02</sub>: There is no statistically significant differences among teachers who have taught for the following number of years in their test administration: 1-5, 6-10, 11-15 and 16 years and above.

- H<sub>03</sub>: There is no statistically significant differences among teachers who have taught for the following number of years: 1-5, 6-10, 11-15 and 16 years and above in terms of their test scoring.
- H<sub>04</sub>: There is no statistically significant differences in test construction, among teachers in the four core subject areas (Mathematics, Science, English and Social Studies).
- H<sub>05</sub>: There is no statistically significant differences in test administration, among teachers in the four core subject areas (Mathematics, Science, English and Social Studies).
- H<sub>06</sub>: There is no statistically significant differences in test scoring, among teachers in the essay-type subject areas (English, Social Studies and Religious and Moral Education).

### **Significance of the study**

The study in general, would inform the Ghana Education Service (GES) and Ministry of Education (MOE) as well as stake holders about the deficiencies that emerge from testing practices for decision making in the sector. This would go a long way to help the GES to provide in-service training to teachers to equip them in their testing practices. A follow up response to the outcome of this research from the MOE could help teachers at the Junior High Schools understand and appreciate the basic rudiments in test construction, administration and scoring practices and equip them in the knowledge of testing practices. It is hoped that the information from this research to the GES would also help them assist teachers to appreciate how well-constructed items, better test administration procedures and following of good test scoring principles could help make valid decisions.

The study reviewed some processes that test developers go through in their testing practices. This finding would go a long way to help the Metro Educational Directorate and GES to understand that some factors can affect the performance of pupils. This will guide the Metro Educational Directorate to ensure that teachers in the Junior High Schools make it a point to go through the right processes in their testing practices. This will assist teachers to reduce errors in test scores and present a score which depicts a fair representation of the students' ability on which valid decisions could be made.

Findings from this research has the potential to assist the policy makers to place much emphasis on testing as part of the curricula in the various colleges of education to ensure that products from the institutions are adequately equipped.

### **Delimitations**

This research looked into the practices of Junior High School teachers pertaining to test construction, administration and scoring. The study did not take into consideration the content of test items constructed by Junior High School teachers with respect to item analysis. How teachers interpret their test scores was also not looked into. There are about 216 districts with six (6) Metropolitans in Ghana with Junior High Schools. Meanwhile the study was delimited to only Junior High Schools in the Cape-Coast Metropolis out of the twenty (20) districts in Central Region. There were 79 Junior High Schools in the Cape Coast Metropolis where teachers also indulge in testing practices, yet the study was delimited to 50 selected Junior High Schools. Participants were selected without consideration of their ethnic, cultural and socio-economic background.

## **Limitation**

The study was challenged by some factors that limited the findings. In the first place, the analysis of the study relied on reported information from the teachers who were part of the study and as such they might either over-reported or under-reported their responses' regarding testing practices in the Metropolis. A further limitation of the study was the lack of classroom observation. This might have given insight to a detailed information into current testing practices of Junior High School teachers in their natural setting. Another limitation was the fact that the study population was restricted to selected teachers in Cape Coast Metropolis in the Central Region of Ghana. Therefore, results of the study would not be generalized to other Metropolis and Districts in the country.

With regard to the research design, descriptive survey data only tells the prevalence and perceived influence of certain variables at the time of data collection. The prevalent situation at the period of the study is subject to change with time. Using the descriptive survey was thus a limitation. A longitudinal data is needed for further clarification of the testing practices of teachers in the Metropolis.

## **Definition of Terms**

For the purposes of clarity, some concepts used in the texts have been defined to assist readers appreciate the context within which they were used in this study.

**Achievement test:** They are tests design to measure attainment of knowledge, skills or ability. Examples are teacher made tests in classroom, homework, class exercise among others.

**Errors in measurement:** They are mistakes which affect individual's score positively or negatively due to chance. Example, administration error, scoring error, content sampling error and errors from fluctuation of individual's behaviour.

**Item analysis:** It is a computation and examination of the statistical property of examinees' responses to an individual test item which help to identify items which are dependable.

**Item format:** The style in which the item has been constructed. It is classified as either selected-response in which the student chooses the correct answer from alternatives provided (e.g. multiple-choice, matching, true/false) or constructed-response in which an answer must be supplied by the student (e.g., essay, short-answer, etc).

**Reliability:** How test scores yield consistent results in a number of successive testing.

**Test:** This refers to an instrument or systematic procedure for observing and describing one or more characteristics or trait of a person.

**Test Item:** It is a test question in a given subject area.

**Validity:** The degree to which the conclusions yielded by a test are meaningful, accurate and useful. The greater the weight of validity evidence that is represented, the more confidence test users can have that they are making accurate decision about the subject in question. It is determined by indicating all practices an assessor will follow to reduce error in measurement.

### **Organization of the Study**

The study is organized into five chapters. The first chapter discusses the introduction, which highlights the background to the study, the research

problem, and the purpose for the study. The research questions were stated, with the significance and delimitations of the study.

Chapter Two reviews the literature related to the study. The review involves empirical studies and conceptual framework. The third chapter describes the methodology used for the study. This involves the research design, population and sampling procedure, the research instrument, the pre-testing procedure, and the procedure for data collection and analysis. In chapter four, the results are discussed while the final chapter summarizes the study and provides conclusions. Recommendations were given in the last section to the chapter based upon the findings of the study.



## **CHAPTER TWO**

### **LITERATURE REVIEW**

#### **Introduction**

This chapter reviews the relevant literature. Information was gathered from journals, abstracts, the internet, books, and works people have done on the present study. For easy referencing, the literature was reviewed under various sub-headings based on the research questions:

1. Theoretical Framework

- Classical True Scores Theory

2. Conceptual framework

- a. History of Testing and its Development
  - b. Importance of Testing
  - c. Classroom Achievement Test

3. Empirical Review

- a. Construction of classroom achievement tests
  - b. Administration of classroom achievement tests
  - c. Scoring of Classroom Achievement Tests

The empirical review deals with available related research findings in the field with respect to construction, administration and scoring of classroom achievement tests.

## **Classical True Score Theory**

A test theory or test model is a symbolic representation of the factors influencing observed test scores and is described by its assumption. Classical true score theory is a simple model that describes how errors of measurement can influence observed score. Classical true score theory states that an observed score ( $X$ ) is equal to the sum of a true score, or true underlying ability ( $T$ ), and the measurement error ( $E$ ) associated with estimating observed scores, or  $X = T + E$ . It is believed that when students take a particular test measuring a construct twice in a succession, it is unlikely that their scores will be identical. This is due to the effect of some factors such as, fatigue, guessing, careless marking, mistakes in scoring. A different form of test would also result in a change in scores because of variation in content. These inconsistencies in individual scores due to the sampling of tasks or occasions must be regarded as measurement error, (Crocker & Algina, 2008). According to Crocker and Algina, the “True Score” can be interpreted as the average of the observed scores obtained over an infinite number of repeated testing of the same test. In the classroom setting, the “true score” is the score a teacher would obtain if he is to take the average score from an infinite number of test administrations. “Of course, in practice, one cannot administer a test an infinite number of times, the vast majority of the time one can get only one chance” (Allen & Yen, 2002 p. 56). According to Crocker and Algina (2008), reliability coefficients are used to estimate both true and error variance associated with observed test scores.

Several assumptions are made about the relationship among these three components (True Score, Observed Score and Error Score). Most of the standard procedures for creating and evaluating classroom teacher made test are

based on a set of assumptions on the Classical true-score theory. The model assumes certain conditions to be true; if these assumptions are reasonable, then the conclusions derived from the model are reasonable. However, if the conditions are not reasonable, then the use of the model leads to faulty conclusions.

### **Assumptions of the Classical True Score Theory**

According to Allen, and Yen (2002, pp. 56-59), the following assumptions of Classical True Score Theory were made:

1.  $X=T+E$  states that, the observed score “X” is the sum of the True score “T” and the error of measurement “E”
2.  $\epsilon(X)=T$ . This states that the expected value (population mean “ $\epsilon$ ”) of “X” is “T”. This assumption is the definition of T: T is the mean of the theoretical distribution of X scores that will be found in repeated independent testing of the same person with the same test
3.  $\rho_{ET} = 0$ . This assumption implies that examinees with high true score do not have systematically more positive or negative error of measurement than examinees with low true score. This assumption will be violated if for example, one administration of a college entrance exam, students with low true scores copied answers from those with high true scores. This situation will create a negative correlation between true score and error score
4.  $\rho_{E_1E_2} = 0$ , where  $E_1$  is the error score for Test 1 and  $E_2$  is the error score for Test 2. This assumption states that, the error scores of two different tests are uncorrelated. That is if a person has a positive error score in Test 1, he or she is not more likely to have a positive or negative error

score in Test 2. This assumption is not reasonable if the test scores are greatly affected by factors such as fatigue, practice effect, the examinee's mood, or effects of the environment

5.  $\rho_{E_1T_2} = 0$ ; This assumption states that, the error scores on one test ( $E_1$ ) are uncorrelated with the true scores on another test ( $T_2$ ). This assumption would be violated if Test 2 measures personality trait or ability dimension that influences error on Test 1. The assumption would also be violated if students with low true scores copied answers from those with high true scores
6. If two tests have observed score,  $X$  and  $X'$  that satisfies assumption 1 through 5, and if, for every population of examinees  $T = T'$  and variance of  $\sigma_E^2 = \sigma_{E'}^2$  then the tests are parallel tests. For,  $\sigma_E^2$  equal to  $\sigma_{E'}^2$  the conditions leading to error of measurement, such as mood, and environmental effect, must vary in the same way for the two tests.
7. If two tests have observed scores  $X_1$  and  $X_2$  that satisfies assumption 1 through 5, and if, for every population of examinees,  $T_1 = T_2 + C_{12}$ , where  $C$  is a constant, then the tests are called  $\tau$ -equivalent test.

The implication of this theory therefore, means that in order to achieve the reliability and validity of classroom teacher made test, the principles of the theory needs to apply. The premise of the theory rest on validity and reliability which are to a large extent, promoters of sound assessment practices.

### **Validity of Test Items**

Validity is “the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of a test” (AERA,

APA, & NCME, 2014, p. 11). Validity according to Nitko (2001) is the “soundness of one’s interpretation and uses of students’ assessment results” (p. 36). This means that for teachers in the Junior High Schools to produce valid results of their students, the student’s results must be supported with many evidences. The results must be devoid of errors and therefore, the soundness of the results. The focus here is not necessarily on scores or items, but rather interpretations made from the results. That is, the behavioural interpretations that one can deduct from test scores is of paramount concern. “In order to be valid, the inferences made from scores need to be appropriate, meaningful, and useful” (Gregory, 1992, p. 117).

Validity is an integrated evaluative judgment on the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment (Messick, 2003). “The validity of classroom assessment depends on, analysing the intended learning and all its embedded elements, having a good match among the assessment approaches, the intended learning, and the decisions that teachers and learners make about learning, ensuring that the assessment adequately covers the targeted learning outcomes including content, thinking processes, skills and attitudes” (Northern Canadian Protocol for Collaboration in Education, 2006, p. 4). There are three types of validity evidence namely, criterion validity, construct validity and content validity.

A measure itself is neither valid nor invalid. Rather, the issue of validity concerns the interpretations and uses of a measure’s scores. “The interpretations and uses of one’s assessment results are also valid only when the values implied by them are appropriate” (Nitko, 2001, p. 37). Essentially, the interpretations

and uses one makes of one's assessment results are also valid when the consequences of these interpretations and uses are consistent with appropriate values Nitko (as cited in Amedahe & Asamoah-Gyimah, 2011). According to Nitko (2001), when the values of the assessment are not in accordance with the consequence of the assessment then this principle is violated. A second important implication of the definition of validity is that validity is a matter of degree, it is not an "all-or-none" issue. That is, the validity of a test interpretation should be conceived in terms of strong versus weak instead of simply valid or invalid (Nitko, 2001).

For test users, validity should be a deciding factor in their choice of psychological tests. Although such choices are based on a number of practical, theoretical, and psychometric factors, a test should be selected only if there is strong enough evidence supporting the intended interpretation and use. A third important facet of validity is that the validity of a test's interpretation is based on evidence and theory. For a test user to be confident in an interpretation and use of test scores, there must be empirical evidence supporting the interpretation and use. In addition, contemporary views on validity emphasize the importance of grounding the interpretation and use of a test in a defensible psychological theory (Nitko, 2001).

### **Validity evidence**

The Standards for Educational and Psychological testing outlined three categories of validity evidence; Content validity, Criterion-related validity and Construct validity, AERA/APA/NCME, (as cited in Miller, McIntire & Lovler, 2011).

### *Content related evidence*

Content Validity is often defined as the extent to which the sample of items, tasks, or questions on a test is representative of the domain of content (Moss, 1992). Bollen (1989) defined content validity as “a qualitative type of validity where the domain of the concept is made clear and the analyst judges whether the measures fully represent the domain” (p.185). But, Wiliam (1993) argues that "content validity should be concerned not just with test questions, but also with the answers elicited, and the relationship between them" (p. 4). Here, Wiliam is advocating for content-related evidence to extend to include the behaviour elicited actually corresponding to the intentions of the assessment task. Wiliam further explained with an example that, “a test claiming to assess students' understanding of forces would be invalidated if it turned out that the reading requirements of the test were so demanding that students with poor reading ability, but a sound understanding of forces, obtained low marks" (p. 4). On the other hand, if a student possesses an understanding of an issue demanded by a test, but fails to show it for reasons of linguistic difficulty then, the results of that test would be invalid.

Wiliam takes this idea from Ackerman and Smith (1988). Ackerman et al. (1988), points out that a test would be considered *biased* and invalid, if it makes different impact on the people who take it because of interfering factors which prevent the appropriate response from being demonstrated. Content-related evidence is therefore, not only demonstrated by the degree to which samples of assessment tasks are representative of some domain of content. It is important for the behaviour elicited by the test item not to have been influenced by factors that conceal the true ability or potential of the student. This could be

an argument in support of school-based teacher assessment as the conditions of assessment can be arranged to provide ecological validity; that means relating the assessments as closely as possible to the learning experiences of the student. As Crooks (2001) point out, "the circumstances under which student performances are obtained can have major implications for the validity of the interpretations from an assessment" (p. 270). Issues such as low motivation, assessment anxiety, and inappropriate assessment conditions can all be threats to the validity of students' assessment results.

Content validity is a general property of a test. A test author who defines the content domain and writes items to represent the domain succeeds to some degree in attaining his/her goal.

According to Miller, et al. (2011 pp. 196-197), there are evidences of validity to be demonstrated based on test content during test development. These evidences include:

1. Defining the test universe which involves the body of knowledge or behaviour that a test presents. They further asserted that, the step involves reviewing other instruments that measure the same construct, and interviewing experts who are familiar with the construct. The purpose is to ensure that one clearly understands and can clearly define the construct being measured. According to Groth-Marnat (1997), evidence of validity based on test content requires that the test cover all the major aspects of the testing universe in the correct proportion.
2. Developing the test specifications/blue print which involves a documented plan containing details about test's content. The specification delineates, the thinking process the test is to measure with



their given proportion, the content area with respect to the subject matter the test is to be measured and the number of questions that will be included to assess each content,

3. Establishing an appropriate test format in which the test will be constructed to elicit the construct of interest,
4. Constructing the test questions. Here test developers are to be careful that each question represents the content area and the objective it is intended to measure.

### ***Criterion-related evidence***

Criterion-related evidence shows the degree of correspondence between a test measure and one or more external referents (criteria), usually measured by their correlation. Criterion-related evidence answers the question, how well the results of an assessment can be used to infer or predict an individual's standing on one or more outcomes other than the assessment procedure itself. Here, the outcome is called the criterion (Etsey, 2012). There are two types of criterion-related evidences. These are concurrent validity and predictive validity. When the criterion exists at the same time as the measure, it means concurrent validity. Concurrent validity refers to the ability of a test to predict an event in the present. In concurrent validity, one is asking whether the test score can be substituted for some less efficient way of gathering criterion data (such as using a score from a group scholastic aptitude test instead of a more expensive-to-gather individual aptitude test score).

For concurrent validity, data are collected at approximately the same time and the purpose is to substitute the assessment result for the scores of a

related variable. For instant a test of swimming ability verses swimming itself to be scored (Etsey, 2012).

When the criterion occurs in the future, it talks about predictive validity. Predictive validity evidence refers to extent to which individual's future performance on a criterion can be predicted from their prior performance on an assessment instrument. For predictive validity, data are collected at different times. Scores on the predictor variables are collected prior to the scores on the criterion variables (Etsey, 2012). The purpose is to predict the future performance of a criterion variable. Etsey further gave example as using first year Grade Point Average (GPA) to predict the final Cumulative Grade Point Average (CGPA) of a University student. Another example is to use students Graduate Management Admonition Test (GMAT) scores to predict their GPA in a graduate programme. Predictive validity can be used to assess the strength of association between the GMAT score with the criterion (i.e., GPA). Although concurrent and predictive validity differ in the time period when the criterion data are gathered, they are both concerned with prediction in a generalizability sense of the term. Both concurrent and predictive reliability would aid one to tell whether an individual behaviour should be reinforced concurrently or based on one's behaviour, one will be able to perform a particular task in the future (Etsey, 2012).

### ***Construct related evidence***

DeVellis (1991) explains that the construct validity of a measure "is directly concerned with the theoretical relationship of a variable (e.g. a score on some scale) to other variables. "It is the extent to which a measure 'behaves'

the way that the construct it purports to measure should behave with regard to established measures of other constructs” (p. 46).

Messick's (1989) definition of construct related evidence captures the breadth of the concept of validity; "validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and *actions* based on test scores or other modes of assessment" (p. 13.). Moss (1992) points out that "the essential purpose of construct validity is to justify a particular interpretation of a test score by explaining the behaviour that the test score summaries" (p. 233). This means asking whether the interpretation given to the test score truly summaries the behaviour. That is, a construct needs to be both operationalized and syntactically defined in order to measure it effectively. The operationalizing of the construct involves developing a series of measurable behaviours or attributes that are posited to correspond to the latent construct. Defining the construct syntactically involves establishing assumed relationships between the construct of interest and other related constructs or behaviours (Benson, 1998; Crocker & Algina, 1986; Gregory, 1992).

According to Trochim (2006), construct related evidence relates to how well you translate or transform a concept, idea, or behaviour that is a construct into a functioning and operating reality.

According to the Standards for Educational and Psychological Testing, AERA, APA, and NCME (2014), validity is not viewed as three separate components but as a unitary concept. They rather focus on interpretation of test scores and “accumulating evidences to provide a sound scientific basis for the proposed score interpretation” (Goodwin & Leech, 2003, p. 3). In the view of

Nitko (2001), for teachers to validate the interpretation and the uses of assessment result, teachers must provide evidences that this interpretation and uses of that test is appropriate. With respect to testing practices of teacher made tests, all the systematic processes the teacher would follow to reduce measurement errors in order to ensure scores obtained by students are the true reflection of the students' ability contributes to validity evidence.

### **Reliability of Tests**

Reliability is defined as the degree of consistency between two measures of the same thing (Ebel & Frisbie, 2001). It is the degree to which assessment results would be similar under slightly different measurement conditions. For instance, if one assesses a student twice, one hopes that he would obtain almost the same score if the student is assessed a day later. Here, if one measures a person's level of achievement, one hopes that the scores will be similar under different administrators, using different scorers, with similar but not identical items.

A reliable assessment is one that consistently achieves the same results with the same (or similar) cohort of learners. According to AERA/APA/NMCE (2014), in a more general sense, "reliability refers to consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported" (p. 33). If the assessment process is reliable, the inferences about a learner's learning outcome should be similar when they are measured by different teachers, when learning is assessed using various methods or when learners demonstrate their learning at different times (Northern Canadian Protocol for Collaboration in Education, 2006). According to William (2008), a reliable test is one in which scores that a learner gets on

different occasions or with a slightly different set of questions on the test, or when someone else does the marking, does not change very much. Various factors affect reliability including ambiguous questions, too many options within a question paper, vague marking instructions and poorly trained markers. Decisions are based on data. These data may come from classroom and standardised test scores, classroom observations, parental reports and many other sources. In using the data for decision making, one should know something about the quality of the data. Here, high-quality data should be weighted more deeply in one's decision than poor-quality data. In principle, data should be reliable, and the inferences one draws from the data should be valid. Reliability is paramount in assessing individuals.

Reliability operates at two levels, i.e, the individual assessed, and that of a number of assessors (Freeman & Lewis, 2008). Reliable assessors make the same decision on a particular assessment whenever they mark it. When more than one assessor is concerned, reliability is achieved if presented with work of the same standard and all assessors make the same judgment. Reliable assessment ensures accurate and consistent comparisons, whether between the performances of different pupils or between a learner's performance and the criteria for success (Freeman & Lewis, 2008). Maizan (2005) postulates that there are three types of reliability that are most relevant to classroom tests. These are internal consistency, inter-scorer and intra-scorer reliability. In the view of Maizan (2005) internal consistency refers to the consistency of objectives among the items of a test while inter-score reliability refers to the consistency between marks given by different teachers. On the other hand, intra-scorer reliability refers to marks given by the same teacher on different

occasions. According to Brown (2007), “the major threat to reliability is the lack of consistency of an individual marker” (p. 78). However, intra –scorer reliability might not in fact be a major concern when a test scorer is supported by rubrics (Johnson & Svnyby, 2007). Consistent grading is essential in order to ensure reliability of test scores.

### **History of Testing and its Development**

The historical development and up-bringing of testing in Africa and in Europe have been interwoven with the development of psychology as a scientific discipline. Test theory evolved from testing covers three major areas of development: civil-service examination, school examination, and the study of individual differences. Civil service testing began in China about 3000 years ago when an Emperor decided to assess the competency of his officials. Later, government positions were filled by persons who scored well on examinations that covered topics such as music, horsemanship, civil law and writing. Such examinations were eliminated in 1905 and were replaced by formal educational requirements. In the West, in England, civil service ability testing was adopted during the middle portion of 19<sup>th</sup> century (Cunningham, 2001; Flanagan et al., 1997). The increase in the use of test in Britain and the United States was attributed to the use of tests as a fair way of selecting among applicants for government jobs (Du Bois, 1970). In the USA, testing began in the later part of the 19<sup>th</sup> century, Du Bois (as cited in Cunningham, 2001; Flanagan et al., 1997). DuBois pointed out that following the successful use in England of the Chinese method of selecting government employees, the method was adopted in USA. The development of academic tests was pioneered in Britain, particularly in the University of London.

Test later became the concern of most nations around the globe. Students in European schools were given civic examination until well after the 20<sup>th</sup> century, when paper began replacing parchment and papyrus. In France, Binet (1905) developed the first individual tests of intelligence as part of his work on the study of individual differences. A German, Stern (1928), developed the intelligent quotient (IQ), which he defined as the ratio of mental (measured) age to chronological (actual) age. Charles Spearman, a British, followed the footsteps of Galton and Pearson, and his work led to the modern concepts of test reliability and factor analysis. Test development, like many other aspects within psychology and education, is a product of many contributors and disciplines throughout history.

### **Importance of Classroom Achievement Testing**

Educational uses of tests have been classified under instructional management decisions, selection decisions, classification decisions, placement decisions, counselling and guidance decisions, and credentialing and certification decisions (Nitko, 2001; Amedahe & Asamoah- Gyimah, 2003). The instructional management decisions refer to all the classroom decisions taken by the teacher on the basis of the assessment results of students. Firstly, tests provide useful information for instructional diagnosis and remediation. The classroom teacher constantly needs to diagnose his instruction and remediate the aspects which have been defective (Amedahe & Asamoah-Gyimah, 2003). This is made possible through feedback from students to the teacher. In instructional diagnosis and remediation, the teacher engages in diagnostic testing to identify which students need remedial help or special attention. According to Nitko (2001), diagnosis involves identifying both the

appropriate content and the features of the learning activities in which a student should be engaged to attain the learning target.

Tests are used in the modelling of learning targets. According to Nitko (2001), “assessments define for students what the teacher wants them to learn”. (p. 9). He continued by noting that students can always compare their current performance on the learning targets with the desired performance. The teacher can then teach his students to detect the ways in which their performance is matching the criterion and the ways in which it is deficient. In this way, the teacher can direct his teaching on the remediation of any identified deficiency and students are also able to know what is important to learn once they are able to evaluate their own performance vis-à-vis the desired learning targets.

Tests are needed for the provision of motivation for students, rewarding those who have prepared well in advance and providing negative consequences for those who have not prepared well. The frequency of an individual behaviour is increased by reinforcement. Hence, it can be reasonably concluded that tests cause students to study more in the sense that the motivation derived from tests as a result of performing well can activate and direct their learning by sustaining their interest (Cunningham, 2001; Ebel & Frisbie, 1991; Gronlund, 2008; Nitko, 2001).

Tests are used for the assignment of grades to students. The grades or symbols (A, B, C) that the classroom teacher reports, represent his /her formal evaluation or judgement of the quality or worth of his/her students ‘achievement of the important learning objectives (Amedahe & Asamoah-Gyimah, 2003; AERA/APA/NCME, 2014; Nitko, 2001). It is worth noting that assessment results of which tests constitute the most important part as it is in the Ghanaian



educational system provide the basis for the assignment of grades. AERA/APA/NCME (2014) have cautioned here that to serve effectively the purpose of stimulating, directing and rewarding students' effort to learn, grades must be valid. To achieve this, the highest grades must go to those students who have demonstrated the highest level of achievement with respect to the course objectives.

On the issue of selection decisions, sometimes, an institution decides whether some persons are acceptable for specific programmes while others are not. Those not acceptable are rejected and are no longer the concern of the institution (Amedahe & Asamoah-Gyimah, 2003; Cronbach, 1960; Nitko, 2001). An educational institution often uses test results to provide part of the information on which selection decisions are based. Typical examples are the selection of candidates for admission into Senior High Schools (SHS) in Ghana which is based on the test scores of students at the end of the Junior High School and university admissions in Ghana which are based on the test scores of students at the end of the SHS.

Tests provide the basis for the grouping of children with reference to their ability to profit from different types of school instruction and the identification of the intellectually retarded and the gifted (Cunningham, 2001). Nitko (2001) has pointed out that sometimes, based on test results, a decision is made that result in a person being assigned to one of several different but unordered categories of programmes. According to Cronbach and Glaser (cited in Nitko, 2001), these types of decisions are called classification decisions. These decisions result in either assigning students in the same classroom to different groups for effective instruction or assigning students to special

education classes. Cunningham (2001) however cautioned test users about the over reliance on test results in assigning students to special education classes by pointing out that intelligence tests are only one component of the assessment procedure used to decide possible placement of students in special classes.

On the issue of placement decisions, Cronbach (1960), Kubiszyn and Borich (1984), and Nitko (2001) have pointed out that placement decisions are made after an individual has been accepted into an educational programme. Corno, Cronbach, Kupermintz, Lohman, Mandinach, Porteus, and Talbert (2001), continued by noting that placement decisions basically involve using assessment results or test data to determine where in a programme an individual is best suited to begin work. Such decisions are characterised by assigning individuals to different levels of the same general type of instruction or education based on their ability, with no one rejected by the institution Cronbach and Glaser (as cited in Nitko, 2001). Promotion in Ghanaian schools from one class or form to another which in most cases is based on the performance in tests of the previous class is an example of a placement decision.

Counselling and guidance decisions involve using assessment results, with test data inclusive, to help students in exploring and choosing careers and in directing them to prepare for the careers they select (Anastasi, 1982; Amedahe & Asomoah-Gyimah, 2003; Kubiszyn & Borich, 1984; Nitko, 2001). Amedahe and Asamoah-Gyimah (2003) have explained that guidance is one of the students' personnel services provided in a non-instructional setting to cater for the needs of students including educational, emotional, and moral and adjustment needs. Nitko (2001) and Amedahe and Asamoah-Gyimah (2003) have agreed with the fact and argued that due to the complexities involved in

guidance and counselling decisions, test data must always be combined with other assessments such as interviews, interest inventories, various aptitude tests and personality questionnaire together with additional background information on students and discussed with students in a series of counselling sessions in order to help students make good decisions.

On credentialing and certification decisions, Nitko (2001) and Amedahe and Asamoah-Gyimah (2003) explained that they are concerned with assuring that a student has attained a certain standard of learning. In Ghana, certification and credentialing of students is done by the West Africa Examination Council (WAEC). With the introduction of the practice of continuous assessment as a result of the educational reforms in 1987, Ghanaian classroom teachers contribute 30% of the total marks for certification of students at the JHS and SHS levels (Amedahe, 2000; Pecku, 2000).

### **Types of Classroom Teacher-Made Tests**

Assessments made by teachers of student's attainment, knowledge and understanding are called variously as teacher-made or classroom made test and school-based assessment (Amedahe, 1989). The predominance of teacher-made tests in every educational set up is given credence by the conclusions of studies by Herman, Dorr-Bremme, Stiggins and Bridgeford (as cited in Mehrens & Lehmann, 2001) that, in the face of the ever-increasing use of portfolios and performance tests to assess student progress, teacher-made tests are mostly the major basis for evaluating student progress in school.

The rationale of teacher-made tests is linked with the constructivist model of learning. The model elaborates on the importance to understand what the student knows and how he/she articulates it in order to develop his/her

knowledge of understanding. With respect to this model, it is learning with understanding which counts and for that matter, information about existing ideas and skills is essential (Chauhan, 2008; Glassman & Hadad, 2013). Work in psychology and learning portrays similarly that for effective learning, the task must be matched to the student's current level of understanding and either pitched at the level to provide practice or slightly higher in order to extend and develop the student's skills (Gipps, 1992a). If the new task given to students is too easy, the students can become bored, difficult items, can also de-motivate (Gipps, 1992b). For content of a course to be adequate and ensure that it is relevant as well, the content should match the understanding level of a particular student (Gipps, 1999). Salvia and Yesseldyke (2001) asserted that, teacher made tests are better when used to evaluate students because they are curriculum matched.

Teacher-made tests can be classified in a variety of ways. According to Mehrens and Lehmann (2001), one type of classification is based on the type of item format used — essay-type versus objective-type. Another classification is based on the stimulus material used to present the tests to students—verbal versus non-verbal, while other classifications may be based on the purposes of the tests and the use of the test results, criterion-referenced versus norm-referenced, achievement versus performance, and formative versus summative. Essentially, there are two main forms of teacher or classroom-made test. These are formal and informal tests. The formal has to do with the process of obtaining information that is used for making decision about students, school, curricula and programmes and educational policies using standardised instrument, example is pencil and paper tests (Gyimah & Anane, 2013). According to

Gyimah, Ntim and Deku (2010) “informal assessment means a procedure for obtaining information that can be used to make judgement about characteristics of children or programmes using means other than standardised instruments” (p. 123).

For example, a teacher who assesses his students through the use of projects, class presentations, experiments, portfolios, interviews, observations, checklist, is using informal assessment. Informal assessment does not follow any strict formalities. Its purpose is to collect information on students that can direct instruction. Teachers may pose questions, observe activities, and evaluate students’ work in a planned and systematic or ad hoc way (Gipps, Brown, McCallum, & McAlister, 1995). Classroom achievement tests are generally teacher-made tests (McDaniel, 1994). These tests are constructed by teachers to test the amount of learning done by students or their attainment at the end of a course unit, term or at the end of an academic year (Amedahe, 1989). Teachers have the responsibility to provide their students with the best instruction possible. This implies that they must have some relevant content procedures or method whereby they can reliably and validly evaluate how effectively their students have learnt what has been taught them (Mehrens & Lehmann, 2009). The pencil and paper or teacher-made test is one such tool.

According to Mehrens and Lehmann (2001), teacher-made tests usually measure attainment in a single subject in a specific class or form or grade. Classroom tests can, be tailored to fit a teachers’ particular instructional objectives, essentially, when one wishes to provide for optimal learning on the part of the pupil and optimal teaching on the part of the teacher (Bejar, 1984). Here, without classroom tests, the objectives that are unique to a particular

school or teacher might not be evaluated. The emphasis on the desirability and importance of the classroom teachers being able to construct their own personal, unique and relevant tests is based on the principles of assessment in education.

A survey conducted by Stiggins and Bridgeford (1985) on the uses of various types of tests reported that, tests are for assigning grades and evaluating the effectiveness of an instructional treatment, for diagnosis, for remedial teaching to motivate students to learn to improve in their work, to provide the basis for guidance in selection and placement in the world, and for certification.

Despite the aforementioned importance of teacher-made tests, a study conducted in the United States of America revealed some deficiencies in teacher-made tests, in the sense that, teachers were only trained to teach but not to assess their students (Gullickson, 2001). The main purpose of teacher-made tests has been delineated by measurement experts (Etsey, 2004; Gronlund, 2008; Mehrens & Lehmann, 2009). All these authorities have agreed with the fact that teacher-made test is to obtain valid, reliable, and useful information concerning students' achievement and thus contribute to the evaluation of educational progress and attainments for the total improvement of classroom teaching and learning. According to Nitko (2001), assessment content is relevant when teacher-made or classroom test comprises choice formats such as (multiple choice, true or false, matching exercise and other formats like greater - less same items), short answers and completion format and essay format (restricted responses and extended responses). Some educators argue that essay tests are more susceptible in scoring than the objective tests. However, classroom teachers exclusively use both since one cannot be used exclusively

to measure all learning outcomes. According to Bartels (2003), with regard to the objective type tests, the multiple choice, short-answer/fill-in-the blanks, matching and true or false types are the major ones used by tutors in the teacher colleges of education in Ghana.

### **Objective -type tests**

The objective-type item was developed in response to the criticism levelled against the essay type tests. Some of the criticisms were, poor content sampling, unreliable scoring, time-consuming to grade, and encouragement of bluffing. The objective test-items normally consist of a large number of items and the responses are scored objectively, to the extent that competent observers can agree on how responses should be scored (Amedahe & Etsey, 2003).

Objective-type item formats are put into two groups; the supply type and the selection type. The supply type format consists of completion type, fill-in-the blanks and short answer. The selection type consists of true-false, matching, and multiple-choice item type. According to Amedahe and Etsey (2003) objective type test items are most useful when class sizes are very large and when there is limited time to submit the results of the test. The short-answer and completion format consists of one or more blanks in which the student writes his answers to the question with a word or, phrase. This type of objective test is also known as constructed response type. It consists of a statement or question and the respondent is required to complete it with a short answer usually not more than one line (Etsey, 2012). It is used for testing knowledge of facts or recall of specific facts (example, “knowledge objective” in Bloom’s taxonomy of educational objectives). Short-answer and completion format can be used to assess higher-level abilities like, making simple interpretations of data and

applications of rules, solving numerical problems in science and mathematics, and to manipulating mathematical symbols and balancing mathematical and chemical equations.

A true or false test consists of a statement to be marked true or false. Here their utilities are placed primarily in assessing knowledge of factual information. True or false items are difficult to prepare (Salvia & Ysseldyke, 2001). True or false test items are made up of four types; simple true or false, (here only two choices; true or false), complex true or false (comprises three choices; true or false and opinion), compound true or false (consists of two choices, true or false plus a conditional completion response) and finally multiple true or false (consist of a stem with three, four or five options and the respondent indicates if the options are true or false (Etsey, 2012).

One of the limitations in constructing the true or false test items is that, the probability of getting right answer by guessing is high. It can be used to assess only a few number of educational objectives, and can be used to evaluate definitions, facts, recognition, and interpretation of charts/graphs. An advantage of true or false test item is that, they can cover a wide range of content within a relatively short period of time.

Matching test format is another choice format item which presents respondents with three things; (a) Directions for matching (b) A list of premises (c) A list of responses. The simple matching exercise requires simple matching based on association that a student must remember. This is basically done to assess respondents' comprehension of concepts and principles. One of advantages of matching test format is that, matching test formats use pictorial materials to assess student's abilities to match words and phrases with pictures



of objects or with locations on maps and diagrams. Nevertheless, the matching sometimes does not help in measuring the higher level of thinking.

A multiple choice item consists of a stem followed by a list of two or more proposed alternatives. Here the respondents are expected to select the correct option from the alternatives. Normally, only one of the options is the correct or best answer to the question. This is called the keyed alternative, keyed answer or basically the key while the remaining incorrect options are called foils or distractors. The purpose is to allow students to demonstrate their knowledge and understanding of the learning targets.

There are two types of multiple-choice tests. These are the single correct type and the “multiple responses” type. The “single correct” type consists of a stem followed by three or more responses and the respondent is to select only one option to complete the stem. The “multiple responses” type consists of a stem followed by several true or false statements or words. The respondent is to select which statement could complete the stem. Multiple-choice tests format do not require students to write out and elaborate their answers and minimize the opportunity for less knowledgeable students to “bluff” or “dress-up” their answers (Wood, 2007).

Etsey (2012), outlined some strengths and weaknesses of objectives items. These strengths include, objectivity and easy in scoring, the potency to allow an extensive coverage of subject content, preventing tests takers from opportunities for bluffing, best suited for measuring lower-level behaviours like knowledge and comprehension, provision of economy of time in scoring, minimising students writing (premium is not placed on writing) and its amenability to item and statistical analysis. He further stipulated that scores

from objective-type tests are not affected by extraneous factors such as the likes and dislikes of the scorer.

Despite the strength of the objective-type tests, Etsey (2012) further gave a number of weaknesses of the objective-type tests. He explained that, they are relatively difficult to construct, writing the items is time consuming and they are susceptible to guessing. He added that using the objective-type to measure higher-order mental processes like analysis, synthesis and evaluation are difficult.

### **Essay-type tests**

According to Amedahe and Etsey (2003), essay-type tests consist of relatively few items, but each require an extended response. Essay test items provide respondents with the freedom to organize their own ideas and respond with limited restriction. Here respondents are asked to speak to a particular issue and for that reason they could not just write a single word as an answer but expresses themselves in terms of what they know about the items. The ability of the respondents to express themselves clearly and fluently and with content required tells the instructor that they have actually mastered the content of the subject. Essay questions are most useful in assessing instructional objectives prepared at a comprehension level or higher order thinking (Salvia & Yesseldyke, 2001). Nitko, (2001) noted that “what is perhaps unique about the essay format is that it offers students opportunity to display their abilities to write about, to organize, to express and to explain interrelationships among ideas” (p.187).

The essay test has two major types; extended and restricted response depending on the amount of scope or freedom given the student to organize

ideas and write answers. Extended-response type of essay questions has no bounds placed on the student as to the point(s) to discuss and the type of organization to use. This type of question permits the student to demonstrate the ability to call on factual knowledge, evaluate factual knowledge organize ideas and present ideas in a logical, coherent written fashion (Etsey, 2012).

The extended response makes the greatest contributions at the levels of synthesis and evaluation of writing skills (style, quality).

Under the restricted-response essay questions, the student is more limited in the form and scope of the answer because it tells specifically the context that the answer is to take. This type of question is of greatest value for measuring learning outcomes at the comprehension, application, and analysis level, and its use is best reserved for these purposes.

Etsey (2012), outlined some strengths of essay-type test items. He postulates that, they provide the respondent with freedom to organize his own ideas and respond within unrestricted limits, they are easy to prepare, they eliminate guessing on the part of the respondents, skills such as the ability to organize material and ability to write and arrive at conclusions are improved. They further encourage good study habits as respondents learn materials in wholes, they are best suited for testing higher-order behaviours and mental processes such as analysis, synthesis and evaluation. Little time is required to write the test Items and the format is practical for testing a small number of students.

The following weaknesses of essay-type test have been elaborated on by Estey (2012). He affirmed that essay-type tests are difficult to score objectively, they provide opportunities for bluffing where students write irrelevant and

unnecessary material, limited aspects of student's knowledge are measured as students respond to few items only, the items are an inadequate sample of subject content and several content areas are omitted. He further explained that premium is placed on writing. Students who write faster all things being equal, are expected to score higher marks. It is said to be time-consuming to both the teacher who scores the responses and the student who writes the responses. Lastly he noted that they are susceptible to the halo effect where the scoring is influenced by extraneous factors such as the relationship between scorer and respondent and also critical reader as well as a competent scorer can only effectively score responses.

### **Principles of Constructing Classroom or Teacher-Made Tests**

. Good and quality test items are not just constructed by test constructors or experts. They require adequate and extensive planning so that the instructional objectives, the teaching strategy to be employed, the textual material, and the evaluative procedures are all related (Mehrens & Lehmann, 1991). Ideally, every test should be reviewed critically by other teachers to minimize the deficiencies identified in. Without adequate and careful planning, one can be fairly certain that one's test will not be very good (Tinkelman, 1971). Tinkelman further noted that "at the very least, inattention to planning can lead to waste and to delay due to failure to coordinate properly the various phrases of test construction" (p. 46).

With respect to carefully development of a good test, Mehrens and Lehmann (2009,) outlined the following stages and steps as being important to the construction of the classroom or teacher-made tests.

1. Specify the course or unit content.

2. List the major course or unit objectives.
3. Define each objective in terms of students' behaviour.
4. Discard unrealistic objectives.
5. Prepare a table of specifications.
6. Decide on the type of item format to be used.
7. Prepare test items that match the instructional objectives.

In addition to the basic principles of test construction, Adamolekun (1985) indicated that, in writing of any classroom or teacher-made tests, it is prudent the teacher considers the following;

1. Identify the purpose of the test i.e. what the teacher wants to achieve by the test.
2. Select the test item type that will best measure the learning outcome.
3. Obtain a representative sample of student behaviour which the teacher would want to evaluate (e.g. in the affective domain; does the teacher wants to know how a student has received a classroom activity, responding, valuing, organization, characterisation by a value complex?)
4. Construct test items of the proper level of difficulty.
5. Try to eliminate factors that are extraneous.

The quality of a test given by a teacher is closely linked with its ability to provide the kind of information needed regarding students' performances. A well written test allows the teacher to accurately and consistently measure students' mastery of specific contents taught in class (Education Up Close, 2005). Results of such tests allow teachers to measure to some degree, how effective their instruction has been. Conversely, poorly designed test items can

lead to inaccurate measurements of learning and provide false information regarding student performance as well as instructional effectiveness (Education Up Close, 2005). Any characteristic of a test item which distracts the examinee from the major point or focus reduces the effectiveness of that item (Frey, 2007). Frey further noted that any item answered correctly or incorrectly because of extraneous factors in the item results in misleading feedback to both examinee and examiner.

Ambiguous question can also affect the reliability of a test. Ambiguous question is when a statement or word have two or more meanings. For example, in essay tests, words such as discuss or explain may be ambiguous in that different pupils may interpret these words differently. Again, excessive wording contributes to difficulty in teacher-made test (Etsey, 2012). Too often teachers think that the more wording there is in a question, the clearer it will be to the student. This does not always happen. The more precise and clear-cut the wording, the greater the probability that the student will not be disorganised. Mostly, teacher-made tests do not cover the objectives stressed and taught by the teacher and do not reflect proportionally the teacher's judgement as to the importance of those objectives. Teacher-made achievement tests are mostly heavily loaded with items that only test the students' ability to recall specific facts and information (Fleming & Chambers, 1983).

Use of inappropriate item formats also contributes to deficiency in teacher-made tests. Some teachers use different item formats like true-false or essay solely because they feel that change or diversity is desirable. But the need for diversity should not govern the type of item to be used therefore; teachers

should be selective and choose the format that is most effective for measuring a particular objective.

Koksal (2004) outlined factors that are inherent in poorly designed tests which if well-handled will lead to quality classroom-based tests. These are:

1. Non specification of the target audience, what skill or area of ability the test intended to measure, how much time allocated for each test item, and what points the test-takers would get for each correct response.
2. Separate sections not clearly stated.
3. Test items having more than one possible answer because they were not conceptualized.
4. Not stating time allocated for each task on the papers. Only the total time available to perform all the tasks is given.
5. Non consideration of level of students in test construction.
6. Unclear instructions.
7. Tasks students are expected to perform not being in concert with the tasks they are told to do during the classroom instructions.
8. Lack of representativeness of items concerning what the teacher intends to test.

### **Administration of Classroom Achievement Tests**

The guiding principle in test administration is to provide all examinees with a fair chance to demonstrate their achievement on what is being measured (Gronlund, 2008; Tamakloe, 2006). The need to maintain uniform conditions in test administration cannot be over-emphasised. This is especially essential for the test to yield consistent, reliable and valid scores without much influence of chance errors. This is emphasised by JCSEPT (1999) by stating that,

“reasonable effort should be made to assure the integrity of the test scores by eliminating opportunities for test takers to attain scores by fraudulent means” (p. 64). This calls for ensuring a congenial psychophysical atmosphere for test taking (Tamakloe, 2006). This was also emphasised by Airasian (as cited in Amedahe & Asamoah-Gyimah, 2003) that test administration is concerned with the physical and psychological setting in which students take their tests.

The first and foremost task of the teacher is to prepare his students in advance for the test (Etsey, 2004). Etsey has emphasised that for students’ maximum performance, they should be made aware of when (date and time) the test will be given, the conditions (number of items, place of test, open or closed book) under which the test will be given, the content areas (study questions or list of learning targets) that the test will cover, the emphasis or weighting of content areas, the kinds of items (objective-types or essay-types) on the test, how the test will be scored and graded, and the importance of the results of the test.

The physical conditions that need to be in place to ensure maximum performance on the part of students include adequate work space, quietness in the vicinity, good lighting and ventilation and comfortable temperature (Etsey, 2004). Adequate work space is very essential for test administration because when tables and chairs are closely arranged, students will not have the independence to work on their own. This will in no doubt lead to students copying from each other. In addition, tables provided for the examination must be conducive to the testing materials being used. For example, in Practical Geography examinations where topographical sheets are used, each student could use two tables or desks in order to get adequate work space (Tamakloe,



2006). Noise and distraction in the testing environment should be kept at the barest minimum if not eliminated completely. Interruptions within and outside the testing room has the tendency of affecting student's performance (Mehrens & Lehmann, 2001). Etsey (2004) has pointed out that it is helpful to hang a "Do Not Disturb. Testing in Progress" sign at the door of the testing room to warn people to keep off. Good lighting is important in effective test administration. This facilitates students' reading of instructions and test items without straining their eyes, thereby working faster. Good ventilation and comfortable temperature should be assured since their absence could create unrest or uneasiness in testees making concentration difficult (Tamakloe et al., 1996). Other basic physical conditions are that, all testing equipment must be in the room and readily available, and also, all possible emergencies during test administration must be expected and well catered for.

The psychological conditions in test administration, on the other hand, include the position of the invigilator, timing of the test, threatening behaviours of invigilators, and interruptions to give instructions and announcements (Etsey, 2004). A study on the examiner as an inhibiting factor, carried out by Bernstein and reported by Amedahe (1989) found out that, the presence of the examiner tended to inhibit the performance of those students who were nervous. The crux of the matter is that if the mere presence of the examiner or invigilator could affect the performance of students who are nervous, then there is no doubt that the position of the invigilator is very significant to the performance of students on examinations. Etsey (2004) has recommended that the invigilator should stand where all students could be viewed and move among the students once a while to check malpractices. Such movements should not disturb the students.

He must be vigilant. Reading novels or newspapers, making of calls and listening to telephone calls, dozing off and chatting are not allowed.

The timing of tests is very important. Tests must not be given immediately before or just after a long vacation, holidays or other important events where students are involved either physically or psychologically. Tests must also not be given when students would normally be doing something pleasant such as having lunch, athletics or other sporting activities as this will hamper students' concentration (Amedahe & Asamoah-Gyimah, 2003; Etsey, 2004). Interruptions during testing, such as giving instruction, must be kept to the barest minimum and should always relate to the test. The time spent and time left to complete the test must be announced at regular intervals to enable students apportion their time to the test items. Where practicable, the time should be written on the chalkboard at 30-minutes intervals until near the end of the test when it could be changed every five minutes. Further, students should start the test promptly and stop on time (Amedahe & Asamoah-Gyimah, 2003; Etsey, 2004).

Teachers should always work at minimising test anxiety in students during testing. They should therefore, avoid, warning students to do their best because the test is important, telling students that they must work faster in order to finish on time, threatening dire consequences of failure in the test, and threatening students with tests if they do not behave (Amedahe & Asamoah-Gyimah, 2003; Etsey, 2004).

## **Guidelines in Administering Achievement Tests**

According to Etsey (2005), in administering test items, classroom teachers are to consider, the following information which are essential in for maximising students' performance.

1. Students must be made aware of the rules and regulations covering the conduct of the test. Penalties for malpractice such as cheating should be clearly spelt out and clearly adhered to.
2. Avoid giving tests immediately before or after a long vacation, holidays or other important events where all students are actively involved physically or psychologically/emotionally.
3. Avoid giving tests when students would normally be doing something pleasant e.g. having lunch etc.
4. The sitting arrangement must allow enough space so that pupils will not copy each other's work.
5. Adequate ventilation and lighting is expected in the testing room.
6. Provision must be made for extra answer sheets and writing materials.
7. Pupils should start the test promptly and stop on time.

## **Scoring of Classroom Achievement Essay Tests**

According to Etsey (2004), essay tests can be scored by using the analytic scoring rubrics (also known as the point-score method) or holistic scoring rubrics (also called global-quality scaling or rating method). In analytic scoring, the main elements of the ideal answer are identified and points awarded to each element. This works best on restricted response essays. In holistic scoring, the model answer serves as a standard. Each response is read for a general impression of its adequacy as compared to the standard. The general

impression is then transformed into a numerical score. To check the consistency of the scoring, a first reading is done to sort the responses into several piles (mostly five A, B, C, D, E) according to the different levels of quality. The analytic, point-score or the trait method basically involves the use in scoring of an already prepared list of points or ideas considered essential to a good answer to the question, together with the number of points (marks) allotted to each idea raised or discussed in the answer (Nitko, 2001; Mehrens & Lehmann, 2001). This is known as a marking scheme, a scoring rubric or a scoring key, (Amedahe & Asamoah-Gyimah, 2003; Etsey, 2004; Tamakloe et al., 2006).

The holistic scoring rubric requires the marker to make judgement about the overall quality of each student's response. Teachers do not mark each specific content elements that student include in the answer. According to Nitko (2001), "the holistic scoring is probably more appropriate for extended response essays involving a student's abilities to synthesize and create and when no definite correct answer can be prespecified" (p. 195). The holistic method is less objective than the analytic method unless one has specified scoring criteria.

The scoring of essay-type tests according to Etsey (2004), is a highly important issue due to the fact that no matter how careful one is in writing the items, without equally taking careful steps to ensure consistency of scoring, the scores will not be reliable. The main reason for utmost care in the scoring of essay-type tests is the subjectivity involved. This is a major difference between the essay- and objective-type tests (Amedahe & Asamoah-Gyimah, 2003; Etsey, 2004; Gronlund, 2008). According to Mehrens and Lehmann (2001), the decision on a method of scoring for essay-type tests depends to some extent on the type of score interpretation desired (norm-referenced or criterion-

referenced) and the amount of diagnostic information needed about individual's responses. It also depends on the time and facilities available for reading the papers and whether the essay is of the restricted- or extended response type.

In order to improve objectivity in the scoring and reliability of the scores of essay-type tests, Mehrens and Lehmann (2001); Amedahe and Asamoah-Gyimah (2003); and Etsey (2004) have suggested the following techniques or principles to be adopted by scorers.

1. Prepare a form of scoring guide. This could either be an analytic scoring guide or a holistic scoring guide.
2. Constantly follow the marking scheme when scoring. It is one thing deciding to score all papers uniformly using a scoring guide and actually following the scoring guide constantly to achieve uniformity. Scorers should follow the marking scheme constantly as they score, as this reduces rater drift, which is the likelihood of either not paying attention to the scoring guide or interpreting it differently as time passes.
3. Scorers must also avoid being influenced by the first few papers they score since this can let them become too lenient or harsh in scoring other papers.
4. Score all responses item by item rather than script by script. Here, scorers must take one item at a time and score all the responses to it throughout before going to the next item. This principle is to minimise the carryover effect on the scores and thereby ensure consistency.
5. Randomly reshuffle the scripts when beginning to score each set of items. This will minimise the bias introduced as a result of the position of one's script. Research by Hales and Tokar (cited in Mehrens and

Lehmann, 2001) has shown that a student's essay grade will be influenced by the position of the paper, especially if the preceding answers were either very good or very poor. Mehrens and Lehmann (2001) have pointed out that randomly reshuffling of scripts is especially significant when teachers are working with high- and low level classes and read the best scripts first or last.

6. Score the scripts anonymously. Scripts should be identified by code numbers or any other means instead of the names of students. This principle is to reduce the halo-effect. This happens when a scorer's general impression of a person influences how the paper is scored.
7. Keep previously scored items out of sight when scoring the rest of the items. This principle is to minimise the carryover effects and ensure consistency of the scores.
8. Try to score all responses to a particular item without interruption. This is to avoid unreliability of the scores as a result of the grader's standards varying markedly due to excessive interruptions in the course of scoring.
9. Score essay-type tests only when you are physically sound and mentally alert. This is to say that essays must be scored at a congenial time. This is because it is known that consistency in scoring essay tests is a function of the time the paper is scored (Karpicke & Roediger, 2008). Over excitement, depression, and any type of psychological or mental disequilibrium will affect the consistency of the scores of essay-type tests.
10. Comments should be provided and errors corrected on the answer scripts for students to facilitate learning. This is especially important in

formative assessments where the comments should be on students 'weaknesses and strengths' in answering various items.

11. The mechanics of expressions such as correct grammar usage, flow of expression, quality of handwriting, orderly presentation of material and spelling should be judged separately from subject matter correctness.

### **Empirical Review**

Challenges in testing practices have been an issue across countries. In England and the United States, a number of researches have proven that teachers had challenges in applying basic principles in their testing practices. It is evident in literature that a lot of studies have not been conducted in the range of research into testing practices in Ghana. The few available ones include that of Amedahe (1989) on the testing practices of secondary school teachers in the Central Region of Ghana, Quaigrain (1992) on teacher competence in the use of essay-type tests in the Western region of Ghana, Oduro (2008) on testing practices of SSS teachers of English Language, Core Mathematics and Integrated Science in the Ashanti Region and Anhwere (2009) on assessment practices among the tutors of Colleges of Education in Ghana. Findings of their results are discussed in the sub headings.

### **Test Construction Practices of Teachers in Ghana**

In the work of Amedahe (1989), it emerged that the problem of insufficient study in the field of classroom achievement testing appears to exist even in the advanced countries like the USA because the emphasis is rather laid on standardised testing. This assertion was established by the research of Gullickson and Ellwein (cited in Amedahe, 1989). The study of Amedahe (1989) indicated that to a great extent, secondary school teachers in the Central

Region of Ghana did not follow the basic suggested principles of classroom test construction. He further averred that there was no significant difference between the procedure used in constructing classroom achievement tests by teachers who received instruction in testing and those who did not, in terms of the accuracy of following prescribed test construction principles.

That notwithstanding, the findings of Anhwere (2009) also revealed similar findings that teacher training college tutors do not follow the basic principle of testing in the construction of teacher made test or classroom tests, and that they perceived the management of assessment in the colleges as a work load to their teaching activities. Anhwere further identified no significant difference in test construction, practices between teachers with respect to their teaching experience. Quaigrain (1992), also arrived at the conclusion that majority of the teachers in the study did planned their essay-type tests in advance. This finding does not support wholly the first finding of Amedahe that, to a large extent, teachers did not follow the basic prescribed principles in construction of classroom test. At the discrepancy of Anhwere and Amedahe's findings, Oduro (2008), identified that teachers followed test construction principles to an acceptable degree where 7 out of 10 principles were indicated to be followed by teachers at the Senior High Schools in the Ashanti Region of Ghana. Another finding of Quaigrain (1992) was that, while some teachers reviewed their essay-type tests items, others did not review them. He found also that majority of the teachers did not indicate the score points which each item attracted on the question paper to guide students. These findings of Quaigrain generally support the first finding of Amedahe that to a great extent, the teachers



in his study did not follow the basic prescribed principles of classroom test construction.

Moving on experience on the job as contributing to competence, the finding of Quaigrain (1992) gave ample evidence that there was no evidence to support any positive relationship between years of teaching and one's competence in the use of essay-type tests. The finding here is also at discrepancy with the third finding of Amedahe that reported a moderate relationship between number of years of teaching and the accuracy with which teachers constructed their classroom achievement tests.

### **Test Administration Practices of Teachers in Ghana**

Under test administration, the result from Oduro (2008) indicated that teachers followed test administration principles to an appreciable level, 12 out of 18 principles were followed. With regards to Amedahe (1989), he discovered that teachers in the study mostly observed good physical and psychological conditions when administering their classroom achievement tests. This was a very good sign for classroom achievement test administration. Anhwere (2009) found that teachers followed test administration processes to an acceptable level and ensured good physical conditions when administering tests. The findings of Anhwere, Amedahe, and Oduro concur that teachers to a higher degree appreciate the principles of test administration and therefore follow them. Anhwere however did not identified any significant difference in test administration practices with respect to years of teaching and how principles of test administration are followed by teachers.

### **Test Scoring Practices of Teachers in Ghana**

On tests scoring, Anhwere (2009) identified no significant difference in test scoring practices among teachers with respect to their years of teaching. Oduro (2008) recounted that, teachers applied six out of nine principles in test scoring. In addition, teachers who had training in testing indicated that they used two out of nine principles more frequently than their counterparts who had no training in testing. Quaigrain (1992) also found out that majority of the teachers in the study prepared their marking scheme after the examination while few prepared their marking scheme before the test was taken. Amedahe (1989) recounted that teachers in the schools under his study used mainly the analytic method in scoring their essay-type tests. Furthermore, teachers in the schools scored their essay-type tests either item by item or script by script. On the part of Quaigrain (1992), he found that majority of teachers in the schools used the analytic method in scoring their essay-type tests. Also, almost half of the teachers scored their essay-type tests item by item while the other half scored them script by script. The analytic method of scoring seems to be very popular with classroom teachers and this may be attributed to the numerous advantages it holds over the holistic method of scoring, especially in formative testing. A comparison of the two studies reviewed above reveals quite similar findings. It could, therefore, be concluded that for a period of three years from 1989 to 1992, Quaigrain's study came to confirm the findings of Amedahe's study to a large extent.

### **Testing Practices of Teachers in the United States of America**

The research findings on the testing practices of teachers in the USA are given below.

Marso and Pigge (1989) conducted a study in Ohio on testing skills and practices of 326 elementary and secondary school teachers. The assessments included direct analysis of teacher made tests as well as perceptual assessments of teachers' testing needs and proficiencies. The findings from their study suggest that teachers were not proficient in their testing practices. Further analyses revealed that a number of 50 or more teacher-made tests were given by teachers within a year. Matching exercises on teacher made tests were particularly prone to error. Most teacher-made tests, except in mathematics and science, functioned at the knowledge level. Administrators' and teachers' perceptual assessments of teachers' testing skills were negatively correlated.

An estimate of 225 studies addressing the knowledge and skills of classroom teachers from kindergarten through grade 12 related to the development and use of teacher-made tests were reviewed by Marso and Pigge (1989). Their findings emerged that, teachers were limited in expertise in testing. Pre-service training in testing was not also organised for teachers. Most teacher-constructed tests also contained many faults, and functioned almost exclusively at the recall level. Teachers typically did not use test improvement strategies such as test blueprints and item analysis.

Another report by the American Association for the Advancement of Science (AAAS, 1998) in describing the current assessment practices in the United States of America pointed out that, teacher-made tests are often limited in measuring student thinking as their standardised counterparts Stiggins and Conklin (as cited in AAAS, 1998). It was found that, teacher-made tests were mostly short-answer or matching items that place far more emphasis on students' recall than on students' thinking ability. They further found that,

because teachers did not receive proper training in effective assessment methods, they did not change the assessment methods they use as assessment needs change. They stipulated that different assessments are needed to measure performance, effort and achievement, for instance, but teachers tended to use the same type of assessment, mainly tests, to measure all three. It also emerged that, due to limited time, teachers usually use the assessments that are found at the end of textbook chapters. They further explained that assessment from books, included mostly short answer questions which only measured the low level thinking skills and simple recall of factual knowledge (Centre for the Study of Testing, Evaluation & Educational Policy [CSTEED] (as cited in AAAS, 1998). The report further stated that even if teachers received the training, time and resources that would allow them to widen their assessment practices, students themselves would have become a barrier. Students, especially high school students had become test-wise. This sometimes make them oppose the more labour intensive format of assessments that entail performance tasks, answering essay-type items or providing possible solutions to open-ended items.

The next was a study to assess teacher-made tests in science and mathematics classrooms in the High Schools. The study was undertaken by Oescher and Kirby (1998) and published by the Educational Resources Information Centre (ERIC). The study covered the nature of classroom assessment, characteristics of teacher-made tests, item format, cognitive levels, quality of test items and teachers' confidence in testing skills. The results of the study indicated that the main areas where teachers lacked competence were the

use of tables of specifications, development of higher order items, item formatting, and empirical analysis of test results.

The apparent reasons for the outcome of the study, according to Stiggins (1999), were that generally, teachers in the USA are not very well prepared. Only a handful of States require competency in assessment as a condition for licensure. Even more troubling is that only three States require competence in assessment for principal certification. He concluded that majority of practicing teachers and administrators in the USA have not had the opportunity to develop the assessment literacy they need as professionals.

### **Testing Practices of Teachers in England**

In England, a review of research findings on a number of published studies by Crooks (1998) and Black (1993) suggest the existence of weak testing practices. The findings suggest that classroom assessment practices encouraged rote learning. They explained that, items were centred on low level of cognitive processing. Hence encouraged rote learning. Also teachers were in a haste when assessing learning such that they do not review constructed test items. Thirdly, teachers over-emphasised the grading function while the learning function was under emphasised. Fourth, there was a tendency on the part of teachers to use normative rather than a criterion approach which emphasised competition between pupils rather than personal improvement of each student. The evidence is that with such practices, the effect of feedback is to teach the weaker pupils that they lack ability, so that they are de-motivated and lose confidence in their own capacity to learn.

According to Black and William (1998), more recent researches have confirmed this general picture. Both in questioning and written work, teachers'

assessments focused on low-level aims, mostly recall. According to Bol and Strage, Pijl, Senk, Beckman and Thompson, (as cited in Black & William, 1998), there has been little focus on such outcomes as speculation and critical reflection. Duschl and Gitomer (as cited in Black & William, 1998) had observed students focus on getting through the tasks and oppose attempts to engage in risky cognitive activities. According to Lorsbach, Tobin, Briscoe and Lamaster (as cited in Black & William 1998), although teachers can foretell the performance of their students on external tests, their own tests do not tell them what they must know about their students' learning.

### **Chapter Summary**

Studies in United States and, England revealed that teachers lacked competences in their testing practices. In the case of Ghana studies have shown discrepancies with respect to particular testing principles that teachers adhere to. Findings from all the studies gave ample evidence to conclude that, in terms of test administration, teachers possess some potentials. However, with respect to test construction and scoring, studies have shown that teachers lacked appreciable competence.

## **CHAPTER THREE**

### **RESEARCH METHODS**

#### **Introduction**

The main rationale for the study was to find out the basic test principles teachers at the Junior High Schools in Cape Coast Metropolis follow in their testing practices with respect to tests construction, administration and Scoring of essay-type tests. The chapter discusses how the study was conducted. It is presented in five sections. The first section deals with research design, and the second deals with the population and sampling procedure. The third section covered the research instrument (including pretesting that was used, validity and reliability of the instruments) while the fourth section deals with data collection procedure. The last section covers how data collected was analysed.

#### **Research Design**

According to De Vos (1998), a research design is a blueprint or a detailed plan of how a research study is conducted. Polit and Beck (2004) indicated that selecting a good research design should be guided by an overarching consideration, namely whether the design does the best possible job of providing trustworthy answers to the research questions.

With this study, descriptive survey research design was used to collect data so that inferences could be made about some characteristics, attitudes or behaviour of the population. Osuala (2001) noted that “descriptive surveys are versatile and practical, especially to the researcher in that they identify present

needs” (p. 35). Descriptive research involves collecting data in order to test hypothesis or answer questions concerning the current status of the subjects of the study. It determines and reports the way things are (Gay, 1992).

Frankel and Wallen (2000) stated that “obtaining answers from a large group of people to a set of carefully designed and administered question, lies at the heart of survey research” (p. 431). According to Polit and Hungler (1999), descriptive study aims at describing, observing and documenting aspects of a situation as it naturally occurs rather than explaining them. A descriptive study provides a more accurate picture of events and seeks to explain people’s perception and behaviour on the basis of data gathered at a point. With respect to this study, the design would explain the prevalent testing practices of teachers at the Junior High Schools in its current situation. The design is appropriate when a researcher attempts to describe some aspects of a population by selecting unbiased samples who are asked to complete questionnaires (Frankel & Wallen, 1993). A major advantage of a descriptive survey is that it determines and reports the way things are (Gay, 1992). McMillan and Schumacher (2006) postulated that apart from being the most commonly used method in educational research, the descriptive design is preferred, because it is objective in data collection at a point in time, it quantifies variables, describes the phenomena using numbers to characterize them.

Irrespective of the strengths of the descriptive survey mentioned earlier, Fraenkel and Wallen (2000) have identified some weaknesses of the descriptive survey. Firstly, they stipulated that descriptive survey has difficulty in ensuring that the questions to be answered are clear and not misleading. Secondly they claim, getting respondents to answer questions thoughtfully and honestly is a



setback, and then added that, getting a sufficient number of questionnaires completed and returned so that meaningful analysis can be made is also a setback as far as descriptive survey is concerned. Osuola (2001) in buttressing the points on the weaknesses of the descriptive research stipulated that, designing a quality investigation requires particular attention to two central factors: appropriate sampling procedures, and precision in defining terms in eliciting information. He continued by adding that, while descriptive research is a prerequisite for finding answers to questions, it is not in itself sufficiently comprehensive to provide answers and that it cannot also provide cause-and-effect relationships.

Despite the setbacks of the descriptive design, it was deemed necessary for the study because judging from the main study; it was the most appropriate design which could help to evaluate the testing practices of Junior High School teachers in terms of their test construction, administration and scoring.

### **Study Area**

Cape Coast Metropolis is one of the six Metropolis among 216 district in Ghana and the only Metro among the 20 districts in the Central Region. The Cape Coast Metropolitan Assembly was formerly known as Cape Coast Municipal. It was raised to the status of municipality in 1987 by LI 1373 and upgraded to metropolitan status in 2007 by LI 1927. The Metropolis is bounded to the South by the Gulf of Guinea, to the West by the Komenda Edina Eguafo Abrem Municipality (at Iture bridge), to the East by the Abura Asebu Kwamankese District, and to the North by the Twifu Heman Lower Denkyira District. It is located on longitude 1° 15'W and latitude 5°06'N. It occupies an Area of approximately 122 square kilometres, with the farthest point at

Brabedze located about 17 kilometres from Cape Coast, the Central Regional capital of Ghana. The population of the Cape Coast Metropolis, according to the 2010 Population and Housing Census, is 169,894 representing 7.7 percent of the region's total population. Cape Coast is endowed with many schools across the length and breadth of the Metropolis, ranging from basic to tertiary institutions. These schools attract people from all over the country and the West Africa Sub-region, who pursue various levels of academic and professional education (Ghana Statistical Service, 2013).

### **Population**

Polit and Beck (2004) defined a population as the entire aggregation of cases that meet a designated set of criteria. The large group to which the researcher wishes to generalize the results of the study becomes the targeted population (Ary, Jacobs, Sorensen, & Razavieh, 2010). According to Amedahe (2000), target population refers to the population that the researcher will ideally like to generalise.

The target population for the study was made up of 1244 teachers of the 79 public Junior High School teachers in Cape Coast Metro. The accessible population was made up of teachers in the 50 selected schools from the five circuits of the Metropolis. The accessible population was made up 300 teachers in the selected 50 public Junior High Schools in the Cape Coast Metropolis.

### **Sampling Procedure**

According to Amedahe (2000), sampling involves the process of selecting a portion of the population to represent the entire population. Sarantakos (2005) postulates that, a sample enables the researcher to study a relatively smaller number of units in place of the target population and to obtain

data that are representative of the target population. There were 79 public Junior High School in Cape Coast Metro which has been grouped into six circuits. There are 1,244 teachers in the 79 public schools in Cape Coast Metro. The first circuit has 12 schools, the second has 16 schools, third has 14, fourth 13 and fifth and the sixth circuit has 10 and 14 schools respectively. Table 1 gives details about the number of circuit and schools with their total number of teachers in the Cape Coast Metro.

Table 1-*Frequency Distribution of Circuits, Public Schools, and number of Teachers in Cape Coast Metro*

Name of Circuit	Number of Schools	Number of Teachers
Cape Coast Circuit	12	222
Aboom Circuit	16	247
Bakaano Circuit	14	211
Pedu/Abura Circuit	13	234
OLA Circuit	10	138
Efutu Circuit	14	192
TOTAL	79	1244

Source: Cape Coast Metro Education Office (2017)

According to Krejcie & Morgan (1970), a population size of 1200 requires a sample size of 291; therefore, a sample of 300 from 1244 was considered for the study. A total of three hundred (300) teachers were selected from fifty (50) Junior High Schools within five circuits of the Cape Coast Metropolis for the study. These constitute 24.2% of the accessible population. According to Amedahe (2002), in most quantitative studies, a sample size of

5% to 20% of the population size is sufficient for generalization purposes, with respect to the size of the population.

The number of schools have been grouped into six circuits in the Metropolis. Each circuit constituted a cluster. The sampling therefore went through three-stage sampling. The first stage involves randomly selecting five (5) out of the six (6) circuits using the lottery method.

In order to obtain a representative sample for the study, I selected ten (10) schools each from the five selected circuits using a table of random numbers. This is because each circuit of the Cape Coast Metro has not less than (10) schools. The researcher purposively selected four teachers from the four core subject areas (English, Mathematics, Science and Social Studies) and randomly selected two other teachers irrespective of the subject taught in the Junior High Schools. The number of the schools were coded.

### **Data Collection Instruments**

Questionnaire was used as the data collection instrument. The items on the questionnaire were carefully selected and constructed from the reviewed literature. I developed the questionnaire with the help of my supervisors. A questionnaire is a set of questions or statements given to a group of people in a written form to illicit their responses. The questions are mainly related to the research questions.

According to Cohen, Manion and Morrison (2000), questionnaire is widely used and is a useful instrument for collecting survey information, providing structured, numerical data and can able to be administered without the presence of the researcher. The choice of a questionnaire is based on the assertion of Osuola (2001) that, they are particularly advantageous whenever

the sample size is large enough to make it uneconomical for reasons of time or funds to observe or interview every subject.

Questionnaire is considered appropriate because it ensures anonymity of respondents and helps to complete the study within a relatively short period. Knowles (2000) also asserts that questionnaires are easy to administer, friendly to complete and fast to score and therefore take relatively less time from researchers and respondents. The questionnaires also provide relatively straightforward information to analyse (Cohen, Morrison & Manion, 2000). The questionnaires further ensure a wider coverage of the construct being measured. The researcher will also be able to approach respondents more easily.

Regardless of the strength of a questionnaire, it has a low response rate and also response biases are more likely to occur (Creswell, 2012). Kothari (2004) also gave a number of disadvantages of questionnaire. Among them includes, “1. Low rate of return of the duly filled in questionnaires, 2. Bias due to no-response is often indeterminate, 3. It can be used only when respondents are educated and cooperating, 4. The control over questionnaire may be lost once it is sent and 5. There is inbuilt inflexibility because of the difficulty of amending the approach once questionnaires have been despatched” (p. 101).

The questionnaire was divided into five sections which comprises of section A, B, C, D, and E (See Appendix A). The items were mainly close ended questions. Section A consisted of the personal data and section B addresses the test format teachers use in assessing their students. Section, C addresses factors to consider in construction of achievement test based on standards in test construction. This section seeks to find out the teacher’s knowledge in test construction, the kind of test they use in assessing their students, the kind of

processes teachers go through in their testing practices, and the rationale behind the use of test and particular test format they use. Section D addresses how tests are administered, and Section E addresses ways by which tests are scored. This section basically looked into the conditions under which Junior High School teachers in Cape Coast Metropolis score essay-type tests, and factors that influence teachers in scoring their essay-type test items.

The scale scores ranged from five (5) for -Always to zero (0) for -Never. The Likert type scale was chosen because according to Asamoah-Gyimah (2002), in measuring the views and impressions of teachers on an on-going practice, it is the simplest, but equally efficient approach when considered alongside with social-distance scales, Thurstone scales and the scalogram analysis. It was adopted also to ensure effective analysis of the data even though it restricts free expression and perception of respondents in a study.

### **Validity of the Instrument**

In order to ensure the validity of the instrument, the questionnaire was given to my supervisors in the Department of Education and Psychology in the University of Cape Coast (UCC) for expert judgement and assessment. This was to ensure that the developed items were related to the construct. The items were pretested and results was given to the supervisors to examine whether they were statistically appropriate to be used. The supervisors also checked to find out that the items were related to the research questions and also comprehensively cover the details of the study. The recommendations and suggestions from the supervisors were considered to fine tune the questionnaire for the purpose of the study.

### **Trial-testing of the Instrument**

The questionnaire was pre-tested in ten public JHS in the Komenda Edina Eguafo Abrem district (K.E.E.A) in the Central Region of Ghana. The respondents were fifty (50) teachers from ten (10) schools in K.E.E.A Municipality. These schools included Ntranoa D/A Junior High School “A”, Ntranoa D/A Junior High School “B”, Abee A.M.E Zion Junior High School, Simiw D/A Junior High School, Mental D/A Junior High School, Nkontrodo A.M.E Zion Junior High School, Elmina Methodist JHS “A” and “B”, Amisano Catholic JHS and Atronkwa D/A JHS. The location and teachers of these schools have similar characteristics with respect to socio-cultural background with that of schools in the Cape Coast Metro in the Central Region of Ghana. Respondents were allowed to ask questions relating to clarity of the questionnaire items. Blank sheets of papers were added to the questionnaire for respondents to express their views in writing on the clarity, ambiguity, biases, inconsistencies and problems in all aspects of the questionnaire. This helped to reduce respondents’ biases and prejudices (Trochim, 2000). Fifty teachers were involved in the pre-testing exercise.

The goal for the pilot test consisted of two parts. The first was to find as many as possible, practical arrangements that might have a negative influence on the success of the research procedure (feasibility of the study). The others include sorting out all practicalities related to measurement instruments as well as the applicability of these instruments to the potential outcomes of the study.

Feedback from the pre-testing was analysed to ensure content validity and helped me to revise items which looked ambiguous and discard items which did not measure the construct it wanted to. The pretesting gave me the

opportunity to assess the validity, the appropriateness and the practicality of the instrument. My attention was drawn to items which looked ambiguous, and items with grammatical errors were corrected. The final instrument for the study was produced after subsequent revisions in the wording of the items. The necessary corrections were effected after the pre-testing.

### **Reliability of the Instrument**

Reliability focuses on the degree to which empirical or measures of a theoretical concept are stable or consistent across two or more attempts to measure the concept (Kothari, 2004). Reliability is a measure of the consistency over time of instruments with groups of respondents and it deals with precision and accuracy (Cohen et al., 2000). Validity and reliability are essential features of any research (Creswell, 2003; Robson, 2002). To obtain the reliability of the instrument, Cronbach ‘s co-efficient alpha was used to estimate the internal consistency. Table 2 provides the summary of the reliability coefficient obtained for each of the study variables.

Table 2- *Summary of the Reliability Coefficient of the Items*

Reliability of the Items	Reliability Coefficient	Number of Items
Item format used	.258	5
Construction of achievement test	.838	20
Administration practices	.581	12
Scoring practices	.802	13
Overall Reliability	.873	50

Source: Field survey, Sasu (2017)

The overall Cronbach’s co-efficient alpha that was obtained for the study was .873, (See Appendix F). According to Fraenkel and Wallen (2000),



the reliability coefficient should be at least 0.70 and preferably higher. Therefore, the reliability obtained on the variables was justifiable for the study.

### **Ethical Consideration**

It is of paramount importance to consider ethical issues for the success of every research. In this research several ethical issues were taken into consideration. The ethical concerns which were addressed were informed consent, anonymity and confidentiality. Informed consent affords prospective participants the opportunity to accept or decline to engage in the research (See Appendix E). It describes the need for participants to understand the aims, objectives and potential harm that their involvement may have on them (Seidman, 2006). In this study, the purpose of the study was carefully reviewed with the participants before they were involved in the study.

Anonymity of study respondents was also highly taken into consideration in the present study. Oliver (2010) pointed out that anonymity is a vital issue in research ethics because it gives the participants the opportunity to have their identity concealed. In this study, codes were adopted instead of names of respondents to ensure anonymity of information. This was to prevent possible victimization of respondents where certain responses may be viewed as unpalatable to other stakeholders. A detailed explanation for the purpose of the study was given to the respondents before involving them in this study.

On the issue of confidentiality, effort was made to ensure confidentiality of the responses from the respondents. Participants were told that their responses would be kept confidential and that no third party known to them would have access to the information provided. Most essentially on the ethical issues, pieces of information that were cited from earlier studies on testing

practices of teachers and the kind of principles teachers follow to support the review of related literature were duly acknowledged through both citation and referencing in order to avoid academic dishonesty otherwise known as plagiarism.

### **Data Collection Procedures**

A letter of introduction was obtained from the Department of Education and Psychology (See Appendix C). The letter spelt out the purpose of the study, the need for individual participation, anonymity as well as confidentiality of respondents' response. I then requested a permission from the Cape Coast Metro Educational Directorate which gave the participating schools confident to participate freely (See Appendix D). I also established the necessary contacts with the head teachers of the selected schools to seek permission to administer the questionnaire.

Consent was sought from participating teachers before including them in the data collection. This was achieved by explaining the purpose of the study to them and giving them an informed consent form to fill, indicating their willingness to be involved in the research (See Appendix E). Respondents were made to be aware that they have the right to discontinue the process at any time they feel responding to certain questions infringe upon their privacy or right.

Respondents were informed that their names and other demographic characteristics which identify them personally would not be disclosed with information they provide to me, ensuring anonymity of data provided. I also allowed the respondents to ask questions for clarity in the items if the need be. They were allowed to seek detailed explanation of items that seemed ambiguous to them. Interaction with the teachers helped to build cordial relationship

between me and the teachers which ensured the commitment of the teachers to answer the items on the form and submit them at the appropriate time.

The questionnaire was administered to the teachers at one sitting with the help of the head teachers. Most of the respondents used a maximum of 20 minute to respond to the items on the questionnaires. For the sake of anonymity, the teachers were asked not to write their names as well as their school name on the questionnaire.

### **Data Processing and Analysis**

The responses to the questionnaires was first edited, typed, coded and scored. The editing procedure was to check whether respondents had followed directions correctly, and whether all items have been responded to. After editing and coding, the data was entered into the computer using the Statistical Package for the Social Sciences (SPSS version 22.0) software.

The demographic variables from the questionnaire were primarily analysed using frequencies and percentages. The frequencies and percentages were based on the teacher's qualification, form taught, years of teaching, and subject taught by the Junior High School Teachers. The second section of the questionnaire was analysed based on the research questions set for the study using inferential statistics (One-Way Analysis of Variance) and descriptive statistics (means and standard deviations). The research questions were analysed and presented systematically.

### **Research Question One**

What principles do Junior High School teachers follow in constructing test items?

Responses to items on this research question were analysed using means and standard deviations, as well as frequencies on the response from the Junior High School teachers. The questionnaires were responded to using four-point Likert scale from “not often” (scored 1) to “always” (scored 4). An item mean of 2.5 was established as the criterion measure of which individual item mean that falls above, represents principles that is often employed by teachers and those that fall below indicate that teachers did not follow that particular test construction principle to an appreciable level. To find the value of the criterion measure, the scores on the four point Likert scale were added together and divided by the total number of scale. That is  $4+3+2+1=10/4=2.50$ .

The overall analysis of the items on test construction practices was analysed using the one-sample t-test at  $p=0.05$  (2-tailed) level of significance. The items were 20 in number with an overall mean of 50 ( $2.5*20$ ). A mean of 50 and above indicates teachers followed the overall test construction principles to an appreciable level and a mean below 50 indicates that teachers did not follow the basic principles in test construction.

### **Research Question Two**

What process do Junior High School teachers follow in administering tests?

Responses to items on this research question were analysed using frequencies, means and standard deviations on the response from the Junior High School teachers. The questionnaires were responded to using four-point Likert scale from “not often” (scored 1) to “always” (scored 4). A test value of 2.5 was established as the criterion measure of which individual item means that fall above represented principles that are often employed by teachers and those that fall on 2.49 and below indicated that teachers did not follow the particular

test administration principle to an appreciable level. To find the value of the criterion measure, the scores on the four point Likert scale were added together and divided by the total number of scale. That is  $4+3+2+1=10/4=2.50$ .

The overall analysis of the items on test administration practices was analysed using the one-sample t-test at  $p=0.05$  (2-tailed) level of significance. The items were 12 in number with an overall mean of 30 ( $2.5*12$ ). A mean of 30 and above indicated teachers followed the overall principles of test administration to an appreciable level and a mean below 30 indicated that teachers did not follow the basic principles in test administration.

### **Research Question Three**

What basic principle of test scoring do Junior High School teachers follow in scoring of essay-type test?

Responses to items on this research question were analysed using frequencies, means and standard deviations on the response from the Junior High School teachers. The questionnaires were responded to using four-point Likert scale from “not often” (scored 1) to “always” (scored 4). A test value of 2.5 was established as the criterion measure of which individual item means that fall above represented principles that are often employed by teachers and those that fall on 2.49 and below indicated that teachers did not follow the particular test scoring principle to an appreciable level. To find the value of the criterion measure, the scores on the four point Likert scale were added together and divided by the total number of scale. That is  $4+3+2+1=10/4=2.50$ .

The overall analysis of the items on test scoring practices was analysed using the one-sample t-test at  $p=0.05$  (2-tailed) level of significance. They were 13 items in number with an overall mean of 32.5 ( $2.5*13$ ). A mean of 32.5 and

above indicated teachers followed the overall test scoring principles to an appreciable level and a mean below 32.5 indicated that teachers did not follow the basic principles in test scoring.

### **Research Hypothesis One**

H<sub>0</sub>: There is no statistically significant differences among teachers who have taught for the following number of years in their test construction practices; 1-5, 6-10, 11-15, 16 years and above.

To find out whether there is statistically significant differences among teachers who have taught up to the following number of years in their test construction practices; 1-5, 6-10, 11-15, 16 years and above, responses to items on this questionnaire were analysed using the One-Way Analysis of Variance (ANOVA) for equality of means at level of significant  $p= 0.05$  (two-tailed).

### **Research Hypothesis Two**

H<sub>0</sub>: There is no statistically significant differences among teachers who have taught for the following number of years in their test administration practices; 1-5, 6-10, 11-15, 16 years and above.

To find out whether there is statistically significant differences among teachers who have taught up to the following number of years in their test administration practices; 1-5, 6-10, 11-15, 16 years and above, responses to items on this questionnaire were analysed using the One-Way Analysis of Variance (ANOVA) at level of significant  $p = 0.05$  (two-tailed).

### **Research Hypothesis Three**

H<sub>0</sub>: There is no statistically significant differences in scoring practices among teachers who have taught for the following number of years; 1-5, 6-10, 11-15, 16 years and above.

The responses to items on this questionnaire were analysed using the One-Way Analysis of Variance (ANOVA) at level of significant  $p = 0.05$  (two-tailed), to find out whether there were any statistically significant differences in scoring of essay-type test among teachers of different level of teaching periods.

#### **Research Hypothesis Four**

H<sub>0</sub>: There is no statistically significant differences in test construction practices, among teachers in the four core subject areas (Mathematics, Science, English and Social studies).

The hypothesis was tested using One-way Analysis of Variance (ANOVA) at level of significant  $p = 0.05$  (two-tailed) to find out whether there were statistically significant differences among teachers in the four core subject areas (Mathematics, Science, English and Social studies) in test construction practices.

#### **Research Hypothesis Five**

H<sub>0</sub>: There is no statistically significant differences in test administration practices, among teachers in the four core subject areas (Mathematics, Science, English and Social studies).

The hypothesis was tested using One-way Analysis of Variance (ANOVA) at level of significant  $P = 0.05$  (two-tailed) to find out whether there were statistically significant differences in test administration practices, among teachers in the four core subject areas (Mathematics, Science, English and Social studies).

### **Research Hypothesis Six**

H<sub>0</sub>: There is no statistically significant differences in test scoring practices, among teachers in the essay-type scored subject areas (Religious and Moral Education, English and Social studies).

The hypothesis was tested using One- way Analysis of Variance (ANOVA) at level of significant  $p = 0.05$  (two-tailed) to find out whether there were statistically significant differences in test scoring, among teachers in the essay-type scored subject areas (Religious and Moral Education (RME), English and Social studies).

### **Chapter Summery**

The research used the quantitative method. Descriptive survey design was employed for the study. The sample comprised of 300 respondents in 50 selected Junior High Schools in the Cape Coast Metropolis. Data on the testing practices of teachers in the Junior High Schools was collected and generated using the survey questionnaires. Data was analysed using the One-Way Analysis of Variance (ANOVA) at  $p = 0.05$  (two tailed) level of significance and descriptive statistics (means, standard deviation) as well as frequencies and percentages. Methodologically, the study was limited by the design for the fact that the design only reported the prevalence situation at the time of the data collection.



## **CHAPTER FOUR**

### **RESULTS AND DISCUSSION**

#### **Introduction**

This chapter presents an analyses of the data gathered from the field in relation to test practices of teachers in the Cape Coast Metro. The study aimed at finding out whether Junior High School teachers follow the basic prescribed principles in the construction, administration and scoring of classroom achievement tests. This chapter again presents the results of the analysed data collected from the respondents. The data were analysed through frequencies, percentages, means and standard deviations, and One-Way Analyses of Variance (ANOVA). The results were presented with discussions.

#### **Analysis of Demographic Variables**

Data gathered on respondents' characteristics were centred on respondent's qualification, class of teaching, years of teaching and subject taught. Respondents' characteristics are presented in Tables 3, 4, 5 and 6.

Item 1 on the questionnaire elicited responses on respondents' qualification. This is presented in Table 3.

Table 3- *Respondents' Qualification*

Qualification	Frequency	Percent (%)
Diploma	79	26.3
Bachelor of Education	173	57.7
Bachelor of Science/Art	15	5.0
Masters	33	11.06
Total	300	100

Source: Field Survey, Sasu (2017)

From Table 3, 173 representing 57.7% of respondents hold Bachelor of Education as their highest qualification. Also, 79 (26.3%) of respondents hold Diploma. Out of the 300 teachers, 11.06% hold Masters Certificate while 5% of respondents hold bachelor of Science/Art as their highest qualification.

Table 4 shows the Form that teachers handle in their various schools.

Table 4 – *Form(s) Taught by Respondents*

Form	Frequency	Percent (%)
One	47	15.7
Two	22	7.3
Three	31	10.3
One and Two	15	5.0
Two and Three	21	7.0
One, Two and Three	164	54.7
Total	300	100

Source: Field Survey, Sasu (2017)

Table 4 indicates that majority of the respondents,  $15+21+164= 200$  (66.7%), taught more than one Form. Among those teaching more than one

Form, 164 (54.7%) teachers taught all three Forms. A total of 47 (15.7%) respondents taught Form One only, 22 (7.3%) taught Form Two only, and only 31 (10.3%) taught Form Three only. In sum, teachers teaching all the three Forms were the dominant group in the study.

Table 5 presents the number of teachers who have taught within the year ranges.

Table 5- *Years of Teaching of Respondents*

Years of teaching	Frequency	%
1-5	58	19.3
6-10	104	34.7
11-15	75	25.0
16 and above	63	21.0
Total	300	100

Source: Field Survey, Sasu (2017)

As indicated in Table 5, 104 representing 35% respondents have taught for 6-10 years. Also 75 (25.0%) respondents have taught for 11 to 15 years while 58 (19.3%) of respondents taught for 1-5 years. It can also be seen from the data in Table 5 that 63 (21.0%) of respondents have taught for 16 years and above. Majority of the respondents had taught for more than five years. This maybe as a result of the fact that Cape Coast is considered as an urban area which attracts more of the teachers who have served beyond their teaching bond of five years. Table 6 shows subjects and the number of teachers in the subject area.

Table 6- *Subject(s) Taught by Respondents*

Subject	Frequency	%
Social Studies	50	16.7
Religious and Moral Education	47	15.7
English	52	17.3
Mathematics	50	16.7
Science	47	15.7
Ghanaian Language (Fante)	17	5.7
French	4	1.3
Basic Design and Technology	12	4.0
I.C.T	21	7.0
Total	300	100

Source: Field Survey, Sasu (2017)

Table 6, indicates that majority of the respondents were core subject teachers representing 66.4% (16.7+17.3+16.7+15.7). Among the core subject teachers, 50 (16.7%) respondents were Mathematics teachers while 47 (15.7%) were Science teachers. Also, 52 representing 17.3% were English teachers while 50 (16.7%) were Social Studies teachers. Moreover, this table shows that 15.7% of teachers were teaching Religious and Moral Education.

### **Research Question One**

What principles do Junior High School teachers follow in constructing test items?

This research question sought to find out the kind of principles that Junior High School teachers followed in the construction of their classroom achievement tests. In addressing this research question, two major issues were

addressed. Teachers were to indicate the item format that they often used and their actual test construction practices. The first section consists of five (5) items which indicated the formats teachers often use. Items in Table 7 were responded to, using a four-point Likert scale with categories ranging from “not often” (scored 1) to “always” (scored 4). A cut-off points was established to determine item format that teachers often used. A mean of 2.50 and above indicates item format often used by teachers while a mean below 2.50 indicates item format that teachers do not often use. The results are presented in Table 7.

Table 7 -*Test Format used by Teachers*

Test format	N	Mean	Std.D
Essay	300	3.32	1.857
Multiple choice	300	3.04	0.792
Short answers	300	2.96	0.878
Matching	300	2.21	0.834
True/false	300	2.50	0.832

Source: Field Survey, Sasu (2017)

Table 7 indicates that teachers mostly used the two major types of test format essay and multiple choice, when assessing students in the Junior High Schools. The results in in Table 7 shows that Essay ( $M = 3.32$ ,  $SD = 1.86$ ) and Multiple choice format ( $M = 3.04$ ,  $SD = 0.79$ ) and Short answers format ( $M=2.96$ ,  $SD=0.873$ ) were often used by teachers when assessing students. However, matching item format ( $M = 2.21$ ,  $SD= 0.83$ ) was less frequently used by teachers when assessing students.

The result is consistent with previous study that suggest that classroom or teacher-made tests in the Senior High Schools in Central Region of Ghana

are mainly essay, short answer and multiple-choice types (Amedahe, 1989). The result further corroborates the assertion made by Bartels (2003) in his study on teachers' practices of continuous assessment in teacher Colleges of Education that the objective type tests are the most commonly used by teachers in Ghana. It is possible that teachers considering class sizes and the limited time to submit the results of the tests might prefer constructing an objective type of test. They might also be using the essay and the multiple choice as a means of following the format of students' final exams set by the WAEC.

The items on construction principles teachers followed were twenty (20). Teachers were given a four point Likert scale items to respond to. The scoring of items was based on the four point Likert scale of measurement ranging from "not often" (scored 1) to "always" (scored 4). For the purpose of clarity, items which indicate negative implications to test construction practices were reversed in coding which ranged from "not often" (scored 4) to "Always" (scored 1). A mean of 2.5 and above on these items means teachers do not practice while a mean below 2.5 means teachers often practice these principles. These items include: "I construct items only when it is time to assess students", "I copy questions from past questions", "I copy questions from text books", "I ask any other colleagues to help me construct items".

The results were ranked in descending order from principle teachers often followed to the principles teachers less often follow and has been presented in Table 8. The results were then discussed using means and standard deviation. Cut-off points using item mean of 2.5 was established to determine test construction practices that teachers often follow and the ones teachers less frequently follow. A mean of 2.50 and above indicates teachers practice test

construction principles to an appreciable level while a mean below 2.5 indicates teachers' do not practice the principles to an appreciable level. Table 8 presents the result of test construction principles that teachers often followed in descending order.

Table 8 – *Test construction practices of teachers*

Test construction practices.	N	Mean	Std. D	Rank
I provide clear and simple instructions on how test is to be answered	300	3.30	.791	1 <sup>st</sup>
I evaluate items given to the students to answer	300	3.04	.932	2 <sup>nd</sup>
I state the purpose of which the test will be used	300	2.81	1.108	3 <sup>rd</sup>
I follow the principles of test construction for each format	300	2.73	1.086	4 <sup>th</sup>
I consider the time individuals will spend on a question	300	2.63	1.142	5 <sup>th</sup>
I specify the construct to be measured	300	2.53	1.019	6 <sup>th</sup>
I do consider students language proficiency	300	2.52	1.203	7 <sup>th</sup>
I match learning outcomes to the items	300	2.45	1.092	8 <sup>th</sup>
I try solving the question myself to determine the time required for the questions	300	2.45	1.265	8 <sup>th</sup>
I copy questions from pupils text books	300	2.34	1.096	10 <sup>th</sup>
I prepare marking scheme after the students have answered the questions	300	2.26	1.498	11 <sup>th</sup>
I use a test specification table	300	2.22	1.046	12 <sup>th</sup>
I write more items than needed before I review and select from them	300	2.18	1.212	13 <sup>th</sup>
I consider variation of students with respect to physical disabilities when writing the items	300	2.17	1.343	14 <sup>th</sup>
I write items at least two weeks before the date of testing	300	2.17	1.217	14 <sup>th</sup>
I copy questions from past questions	300	2.15	1.157	16 <sup>th</sup>

I construct items only when it is time to assess students	300	2.08	1.292	17 <sup>th</sup>
I ask other colleagues in the subject area to review the items by reading it over	300	1.71	1.043	18 <sup>th</sup>
I consider meaning of wording against different ethnic background of the students	300	1.67	1.325	19 <sup>th</sup>
I ask other colleagues to help me construct items	300	1.60	1.262	20 <sup>th</sup>

Table 8- Continued

Source: Field Survey, Sasu (2017)

From the results in Table 8, practices which were mostly followed by teachers when constructing test items include the following: “I provide clear simple instructions on how test is to be answered” ( $M = 3.30$ ,  $SD = 0.79$ ); “I evaluate items given to students” ( $M = 3.04$ ,  $SD = 0.93$ ); “I state the purpose of the test” ( $M = 2.8$ ,  $SD = 1.11$ ) and “I follow the principle of test construction for each item format” ( $M = 2.73$ ,  $SD = 1.09$ ). However, test construction practices that teachers did not often or frequently follow when constructing test items include the following: “I ask colleagues in the same subject area to review items” ( $M = 1.71$ ,  $SD = 1.02$ ); “I consider meaning of words against different ethnic background” ( $M = 1.67$ ,  $SD = 1.33$ ). Based on the reversal of the scale, “I construct items only when it is time to assess students” ( $M = 2.08$ ,  $SD = 1.292$ ) and “I ask any other colleagues to help me construct items” ( $M = 1.60$ ,  $SD = 1.26$ ), were items with negative implications to tests construction that was often practice by teachers.

The result from Table 8 shows that teachers consider the language proficiency of students when constructing test item ( $M = 2.52$ ,  $SD = 1.203$ ). This finding supports the assertion of Wiliam (1993), that, to increase the validity



of a test, teachers must consider the student's language proficiency. He further stipulated that "test would be invalidated if it turned out that the reading requirements of the test were so demanding that students with poor reading ability, but a sound understanding obtained low marks" (p. 4). On the other hand, if a student possesses an understanding of an issue demanded by a test, but fails to show it for reasons of linguistic difficulty then, the results of that test would be invalid. Teachers must therefore be commended for following such principles which would help to appreciate the validity of tests at the Junior High Schools in the Metropolis.

However, the result indicated that teachers did not often consider meaning of words against different ethnic background of their students when constructing test items ( $M= 1.65$ ,  $SD= 1.325$ ). When teachers fail to consider meaning of words against different ethnic background, the interpretation made from test may lead to faulty conclusions (Tom & Gary, 2003). The possible cause of teachers not considering the meaning of words against ethnic background of students may be as a result of the limited time and excessive workload on teachers which may lead them to give less attention to wording of test items with little consideration to students' ethnic background.

The study further revealed that teachers often asked other colleagues who are not in the subject area to help them construct test items. This attitude might have a great deal of implication to validity of test results. This is because the teacher assessing the students might not appropriately measure the real competence of the students since he/she might not know the detail of the content coverage and the thinking process to assess on a particular topic.

On the issue of reviewing test items, evidence from Table 8 clearly indicates that teachers did not review their tests ( $M=1.71$ ,  $SD=1.71$ ). This corroborate with findings in England, on review of research findings on a number of published studies by Crooks (1998) and Black (1993), which suggests that most teachers do not review their tests. The finding further supports the findings of Quaigrain (1992) on teacher-competence in the use of essay type tests in the secondary schools in the Western Region of Ghana who indicated that some teachers do not review their test.

From Table 8, there was an evidence that, teachers often use questions from students' text books. This finding is consistent with findings in the USA, which noted that due to limited time, teachers usually use the assessments that are found at the end of textbook chapters (AAAS, 1998).

In order to appreciate if teachers followed the principles of test construction practices to an appreciable level, one sample t-test was conducted with a hypothetical mean of 50 as the test value. This value was established by multiplying 2.5 as the item mean to the number of items (20). Thus  $2.5 \times 20 = 50$ . Therefore, overall mean of 50 and above indicates that teachers follow the basic principles in their test construction practices. A mean below 50 indicates that teachers do not follow test construction principles to an appreciable level. The results are presented in Table 9.

Table 9- *One Sample t- Test results of Test Construction Practices of Teachers*

Test value =50				
Mean	SD	t-value	Df	sig-value
47.0100	10.63612	-4.869	299	.000

Source: Field Survey, Sasu (2017)

A one sample t-test (two-tailed) was conducted to identify if teachers at the Junior High Schools follow the basic principles of test construction practices. The result from Table 9 indicates that teachers do not follow test construction principles to an appreciable level (Mean = 47.01, SD =10). This is because the test value (50) is higher than the mean score (47.01)

The results support the findings of Amedahe (1989) who indicated that to a great extent, secondary school teachers in the Central Region did not follow the basic suggested principles of classroom test construction. This finding also support the findings of Anhwere, (2009) who also revealed similar findings that teacher training college tutors do not follow the basic principle of testing in the construction of teacher made test or classroom tests. However, the finding does not support Oduro (2008), who found out that teachers in Senior High Schools in the Ashante Region of Ghana follow the principles of test construction.

### **Research Question Two**

What process do Junior High School teachers follow in administering test items?

This research question sought to find out the principles that Junior High School teachers follow in administering test items to students. In addressing this research question, teachers were given a five point Likert scale items to respond to.

The items on administration principles teachers follow were twelve (12). The scoring of items was based on the five point Likert scale of measurement ranging from “not often” (scored 1) to “always” (scored 4). For the purpose of clarity, items which indicate negative implications to test administration practice were reversed in coding which ranged from “not often” (scored 4) to

“Always” (scored 1). These items include: “I give more instructions during the time the students are taking the test” and “I make sure tests are given immediately before or after a long vacation, holidays or other important events. The results were ranked in descending order from principle teachers often follow to the principles teachers do not often follow. The results were then discussed using means and standard deviation. Cut-off point using of 2.5 was established to determine test administration practices that teachers often follow and the ones teachers do not often follow. A mean of 2.50 and above indicates teachers practiced test administration principles to an appreciable level while a mean below 2.5 indicates teachers did not practice the principles to an appreciable. The results are presented in Table 10.

Table 10- *Test Administration Practices of Teachers*

Test Administration Practices.	N	Mean	Std. D	Rank
I make my students aware of the rules and regulations covering the conduct of the test	300	3.63	.670	1 <sup>st</sup>
I make provision for extra answer sheets and writing materials during examination time	300	3.58	.711	2 <sup>nd</sup>
I make sure pupils start the test promptly and stop on time	300	3.48	.729	3 <sup>rd</sup>
I ensure that there is adequate ventilation and lighting expected in the testing room	300	3.36	.853	4 <sup>th</sup>
I proof read all my test items to eliminate ambiguities and unclear statement before administering the test	300	3.16	1.077	5 <sup>th</sup>
I inform the students about the test format	300	2.93	1.153	6 <sup>th</sup>

I inform students in advance the topics and the areas that the test will cover	300	2.62	1.222	7 <sup>th</sup>
I make provision for emergencies during the time the test is taken	300	2.30	1.230	8 <sup>th</sup>

Table 10 continue

I give more instructions during the time the students are taking the test	300	2.25	1.241	9 <sup>th</sup>
I make sure tests are given immediately before or after a long vacation, holidays or other important events	300	2.07	1.358	10 <sup>th</sup>
I prepare the classroom a day before the test is taking	300	1.95	1.295	11 <sup>th</sup>
I use "Do not Disturb" sign at the entrance of the examination classroom	300	1.35	1.447	12 <sup>th</sup>

Source: Field Survey, Sasu (2017)

From Table 10, practices that were agreed to have been mostly used by teachers in administering constructed test items to students include the following: “I make provision for extra sheets and writing materials” ( $M = 3.68$ ,  $SD = 1.84$ ); “I make students aware of the rule and regulations covering the test” ( $M = 3.63$ ,  $SD = 0.67$ ); “Students start and stop test on time” ( $M = 3.48$ ,  $SD = 0.73$ ); “I ensure adequate ventilation and lighting” ( $M = 3.36$ ,  $SD = 0.85$ ); “Proof read all test items” ( $M = 3.16$ ,  $SD = 1.08$ ). However, test practices that teachers did not often or frequently use when administering test to students include the following: “I prepare classroom a day before test is taken” ( $M = 1.95$ ,  $SD = 1.29$ ); “I use “Do Not Disturb” sign at entrance of classroom ( $M = 1.35$ ,  $SD = 1.45$ ). Base on the reversal of the scale, one item which had negative implication to test administration was often practiced. This is “Tests are given after a long vacation or important holidays” ( $M = 2.07$ ,  $SD = 1.36$ ).

From Table 10, it was evident that, teachers often followed quite a number of principles to an appreciable level. This support previous findings which agreeably confirm teachers did follow tests administration principles (Amedahe, 1998; Quagrain, 1992; & Oduro, 2008).

However, there was an evidence that teachers did not often follow the advance preparation of the class before administering test. This might lead to improper arrangement of the environment for a test which can hinder maximisation of individual test scores. This is because students trying to find a proper place to sit, due to improper arrangement of desks and poor lighting, may emotionally affect students. Notwithstanding the cause of this practice might be from the fact that, most of the Junior High Schools do not have adequate facilities in terms of classroom and desks to accurately administer tests without interrupting the learning process in other classes with respect to space, desks, and lighting. This is not consistent with Anhwere (2009), whose earlier findings suggested that Tutors at the Colleges of Education level had adequate facilities and also put in much effort to organise classroom appropriately when administering tests.

The findings further revealed that teachers did not adequately control noise when administering tests. This practice is not consistent with the assertion made by Mehrens and Lehmann (2001). According to Mehrens and Lehmann, noise and distraction in the testing environment should be kept at the barest minimum if not eliminated completely. Interruptions within and outside the testing room has the tendency of affecting student's performance. Etsey (2004) also affirmed that it is helpful to hang a "Do Not Disturb. Testing in Progress" sign at the door of the testing room to warn people to keep off. The distraction

from outside can divert the attention of test takers which could contribute to low performance of students.

The result also indicated that teachers often give tests immediately after a long vacation or an important holiday. This practice does hinder validation of tests. The practice is inconsistent with the assertion made by Amedahe and Gyimah (2003) and Etsey (2004) who concur that tests must not be given immediately before or just after a long vacation, holidays or other important events where students are involved either physically or psychologically. Amedahe and Gyimah (2003) went on to say that tests must also not be given when students would normally be doing something pleasant such as having lunch, athletics or other sporting activities as this will hamper students' concentration.

Therefore, teachers in the field of testing must recognise that the implication from the interpretation made of tests has greater consequences on the students than the teachers' idea of getting a score to represent assessment. Therefore, it would be prudent for teachers to ensure that scores from students' successive tests yield an appreciable consistency. According to Crocker and Algina (2008), psychological measurement should focus on a way of reducing systematic errors which may result from factors such as "fatigue, boredom, forgetfulness, guessing" among others (p. 6).

In order to find the overall practice of teachers on how they follow test administration principles in the Junior High Schools, one sample t-test (two-tail) was conducted with a hypothetical mean of 30 as the test value. This value was established by multiplying 2.5 as the item mean to the number of items (12). Thus ( $t = 2.5 * 12 = 30$ ). Therefore, overall mean of 30 and above indicates

that teachers follow the basic principles in their test administration practices. A mean below 30 shows low performance of teachers in following test administration principles to an appreciable level. The results are presented in Table 11.

Table 11- *One Sample t- Test results of Test Administration Practices of Teachers*

Test value =30					
Mean	SD	t-value	Df	sig-value	
32.6867	6.13574	7.584	299	.000	

Source: Field Survey, Sasu (2017)

Table 11 shows the result of one sample t-test (two-tail) of teachers at the Junior High Schools practices on principles of test administration. From Table 11 the result shows that majority of the teachers followed the principles of test administration practices. The test value of 30 was less than the computed mean of 32.69. This was coupled with one sample t-test results  $t(299) = 7.584$ ,  $p = .000$  which is statistically significant. This is because the p value is less than .05.

### **Research Question Three**

What basic principles of test scoring do teachers in the Junior High Schools follow in scoring of essay-type test items

This research question sought to find out the kind of test principles in a total of 13 items that Junior High School teachers consider when scoring essay-type items. In addressing this research question, teachers were given a five point Likert scale items to respond to. The scoring of items was based on the five point Likert scale of measurement ranging from “not often” (scored 1) to



“always” (scored 4). Items with negative implication to test scoring practices were reversed in coding ranging from “not often” (scored 4) to “Always” (scored 1). These items include; “I give extra marks to students besides the subject matter based on handwriting, gender, grammatical expression, length of students’ essay”, and “I am influenced by the first few papers read when scoring tests items”.

The results were ranked in descending order from principle teachers often follow to the principles teachers do not often followed. The results were then discussed using means and standard deviations. An item mean of 2.5 was established to determine test scoring practices that teachers often follow and the ones teachers did not often follow. A mean of 2.50 and above indicates teachers practiced test construction principle to an appreciable level while a mean below 2.5 indicates teachers did not practice the principle to an appreciable. The result is presented in Table 12.

Table 12-*Tests scoring practice of teacher*

Test Scoring Practice	N	Mean	Std. D	Rank
I promptly mark my students papers just after the test is taken	300	3.15	.940	1 <sup>st</sup>
I constantly follow the scoring guide as i score test items	300	3.01	1.065	2 <sup>nd</sup>
I provide comments and errors corrected on the scripts for class tests to facilitate learning	300	2.93	1.022	3 <sup>rd</sup>
I usually prepare a form of scoring guide, either an analytic scoring rubric or a holistic scoring rubric	300	2.76	1.156	4 <sup>th</sup>
I score the essay test when i am physically sound, mentally alert and in an environment with very little or no distraction	300	2.73	1.203	5 <sup>th</sup>

I keep scores of previously graded items out of sight when evaluating the rest of the items	300	2.32	1.308	6 <sup>th</sup>
I grade the responses item by item and not script by script after the tests are taken	300	2.24	1.276	7 <sup>th</sup>

Table 12 continue

I make sure the scripts are shuffled before starting to score	300	1.96	1.337	8 <sup>th</sup>
I give extra marks to students besides the subject matter based on handwriting, gender, grammatical expression, length of students essay	300	1.89	1.354	9 <sup>th</sup>
I score a particular question or item on all papers at one sitting	300	1.76	1.288	10 <sup>th</sup>
I make sure test takers are kept as anonymous as possible. that is different forms of identification are used instead of names	300	1.74	1.328	11 <sup>th</sup>
I am influenced by the first few papers read when scoring tests items	299	1.52	1.296	12 <sup>th</sup>
I periodically rescore previously scored papers	300	1.50	1.239	13 <sup>th</sup>

Source: Field Survey, Sasu (2017)

From Table 12, practices that teachers mostly undertook when scoring essay-type items include the following: “I mark papers just after the test is taken” ( $M = 3.15$ ,  $SD = 0.94$ ); “I constantly follow scoring guide” ( $M = 3.01$ ,  $SD = 1.07$ ). However, test practices that teachers did not often or frequently consider when scoring essay-type items include the following: “I shuffle scripts before scoring” ( $M = 1.96$ ,  $SD = 1.34$ ); “I score a particular item on all papers at a sitting” ( $M = 1.76$ ,  $SD = 1.29$ ); and “I make sure test takers are kept anonymous” ( $M = 1.74$ ,  $SD = 1.33$ ). Based on the reversal of the scale, the items

“I am influenced by the first few papers read” ( $M = 1.52$ ,  $SD = 1.29$ ) and “I give extra marks to students based on handwriting, gender etc.,” ( $M = 1.83$ ,  $SD = 1.35$ ) were often practiced. These items have negative implications to tests scoring.

Teachers indicated that they constantly follow the scoring guide when marking their tests. This process must be hailed for following some of the principles to an extent. This finding supports the assertion that admonishes teachers to constantly follow the marking scheme as they score tests items, as this reduces rater drift, which comes from the likelihood of either not paying attention to the scoring guide or interpreting it differently as time passes (Mehrens & Lehmann, 2001; Amedahe & Gyimah, 2003; and Etsey, 2004).

Notwithstanding, the result from the research also indicated that, teachers did not often consider reshuffling script when scoring their test. The finding does not follow the assertion of Mehrens and Lehmann (2001) that, randomly reshuffling of scripts when beginning to score each set of items will minimise the bias introduced as a result of the position of one's script. Research by Hales and Tokar (as cited in Mehrens and Lehmann, 2001) has shown that a student's essay grade will be influenced by the position of the paper, especially if the preceding answers were either very good or very poor. Mehrens and Lehmann pointed out that randomly reshuffling of scripts is especially significant when teachers are working with high- and low level classes and read the best scripts first or last.

Another finding of the research indicated that teachers did not often score a particular item on all papers at a sitting. Mehrens and Lehmann (2001), Amedahe and Gyimah (2003), and Etsey (2004) agreeably asserted that

responses of item should be scored item by item rather than script by script. This principle is to minimise the carryover effect on the scores and thereby ensure consistency. However, this finding is not consistent with the findings of Amedahe (1989), who recounted that teachers in the Senior Secondary Schools in the Central Region of Ghana used mainly the analytic method in scoring their essay-type tests. He further asserted that, teachers in the schools scored their essay-type tests either item by item or script by script.

With regards to scoring, teachers also indicated that they give extra marks to students based on handwriting, gender etc. This practice has been elaborated by Amedahe and Gyimah (2003) and Etsey (2004), who indicated that, the mechanics of expressions such as correct grammar usage, flow of expression, quality of handwriting, orderly presentation of material and spelling should be judged separately from subject matter correctness. When teachers are influenced by factors other than the subject matter, the marks awarded would represent construct irrelevant or construct mis-representativeness. This simply means higher scores on tests might not necessary reflect the ability of students on the subject matter but rather discriminate students in proficiencies they have over other students.

The results also indicated that, anonymity was not ensured when teachers score their test. This finding flouts the recommendation of Etsey (2004) who indicated that scripts must be scored anonymously. He suggested scripts should be identified by code numbers or any other means instead of the names of students. This principle is to reduce the halo-effect. This happens when a scorer's general impression of a person influences how the paper is scored.

In order to find the overall practice of teachers on how they followed test scoring principles in the Junior High Schools, one sample t-test was conducted with a hypothetical mean of 32.5 as the test value. This value was established by multiplying 2.5 as the item mean to the number of items (13). Thus ( $t = 2.5 * 13 = 32.5$ ). Therefore, overall mean of 32.5 and above indicates that teachers followed the basic principles in scoring their essay-type tests. A mean below 32.5 indicates that teachers did not follow test scoring principles to an appreciable level. The results are presented in Table 13.

Table 13- *One Sample t- Test results of Test Scoring Practices of Teachers*

Test value					
=32.5					
Mean	SD	t-value	df	sig-value	
29.5133	7.87548	-6.569	299	.000	

Source: Field Survey, Sasu (2017)

From Table 13 the result shows that majority of the teachers did not follow the principles of test scoring practices to an appreciable level. The test value of 32.5 was greater than the computed mean of 29.51 which gives evidence to the fact effect. The results of the one sample t-test  $t(299) = -6.569$ ,  $p = 000$  was statistically significant since the p value was less than .05.

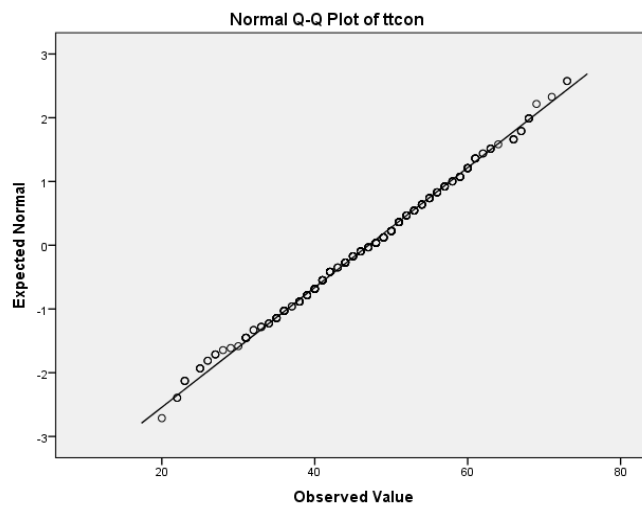
### **Research Hypothesis One**

There is no statistically significant differences among teachers who have taught for the following number of years in their test construction practices; 1-5, 6-10, 11-15 and 16 years and above.

This research hypothesis was used to find out whether there is any significant differences among teachers who have taught within the year ranges

of 1-5, 6-10, 11-15 and 16 years and above in Junior High School on the test practices they follow when constructing test items.

Using the number of years as the independent variable and the test construction practices as dependent variable, a One-Way Analyses of Variance (ANOVA) was conducted to determine whether there existed any significant difference among those who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above in Junior High School. The results are shown in Tables 15 and 16. Prior to conducting the One-way ANOVA, assumptions were checked. These assumptions were normality and linearity and homogeneity of variance. Figure 1 presents the normality and Table 14 presents the results of the homogeneity of variance.



*Figure 1: Normal Q-Q Plot of Test Construction*

According to Pallant (2007), a straight normal probability plot is an indication of normality. From Figure 1, a reasonable straight line could be seen from the plot demonstrating normality of the data. Table 14 shows the result of Levene test for homogeneity of variance.

*Table 14- Test of Homogeneity of Variances*

	Levene Statistic	df1	df2	Sig.
Test Construction	1.595	3	296	0.191

Source: Field Survey, Sasu (2017)

From Table 14, the Levene's test shows a sig. value greater than 0.05. This suggests that variances are assumed equal within age of teaching. Therefore, the assumption of homogeneity of variance was met.

Table 15 depicts the results of the descriptive statistics of teachers' years of teaching and the mean scores of their test construction practices.

Table 15- *Descriptive Statistics of Teachers Test Construction Practices*

Years of Teaching	N	Mean	Standard deviation
1-5	58	44.81	11.304
6-10	104	46.44	9.751
11-15	75	49.15	12.154
16 and above	63	47.43	9.113

Source: Field Survey, Sasu (2017)

From Table 15, the results show that there are differences in the mean scores. The table indicated the teaching years of experience of 11-15 had the highest mean score followed by 16 and above and 6-10 years while, 1-5 had the least mean score. However, in order to give more statistical evidence, One-Way Analyses of Variance (ANOVA) was conducted to find out whether there were statistical significant differences among teachers within the teaching years of experience. The result is presented in Table 16.

Table 16-*ANOVA Summary Table for Teachers Test Construction Practices*

Source	SS	Df	MS	F	Sig.
--------	----	----	----	---	------

Between	667.587	3	222.529	1.987	0.116**
groups					
Within Groups	33157.383	296	112.018		
Total	33824.970	299			

\*\*Not Significant at  $p > 0.05$

A one-way ANOVA was conducted to compare mean scores on test construction practices of teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above. The result showed that, there were no statistically significant differences among teachers who have taught within the age ranges of 1-5, 6-10, 11-15 and 16 years and above with respect to their test construction practices ( $F(3, 296) = 1.987, p > .05$ ). Therefore, the null hypothesis is not rejected. This result is consistent with previous study conducted by Anwhere (2009) using the Tutors in the Teacher Colleges of Education in Ghana. In this study, no statistically significant difference was found among teachers' test construction practices with respect to years of teaching among Tutors in Colleges of Education. It is therefore possible that teachers irrespective of years of teaching follow similar practices when constructing test items. However, the finding here is also at discrepancy with the finding of Amedahe (1989) who found that a moderate relationship exists between number of years of teaching and the accuracy with which teachers constructed their classroom achievement tests among teachers in Senior High Schools in Cape Coast.

### **Research Hypothesis Two**



There is no statistically significant differences among teachers who have taught for the following number of years in their test administration practices; 1-5, 6-10, 11-15 and 16 years and above.

This research hypothesis was used to find out whether there were statistically significant differences among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above in Junior High School on the test practices they follow when administering constructed test items to students.

Using the number of years as the independent variable and the test practices adhered to when administering test as dependent variable, a One-Way Analyses of Variance (ANOVA) was conducted to determine whether there existed any significant differences among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above in Junior High School. The results are presented in Tables 18 and 19. Prior to conducting the One-way ANOVA, assumptions were checked. These assumptions were normality, and homogeneity of variance. Figure 2 presents the normality and Table 15 presents the results of the homogeneity of variance.

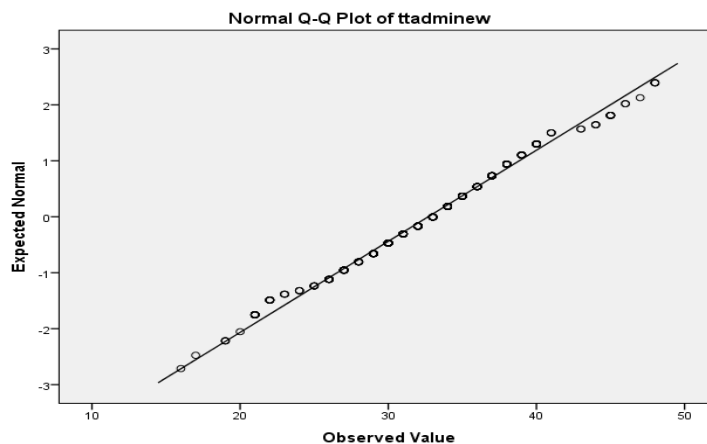


Figure 2: Normal Q-Q Plot of Test Administration

According to Pallant (2007), a straight normal probability plot is an indication of normality. From Figure 2, a reasonable straight line could be seen from the plot demonstrating normality of the data. Table 17 shows the result of Levene test for homogeneity of variance.

Table 17- *Test of Homogeneity of Variances*

	Levene Statistic	df1	df2	Sig.
Test Administration	.633	3	296	.594

Source: Field Survey, Sasu (2017)

From Table 17, the Levene's test shows a sig. value greater than 0.05. This suggest that variances are assumed equal among years of teaching. Therefore, the assumption of homogeneity of variance was met.

Table 18 presents the result of a one-way ANOVA which was conducted to compare mean scores on test administration practices among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above.

Table 18- *ANOVA Summary Table for Teachers Test Administration Practices*

Source	SS	Df	MS	F	Sig.
Between groups	135.281	3	45.094	1.200	0.310**
Within Groups	11121.266	296	37.572		
Total	11256.547	299			

\*\*Not Significant at  $p > 0.05$

From table 18, there was no statistically significant difference among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above with respect to their test administration practices ( $F(3, 296) = 1.696, p > .05$ ). Therefore, the null hypothesis is not rejected. This finding is therefore in line with findings obtained by Anhwere (2009) that teachers do not differ in their test administration practices with respect to years of teaching. The findings also support that of Quaigrain (1992) that a zero relationship exist between years of teaching and one's competence in tests administration practices.

Table 19 presents the descriptive statistics on the number of years that teachers have taught with their mean and standard deviation with respect to how test administration principles are followed.

Table 19- *Descriptive Statistics for Teachers Test Administration Practices*

Years of teaching	N	Mean	Standard deviation
1-5	58	31.86	6.495
6-10	104	32.60	5.970
11-15	75	33.76	5.603
16 and above	63	32.32	6.626

Source: Field Survey, Sasu (2017)

From the descriptive statistics, one may find differences in the mean scores of teacher who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above. Statistically, the one-way Analysis of Variance (ANOVA) showed that there was no statistically significant difference in test administration process with respect to the number of years taught.

### **Research Hypothesis Three**

There is no statistically significant difference among teachers who have taught for the following number of years; 1-5, 6-10, 11-15 and 16 years and above in terms of their essay test scoring practices.

This research hypothesis was to find out whether there were any significant differences among teachers who have taught within the year range of 1-5, 6-10, 11-15 and 16 years and above in Junior High School on practices they follow when scoring essay test items.

Using the number of years as the independent variable and the test practice adhered to when scoring essay test items as dependent variable, a One-Way Analysis of Variance (ANOVA) was conducted to determine whether there existed any significant differences among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above in Junior High School. The results are shown in Tables 21 and 22. Prior to conducting the One-way ANOVA, assumptions were checked. These assumptions were normality, and homogeneity of variance. Figure 3 presents the normality and Table 20 presents the results of the homogeneity of variance.

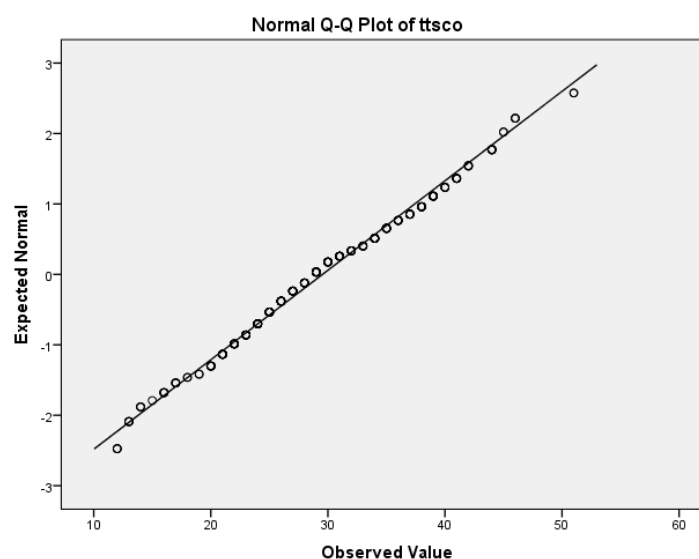


Figure 3: Normal Q-Q Plot of Essay Test Scoring

According to Pallant (2007), a straight normal probability plot is an indication of normality. From Figure 3, a reasonable straight line could be seen from the plot demonstrating normality of the data.

Table 20- *Test of Homogeneity of Variances*

	Levene Statistic	df1	df2	Sig.
Essay Test Scoring	1.277	3	295	0.282

Source: Field Survey, Sasu (2017)

From Table 20, the Levene’s test shows a sig. value greater than the 0.05. This suggests that variances are assumed equal among age of teaching. Therefore, the assumption of homogeneity of variance was met.

Table 21 presents the result of a one-way ANOVA was conducted to compare mean scores on essay test scoring practices of teachers who have taught among the year ranges of 1-5, 6-10, 11-15 and 16 years and above

Table 21- *ANOVA Summary Table for Teachers Test Scoring Practices*

Source	SS	Df	MS	F	Sig.
Between groups	331.616	3	110.539	1.798	0.148**
Within Groups	18140.611	295	61.494		
Total	18472.227	298			

\*\*Not Significant at  $p > 0.05$

From table 21, there was no statistically significant difference among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above with respect to their test scoring practices ( $F(3, 295) = 1.798, p > .05$ ).

The means and standard deviations are presented in Table 22.

Table 22- *Descriptive Statistics for Teachers Test Scoring Practices*

Years of teaching	N	Mean	Standard deviation
1-5	58	30.31	8.200
6-10	104	29.53	7.034
11-15	74	30.55	7.725
16 and above	63	27.67	8.850

Source: Field Survey, Sasu (2017)

Table 22 shows differences in mean score among the teaching years of teachers ranging from 1-5, 6-10, 11-16, and 16 years and above. Further statistical analysis using the ANOVA proved no statistically significant difference among teachers with respect to the years taught. Therefore, the null hypothesis is not rejected.

The findings suggest that irrespective of years of teaching, teachers probably fail to abreast themselves with technicalities to ensure they follow good practices in scoring essay-type tests.

#### **Research Hypothesis Four**

There is no statistically significant difference in test construction practices, among teachers in the four core subject areas (Mathematics, Science, English and Social studies).

This research hypothesis was to examine whether there were significant differences among teachers who teach the core subjects (Mathematics, Science, English and Social Studies) in Junior High School with respect to their test construction practices.

Using subject taught as the independent variable and the test practice adhered to when constructing test items as dependent variable, a One-Way

Analyses of Variance (ANOVA) was conducted to determine whether there existed any significant differences among those who teach the four core subjects in Junior High School. Prior to conducting the One-way ANOVA, assumptions were checked. These assumptions were normality, and homogeneity of variance. Figure 4 presents the normality and Table 24 presents the results of the homogeneity of variance. The descriptive statistics was presented in Table 23.

Table 23- *Descriptive Statistics for Teachers Test Construction Practices in Subject Taught*

Subject Area	N	Mean	Standard deviation
Social Studies	50	42.92	9.820
English	52	45.33	10.718
Maths	50	50.44	10.766
Science	47	47.15	11.000

Source: Field Survey, Sasu (2017)

The descriptive statistics shows differences in mean score of teachers in the core subject area with respect to their test construction practices. A further analysis was conducted using the one-way Analysis of Variance (ANOVA) to find out whether the differences among the core subject teachers with respect to their test construction practices is significant. The result is presented in Table 23.

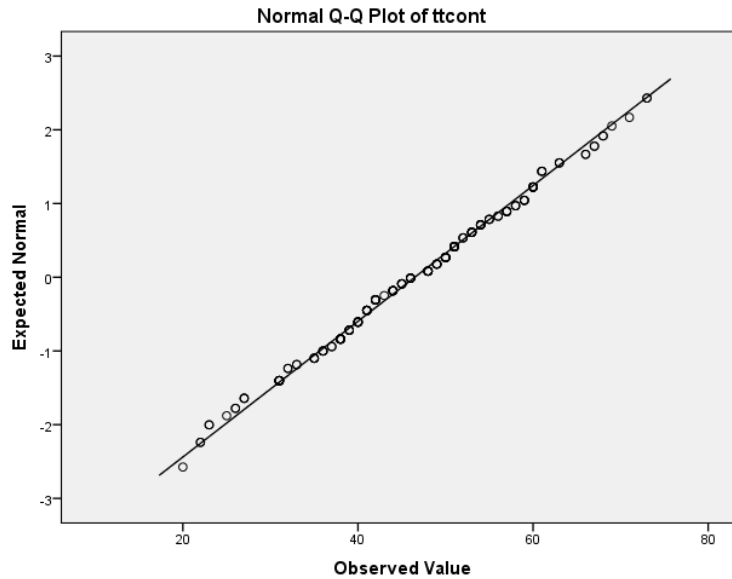


Figure 4: Normal Q-Q Plot for Test Construction

According to Pallant (2007), a straight normal probability plot is an indication of normality. From Figure 4, a reasonable straight line could be seen from the plot demonstrating normality of the data. Table 24 presents the Levene statistic for homogeneity of variance.

Table 24- *Test of Homogeneity of Variances*

	Levene Statistic	df1	df2	Sig.
Test Construction	0.403	3	195	0.751

Source: Field Survey, Sasu (2017)

From Table 24, the Levene's test shows a sig. value greater than the 0.05. This suggest that variances are assume equal among core subject teachers. Therefore, the assumption of homogeneity of variance was met.

Table 25- *ANOVA Summary Table for Teachers Test Construction Practices in the Core Subjects*

Source	SS	Df	MS	F	Sig.
--------	----	----	----	---	------



Between	1507.565	3	502.522	4.488	0.005**
groups					
Within Groups	21835.400	195	111.976		
Total	23342.965	198			

\*\*Significant at  $p < 0.05$

A one-way ANOVA was conducted to compare mean scores on test construction practices of teachers who teach the four core subject areas (Mathematics, Science, English and Social Studies). The result showed a statistically significant difference among teachers who teach the four core subject areas (Mathematics, Science, English and Social Studies) with respect to their test construction practices ( $F(3, 195) = 4.488, p < .05$ ). Since difference among independent group was statistically significant, the null hypothesis is rejected. A further analysis was done to find the between group means where the significant differences exists using Tukey's test. The result is shown in Table 26.

Table- 26 *Multiple Comparisons (Post-hoc/Follow up)*

Subject taught (I)	Subject taught (J)	Mean difference	Sig (2-tailed)
Social Studies	English	-2.407	0.660
	Maths	-7.52	0.003**
	Science	-4.23	0.204
English	Social Studies	2.41	0.660
	Maths	-5.11	0.073
	Science	-1.82	0.828
Maths	Social Studies	7.52	0.003**

	English	5.11	0.073
	Science	3.29	0.421
Science	Social Studies	4.23	0.204
	English	1.82	0.828
	Maths	-3.29	0.421

---

\*\*Significant at  $p < 0.05$

Table 26 presents the multiple comparisons (Post Hoc/ Follow-up test) using, Tukey's test to find the differences in test construction practices among teachers in the four core subject areas. According to Field (2009), Tukey's test as a post-hoc comparison is appropriate when equal variance assume. Considering the Levene's test for this data, the assumption of homogeneity of variance was met. The Tukey's test indicated that the mean score for Maths subject teachers ( $M = 50.44$ ,  $SD = 10.77$ ) was significantly different from Social Studies teachers ( $M = 42.92$ ,  $SD = 9.82$ ). However, considering the direction of their performance with respect to the mean scores, Maths subject teachers did better than Social Studies teachers. English and Science teachers did not differ significantly from either of the other subject teachers. It is possible that Mathematics teachers devote more attention to test construction process because the subject involves computations, hence a little twist may lead to faulty interpretation to the students.

#### **Research Hypothesis Five**

There is no statistically significant differences in test administration, among teachers in the four core subject areas (Mathematics, Science, English and Social Studies).

This research hypothesis was used to find out whether there was any significant difference among teachers who teach the core subjects (Mathematics, Science, English and Social studies) in Junior High School with respect to their test administration practices.

Using subject taught as the independent variable and the test practice adhered to when administering test as dependent variable, a One-Way Analyses of Variance (ANOVA) was conducted to determine whether there existed any significant differences among those who teach the four core subjects in Junior High School. The results are shown in Tables 28 and 29. Prior to conducting the One-way ANOVA, assumptions were checked. These assumptions were normality, and linearity and homogeneity of variance. Figure 5 presents the normality and Table 27 presents the results of the homogeneity of variance.

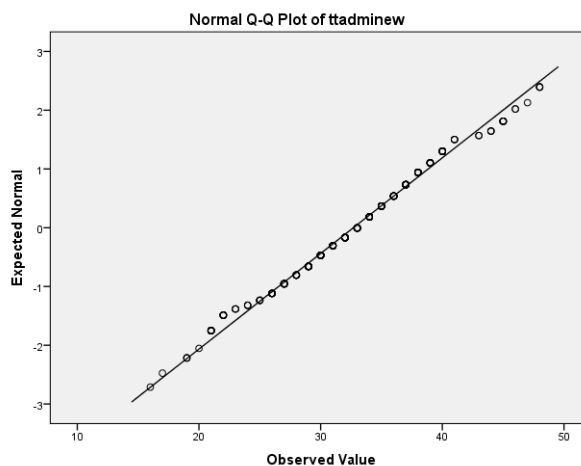


Figure 5: Normal Q-Q Plot for Test Administration

According to Pallant (2007), a straight normal probability plot is an indication of normality. From Figure 5, a reasonable straight line could be seen from the plot demonstrating normality of the data.

Table 27- *Test of Homogeneity of Variances*

	Levene Statistic	df1	df2	Sig.
Test Administration	0.132	3	195	0.941

Source: Field Survey, Sasu (2017)

From Table 27, the Levene's test shows a sig. value greater than the 0.05. This suggest that variances are assumed equal among core subject teachers. Therefore, the assumption of homogeneity of variance was met. Table 28 presents the descriptive statistics on the subject taught by teachers and their test administration practices.

Table 28- *Descriptive Statistics for Teachers Test Administration Practices in Subject Taught*

Subject area	N	Mean	Standard deviation
Social Studies	50	32.02	5.316
English	52	32.04	6.278
Maths	50	35.54	6.991
Science	47	32.02	5.962

Source: Field Survey, Sasu (2017)

From the descriptive statistics in Table 28, there exist a difference among the teachers in the core subject areas with respect to their test administration processes. A further statistical tool was applied to find out

whether the differences among the teachers of the various categories were significant. The results are presented in Table 29.

Table 29- ANOVA Summary Table for Teachers Test Administration Practices in the Core Subjects

Source	SS	Df	MS	F	Sig.
Between groups	462.070	3	154.023	4.045	0.008**
Within Groups	7424.302	195	38.073		
Total	7886.372	198			

\*\*Significant at  $p < 0.05$

A one-way ANOVA was conducted to compare mean scores on test administration practices of teachers who teach the four core subject areas (Mathematics, Science, English and Social studies). The result showed a statistically significant difference among teachers who teach the four core subject areas (Mathematics, Science, English and Social studies) with respect to their test administration practices ( $F(3, 195) = 4.045, p < .05$ ). Since difference among independent group was statistically significant, the null hypothesis was rejected. A further analysis using Tukey's test was done to find where the significant differences exist. The result from the Multiple Comparisons (Post-hoc/Follow up) test is presented in Table 30.

Table 30- Multiple Comparisons (Post-hoc/Follow up)

Subject taught (I)	Subject taught (J)	Mean difference	Sig (2-tailed)
Social studies	English	-0.185	1.00
	Maths	-3.520	0.025**

	Science	-0.0013	1.00
English	Social studies	0.0185	1.00
	Maths	-3.502	0.024**
Maths	Science	0.017	1.00
	Social studies	3.520	0.025**
	English	3.502	0.024**
Science	Science	3.519	0.028**
	Social studies	0.0013	1.00
	English	-0.017	1.00
	Maths	-3.519	0.028**

\*\*Significant at  $p < 0.05$

Table 30 presents the multiple comparisons (Post Hoc/ Follow-up test) using, Tukey's test to find the differences in test administration, among teachers in the four core subject areas (Mathematics, Science, English and Social Studies). According to Field (2009), Tukey's test as a post-hoc comparison is appropriate when variance assume equal. Considering the Levene's test for this data, the variances assumed equal.

The post-hoc comparison indicated that the mean score for Social Studies subject teachers ( $M = 32.02$ ,  $SD = 5.32$ ) was significantly different from Maths teachers ( $M = 35.54$ ,  $SD = 6.99$ ),  $p < 0.5$ . Considering the direction of their performance with respect to the mean scores, Maths teachers performed better than Social Studies teachers.

Also, there was a significant difference between the mean score for English teachers ( $M = 32.04$ ,  $SD = 6.28$ ) and Maths teachers ( $M = 35.54$ ,  $SD =$

6.99). Considering the direction of their performance with respect to the mean score, Maths teachers performed better than English teachers.

Again, a significant difference was indicated between the mean score for Maths teachers ( $M=35.54$ ,  $SD= 6.99$ ) and Science teachers ( $M=32.02$ ,  $SD= 5.96$ ). However, considering the direction of their performance with respect to the mean score, Maths teachers performed better than Science teachers.

No statistically significant difference was found between Social Studies teachers and science teachers. Also, a pairing between English teachers and Social Studies teachers showed no statistically significant difference between teachers. Furthermore, a comparison between the mean score of English teachers and Science teachers indicated no significant difference between them. It is possible that Mathematics teachers find test administration process as a contributing factor to students' performance than other subject areas do. This might be the reason they adhere to test administration principles more than the other subject areas.

### **Research Hypothesis Six**

There is no statistically significant difference in essay test scoring, among teachers in the essay-type subject areas (English, Social Studies and Religious and Moral Education (RME)).

This research hypothesis was used to find out whether there was any significant difference among teachers who teach essay-type subjects (English, Social studies and Religious and Moral Education) in Junior High School with respect to their test scoring practices.

Using subject taught as the independent variable and the test practice adhered to when scoring test as dependent variable, a One-Way Analyses of

Variance (ANOVA) was conducted to determine whether there existed any significant difference between those who teach essay-type subject areas (English, Social Studies and Religious and Moral Education) in Junior High School. The results are shown in Tables 32 and 33. Prior to conducting the One-way ANOVA, assumptions were checked. These assumptions were normality, and homogeneity of variance. Figure 6 presents the normality and Table 31 presents the results of the homogeneity of variance.

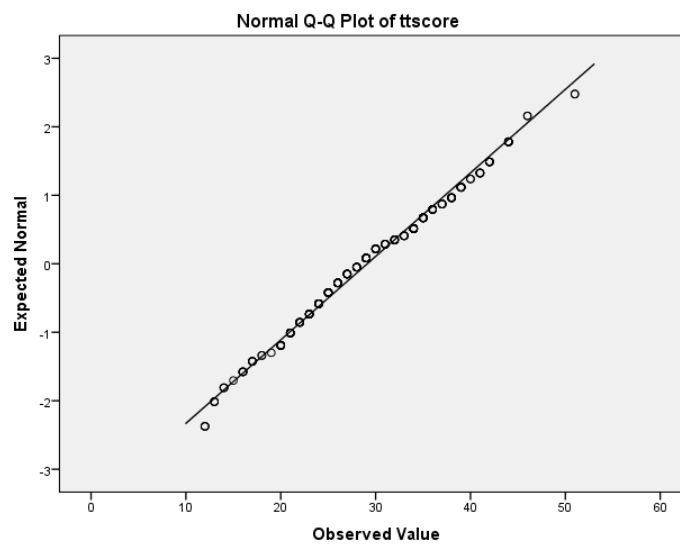


Figure 6: Normal Q-Q Plot for Essay Test Scoring

According to Pallant (2007), a straight normal probability plot is an indication of normality. From Figure 6, a reasonable straight line could be seen from the plot demonstrating normality of the data. Table 31 presents the Levene test for homogeneity of variance.

Table 31- *Test of Homogeneity of Variances*

	Levene Statistic	df1	df2	Sig.
Essay Test Scoring	2.054	2	145	0.132

Source: Field Survey, Sasu (2017)



From Table 31, the Levene's test shows a sig. value greater than the 0.05. This suggest that variances are assume equal among essay-type subject areas (English, Social Studies and Religious and Moral Education (RME)). Therefore, the assumption of homogeneity of variance was met.

Table 32- *Descriptive Statistics for Teachers Test Scoring Practices in Essay-type Subjects*

Subject area	N	Mean	Standard deviation
Social Studies	50	28.06	8.119
English	52	26.25	7.321
RME	47	30.51	8.119

Source: Field Survey, Sasu (2017)

Table 32 showed that there were differences in mean scores among the essay-type subject areas. A further statistical analysis using one-way Analysis of Variance (ANOVA) was conducted to find out if the differences among the means were statistically significant. Table 33 presents the results.

Table 33- *ANOVA Summary Table for Teachers Test Scoring Practices in Essay-type Subjects*

Source	SS	Df	MS	F	Sig.
Between groups	449.608	2	224.804	3.353	0.038**
Within Groups	9720.311	145	67.037		
Total	10169.919	147			

\*\*Significant at  $p < 0.05$

A one-way ANOVA was conducted to compare mean scores on test scoring practices of essay-type subject areas (English, Social Studies and Religious and Moral Education). There was a statistically significant difference among teachers on their test scoring practices on essay-type subject areas ( $F(2, 145) = 3.353, p < .05$ ). Since the difference among the independent groups was statistically significant, the null hypothesis was rejected. A further analysis using Tukey's test was done to find where the significant difference exists. Table 34 shows the results from the comparison.

Table 34- *Multiple Comparisons (Post-hoc/Follow up)*

Subject taught (I)	Subject taught (J)	Mean difference	Sig (2-tailed)
Social Studies	Religious and Moral Edu.	-2.449	0.308
	English	1.811	0.449
RME	Social Studies	2.449	0.308
	English	4.261	0.022**
English	Social Studies	-1.811	0.449
	Religious and Moral Edu.	-4.426	0.022**

\*\*Significant at  $p < 0.05$

Table 34 presents the multiple comparisons (Post Hoc/ Follow-up test) using, Tukey's test to find the differences in test scoring practices among teachers on essay-type subject areas (English, Social Studies and Religious and Moral Education). According to Field (2009), Tukey's test as a post-hoc comparison is appropriate when the data meets the assumption of variance. Considering the Levene's test for this data, the assumption of homogeneity of

variance was met. The post-hoc comparison test indicated that the mean score for Religious and Moral Education subject teachers ( $M= 30.51$ ,  $SD= 9.12$ ) was significantly different from English teachers ( $M=26.25$ ,  $SD= 7.32$ ). Considering the direction of their performance with respect to the mean score, Religious and Moral Education subject teachers performed better than English teachers. English and Social teachers did not differ significantly from each other with respect to scoring of essay-type test. The possible reason might be that, marking of (RME) demands more objectivity in relation to the concept to measure which makes teachers very objective in following the principles of scoring.

### **Chapter Summary**

The findings from this study revealed that teacher made tests were mainly essay, short answer and multiple-choice type. Also, the finding from this study suggested that teachers at the Junior High Schools performed low in following the basic principles in constructing teacher-made test. The research further suggested that, teachers' tests practices in relation to test administration were considered to be fair. With respect to scoring essay items, the result suggested teachers followed relatively low practices with respect to the principles of scoring essay test items. No significant differences were observed among teachers who have taught within the age ranges of 1-5, 6-10, 11-15 and 16 and above with respect to test construction, test administration and essay-type test scoring. A significant difference was found among teachers who taught in the four core subject areas with respect to test construction, administration and essay-type test scoring.

## **CHAPTER FIVE**

### **SUMMARY, CONCLUSIONS AND RECOMMENDATIONS**

#### **Overview of Study**

The study sought to investigate the testing practices with regards to basic principles JHS teachers follow in the construction, administering and scoring of classroom or teacher-made tests with specific reference to the practices in Core Subject areas and scoring of essay-type test items of Junior High School teachers. A descriptive sample survey was conducted in the Cape Coast Metropolis using a self-constructed questionnaire items which was administered to 300 teachers from five circuits. The study was guided by the following research questions and hypotheses:

1. What principles do Junior High School teachers follow in constructing test items?
2. What process do Junior High School teachers follow in administering test items?
3. What basic principles of test scoring do teachers in the Junior High Schools follow in scoring of essay-type test items?
4. There are no statistically significant differences among teachers who have taught for the following number of years in their test construction practices; 1-5, 6-10, 11-15 and 16 years and above.

5. There are no statistically significant differences among teachers who have taught for the following number of years in their test administration practices; 1-5, 6-10, 11-15 and 16 years and above.
6. There are no statistically significant differences among teachers who have taught for the following number of years; 1-5, 6-10, 11-15 and 16 years and above in terms of scoring their essay-type test.
7. There are no statistically significant differences in test construction practices, among teachers in the four core subject areas (Mathematics, Science, English and Social studies).
8. There are no statistically significant differences in test administration, among teachers in the four core subject areas (Mathematics, Science, English and Social studies).
9. There are no statistically significant differences in test scoring, among teachers in the essay-type subject areas (English, Social studies and Religious and Moral Education)

## **Summary of Key Findings**

### **Research Question One**

What principles do Junior High School teachers follow in constructing test items?

The finding on this research question showed that teachers engage in some good principles such as providing clear and simple instructions on how test is to be answered, evaluating items given to the students to answer and stating the purpose of which the test will be used. However, the overall mean score for test construction principles indicated that, teachers did not follow the principles to an appreciable level.

### **Research Question Two**

What process do Junior High School teachers followed in administering test items?

The finding on these respondents agreed to the statement that teachers follow basic principles when administering tests to students. The test administration principles that were commonly practiced among teachers include; making students aware of the rules and regulations covering the conduct of the test, making provision for extra answer sheets and writing materials during examination time and then making sure pupils start the test promptly and stop on time.

### **Research Question Three**

What basic principles of test scoring do teachers in the Junior High Schools follow in scoring of essay-type test items?

The finding on these respondents agreed to the statement that, the major scoring practices that teachers followed were, promptly marking their students' papers just after the test is taken, constantly follow the scoring guide as they score test items and providing comments and errors corrected on the scripts for class tests to facilitate learning. However, the overall mean score for principles teachers followed when scoring essay-type tests indicated that teachers did not follow the principles to an appreciable level.

### **Research Hypothesis One**

There is no statistically significant differences among teachers who have taught for the following number of years in their test construction practices; 1-5, 6-10, 11-15 and 16 years and above.

The findings from this research hypothesis showed no statistically significant differences among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above with respect to their test construction practices.

### **Research Hypothesis Two**

There is no statistically significant differences among teachers who have taught for the following number of years in their test administration practices; 1-5, 6-10, 11-15 and 16 years and above. The findings from this research hypothesis showed no statistically significant differences among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above with respect to their test administration practices.

### **Research Hypothesis Three**

There is no statistically significant differences among teachers who have taught for the following number of years; 1-5, 6-10, 11-15 and 16 years and above in terms of scoring their essay-type test.

The finding from this research hypothesis showed no statistically significant differences among teachers who have taught within the year ranges of 1-5, 6-10, 11-15 and 16 years and above with respect to their test scoring practices.

### **Research Hypothesis Four**

There is no statistically significant differences in test construction practices, among teachers in the four core subject areas (Mathematics, Science, English and Social studies).

The finding from this research hypothesis revealed a statistically significant differences among teachers who teach the four core subject areas

(Mathematics, Science, English and Social studies) with respect to their test construction practices. A Post hoc analysis revealed a significant difference between Maths teachers and Social Studies teachers. The results indicated that Mathematics teachers performed better than Social Studies Teachers in their test construction practices.

#### **Research Hypothesis Five**

There is no statistically significant differences in test administration practices, among teachers in the four core subject areas (Mathematics, Science, English and Social studies).

The finding from this research hypothesis revealed a statistically significant difference among teachers who teach the four core subject areas (Mathematics, Science, English and Social studies) with respect to their test administration practices. A Post hoc analysis revealed a significant difference between Mathematics teachers and the other core subject area teachers (English, Science and Social Studies). The result indicated that Mathematics teachers performed better than teachers in the other core subject areas (English, Science and Social Studies) in their test administration practices. The result did not show statistically significant differences among the mean scores for (English, Science and Social Studies) teachers.

#### **Research Hypothesis Six**

There is no statistically significant differences in essay-type test scoring practices, among teachers in the essay-type subject areas (English, Social studies and Religious and Moral Education).

The finding from this research hypothesis revealed a statistically significant difference among teachers in their test scoring practices on essay-



type subject areas. A Post hoc analysis revealed that Religious and Moral Education subject teachers performed significantly better than English teachers.

### **Conclusions**

It is evident from the findings of the study that responses of teachers at the Junior High Schools on how test administration processes are followed was appreciably good. However, the results from the construction and scoring practices was not followed to an appreciable level. Teachers in the Junior High Schools have a sensitive responsibility of assessing and making decisions concerning students' academic progress. They are expected to be professional to an appreciable level in their testing practices. However, low performance in teachers testing practices on the construction and scoring practices may be from the fact that they are comfortable with negative practices without recognising the impact of their practices on issues of validity and reliability.

The result from this study further suggest that, aside teachers' deficiencies in testing practices there is discrepancies with respect to the subject taught and how teachers adhere to testing practices in the Junior High Schools.

### **Recommendations**

With respect to the findings resulting from the study, the following recommendations are made for the improvement of testing practices of teachers in the Junior High Schools in Ghana.

1. I suggest more in-service training, both school based and Metropolitan organised training, should be organized for teachers in Junior High Schools with respect to their testing practices (construction, administration and scoring of test). This could be achieved through the collaboration of the Ministry of Education, the Colleges of Education,

University of Cape Coast and other stakeholders of education. The rationale is to assist teachers to improve on good testing practices and also sustain the testing practices that most teachers agreed to have followed to an appreciable level. This is because testing forms an integral part of the teaching profession since it is the most widely used as a tool for assessing students in Ghana.

2. From the office of the Educational Directorate, Junior High School teachers in the Cape Coast Metropolis should be sensitized on regular basis by Lead teachers, Circuit supervisors and staff from the Metropolitan Education office, on the importance of their testing practices with regard to construction, administration and scoring. Teachers in the Metropolis should know about the implication of their testing practices and its effect on validity and reliability which will adversely affect how decisions are made on their students. Teachers should be made aware of how a misleading score could affect the future of a student.
3. A guideline on test construction should be design by the Curriculum Research and Development Division in collaboration with tertiary institutions such as University of Cape Coast and University of Education, Winneba which could be made available for teachers to guide them in test construction.

### **Suggestions for Future Research**

The following are suggested for future research:

1. A study should be carried out to look into testing practices in terms of item analyses of objective type test of teachers.

2. A study also needs to be carried out to look at the perception of teachers in following the principles of testing practices and its effect on their practices.
3. The study can further be replicated to cover a wider range of population to establish the extent to which teachers in Ghana follow the basic principles of test construction, administration and scoring.

## REFERENCES

- Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice and free response writing test. *Applied Psychological Measurement, 12*, 117-128.
- Adamolekun, J. (1985). Effective test construction: A case for the senior secondary school mathematics curricula. *Journal of the School of Pure Science, 1*(3), 197-224.
- AERA/APA/NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Agu, N. N., Onyekuba, C., & Anyichie, A. C. (2013). Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: Need for a test construction skill inventory. *Academic Journals, Educational Research and Review, 8* (8), 431-436.
- Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. USA: Waveland Press Inc.
- Amedahe, F. K. (1989). *Testing practices in secondary schools in the Central Region of Ghana*. Unpublished master's thesis. University of Cape Coast, Cape Coast, Ghana.
- Amedahe, F. K. (2000). *Continuous assessment*. Unpublished paper, University of Cape Coast, Cape Coast, Ghana.
- Amedahe F. K. (2002). *Fundamentals of educational research methods*. Mimeograph, University of Cape Coast (Unpublished).
- Amedahe, F. K., & Etsey, Y. K. A. (2003). *Assessment criteria, feedback and evaluation*. Unpublished paper presented at the University of Cape Coast, Ghana, October 2003.

- Amedahe, F. K. & Asamoah-Gyimah, K. (2003). *Measurement and evaluation*. Cape Coast, Ghana: Centre for Continuing Education.
- Amedahe, F. & Asamoah-Gyimah, K. (2011). *Introduction to measurement and evaluation*. Centre for Continuing Education, University of Cape Coast, Ghana: University Printing Press.
- American Association for the Advancement of Science (AAAS). (1998). *Blueprint online-project 2061*. Retrieved from: <http://www.project2061.org>. online.
- Anamua-Mensah, J. & Quagrain K. A. (1998). Teacher competence in the use of essay-tests. A study of secondary schools in the Western Region of Ghana. *The Oguaa Educator*, 12 (1), 31-43.
- Anastasi, A. (1982). *Psychological testing*. New York: Macmillan Publishing Company.
- Anhwere, Y. M. (2009). *Assessment practices of teacher Colleges of Education tutors in Ghana*. Unpublished Master's thesis. University of Cape Coast. Cape Coast, Ghana.
- Ary, D., Jacobs, L. C., Sorensean, C., & Razavieh, A. (2010). *Introduction to research in education* (8<sup>th</sup> ed.). Canada: Wadsworth, Lengage Learning.
- Asamoah-Gyimah, K. (2002). *An evaluation of the practice of continuous assessment in the senior secondary schools in the Ashanti Region of Ghana*. Unpublished Master's thesis, University of Cape Coast, Cape Coast, Ghana.
- Bartels, E. K. (2003). The practice of continuous assessment in teacher training colleges in Ghana. *Journal of Educational Development and Practice*, 1 (1), 59-72.

- Bejar, I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*, 2 (1), 175-189.
- Benson, J. (1998). Developing a strong program of construct validation: a test anxiety example. *Educational Measurement: Issues and Practice*, 17, 10-17.
- Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Annee Psychologique*, 12, 191–244. In E. S. Kite (Ed.), *The Development of Intelligence in Children* (pp. 15-35). Vineland, NJ: Publications of the Training School at Vineland.
- Black, P. J. (1993) Formative and summative assessment by teachers. *Studies in Science Education*. 21, 49 – 97.
- Black, P. & William, D. (1998). Teachers' practices in formative evaluation. *Assessment in Education: Principles, policy and Practice* 5 (1), 7-68.
- Bollen, P. J. (1989) Formative and summative assessment by teachers. *Studies in Science Education*. 21, 49 – 97.
- Brown, B. S. (2007). *Handbook on formative and summative evaluation of student learning*. New York: McGraw- Hill, Inc.
- Chauhan, S. S. (2008). *Advanced educational psychology* (7<sup>th</sup> ed.). Shimla: Vicas Publishing House PVT Ltd.
- Childs, R. A. (1989). Constructing classroom achievement tests. *ERIC* 2, 12-24.
- Creswell, J. W. (2003). *Research design, qualitative, quantitative and mixed methods approaches* (2<sup>nd</sup> ed.). Thousand Oaks, CA: Sage Publications, Inc.

- Creswe, J. W. (2012). *Planning, conducting and evaluating quantitative and qualitative research* (4<sup>th</sup> ed.). Boston: Pearson Education Inc.
- Cohen, L., Manion, L. & Morrison, (2000). *Research methods in education* (5<sup>th</sup> ed.). London: Routledge Falmer.
- Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W. & Talbert, J. E. (2001). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mahwah, N.J: Erlbaum.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth: Holt, Rinehart and Winston.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. USA: Lengage learning.
- Cronbach, L. J. (1960). *Essentials of psychological testing* (2<sup>nd</sup> ed.). New York: Harpers and Brothers Publishers.
- Crooks, T. J. (1998). The impact of classroom evaluation practices on students. *Review of Educational Research*. 58, 438– 481.
- Crooks, T. (2001). *The validity of formative assessment*. Educational Assessment Unit, University of Otago, Dunedin, New Zealand. A paper presented at the British Educational Research Association annual Conference, University of Leeds, 13-15 September 2001.
- Cunningham, G. K. (2001). *Educational and psychological measurement*. New York: Macmillan Publishing Company.
- DeVellis, R. F. (1991). *Scale Development: Theory and applications*. Newbury Park: Sage.

- De Vos, A. S. (1998). Conceptualisation and operationalisation. In A.S. De Vos (Ed.), *Research at grass roots* (pp. 41-45). A primer for the caring professions, South Africa, Pretoria: Van Schaik Publishers.
- DuBois, P. H. (1970). *A history of psychological testing*. Boston: Allyn and Bacon Inc.
- Ebel, L. R., & Frisbie, A. D. (1991). *Essentials of educational measurement*. Englewood Cliffs, New Jersey: Prentice Hall.
- Ebel, R. L., & Frisbie, D. A. (2001). *Essentials of educational measurement* (5<sup>th</sup> ed.). Englewood Cliffs, New Jersey: Prentice Hall.
- Education Up Close (2005). *Writing effective tests: A guide for teachers*. Retrieved from <http://www.glencoe.com/sec/teachingtoday/educationup.p>.
- Etsey, Y. K. A. (2003). Pre-Service teacher's knowledge of continuous assessment teaching in Ghana. *Journal of Educational Development and Practice, 1* (1), 1-18.
- Etsey, Y. K. (2004) Assessing performance in schools: Issues and practice. *Ife Psychologia, 13* (1), 123-135.
- Etsey, Y. K. A. (2004). *Educational measurement and evaluation*. Lecture notes on EPS 203. Unpublished document, Educational Foundations, University of Cape Coast, Ghana.
- Etsey, K. (2005). *Causes of low academic performance of primary school pupils in the Shama sub-metro of Shama Ahanta East Metropolitan Assembly (SAEMA) in Ghana*. Paper presented at the Regional Conference on Education in West Africa, Dakar, Senegal, 1–2 November 2005.



- Etsey, Y. K. A. (2012). *Assessment in education*. Unpublished, University of Cape Coast, Cape Coast, Ghana.
- Field, A. (2009). *Discovering statistics using SPSS* (3<sup>rd</sup> ed.). London: SAGA Publication Ltd.
- Flanagan, D, Genshaft, J. L. & Harrison, P. L. (1997). *Intellectual assessment, tests, and issues*. New York: The Guilford Press.
- Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. *New Directions for Testing and Measurement*, 19(3), 29-38.
- Fraenkel, J. R, & Wallen, N. E. (1993). *How to design and evaluate research in education*. New York: McGraw Hill.
- Fraenkel, J. R, & Wallen, N. E. (2000). *How to design and evaluate research in education* (6<sup>th</sup> ed.). Boston: McGraw Hill.
- Freeman, H. G. & Lewis S. (2008). *Theory and practice of psychological testing*. New Delhi: Oxford & Ibh Publishing.
- Frey, B. B. (2007). *An introduction to quality test construction*. Retrieved from <http://www.specis1connection.ku.edu/cgi-bin/cgiwrap/speccom/main.php?cat=assessment&section=main&section=qualitytest/main>.
- Gay, R. L. (1992). *Educational research: Competencies for analysis and application* (4<sup>th</sup> ed). New York: Macmillan Publishing Company.
- Ghana Statistical Service (2013). *District and housing census 2010: District analysis report*. Cape Coast, Central Region, Ghana: Sowah C. & Livingbird, B.
- Gipps, C. (1992a). *National testing at seven. What can it tells us?* London: Hodder and Stoughton Publishing Company.

- Gipps, C. (1992b). *What we know about effective primary teaching*. London: Tufnell Press.
- Gipps, C., J., Brown, M., McCallum, B., & McAlister, S. (1995). *Imitation or evidence? Teachers and national assessment of seven-year-olds*. Buckingham: Open University Press.
- Gipps, C. V. (1999). Socio-cultural aspects of assessment. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 355-392). Washington, DC: American Educational Research Association.
- Glassman, W. E. & Hadad, M. (2013). *Approaches to psychology* (6<sup>th</sup> ed.). New York: McGraw-Hills Education LTD.
- Goodwin, L. D., & Leetch, N. L. (2003). The meaning of validity in the new Standard for Educational and Psychological Testing: Implication for measurement courses. *Measurement and Evaluation in Counseling and Development, 35*, 181-183.
- Gregory, R. J. (1992). *Psychological testing: History, principles and application*. Boston: Allyn and Bacon.
- Gronlund, N. E. (2008). *How to construct achievement tests* (4<sup>th</sup> ed.). New Jersey: Prentice-Hall, Inc.
- Groth-Marnat, G. (1997). *Handbook on psychological assessment* (3<sup>rd</sup> ed.). New York: John Wiley & Sons.
- Gulliksen, H. (2001). Basic equations derived from a definition of random error. In H. Gulliksen (Ed), *Theory of mental tests* (pp. 4-28). Hillsdale, NJ: Lawrence Erlbaum.
- Gyimah, E. K. & Anane, E. (2013). *Assessment in schools*. Unpublished, University of Cape Coast, Cape Coast, Ghana.

- Gyimah, E. K., Ntim, E. K., & Deku, P. (2010). *Assessment techniques in special education*. Center for Continuing Education, University of Cape Coast. Cape Coast: Hamton Press Ltd.
- Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, 59 (3), 297–313.
- Izard, J. (2005). Overview of test construction. In K. N. Ross (ED). *Quantitative research methods in educational planning* (p. 7). Paris: UNESCO International Institute for Educational Planning.
- Johnson, R. N. & Synyby F. L. (2007). *A summary of published research: Classroom teachers' knowledge and skills related to the development and use of teacher-made tests*. Retrieved from <http://eric.ed.gov/ERICWeb>.
- Joint Committee on Standards for Educational and Psychological Testing (JCSEPT). (1999). *Standards for educational and psychological testing*. Washington D.C: America.
- Karpicke, J. D., & Roediger, H. L., (2008). The critical importance of retrieval for learning. *Science*, 15, 966 –968.
- Koksal, D. (2004). Assessing teachers' testing skills in Elt and enhancing their professional development through distance learning on the net. *Turkish Online J. Distance Education*, 5 (1), 32-44.
- Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, Massachusetts: Harvard University Press.
- Kothari, C. R. (2004). *Research methodology: Methods and techniques* (2<sup>nd</sup> ed.). New Delhi, India: New age international (P) LTD, Publishers.

- Knowles, M. S. (2000). *The modern practice of adult education: From pedagogy to andragogy*. Upper Saddle River Merrill, New York: Association Press, Pearson Education Inc.
- Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, 30, 607-610.
- Kubiszyn, T. & Borich, G. (1984). *Educational testing and measurement: Classroom application and practice*. New Jersey: Scott and Foresman Company.
- Leinhardt, G. (1983). Novice and expert knowledge of individual students' achievement. *Educational Psychologist*, 18, 165–179.
- Maizan, K. K. (2005). *Districts electives papers*. *Education week*. Retrieved from <http://www.edweek.org/ew/articles/1996/12/11>.
- Marso, R. N. & Pigge F. L. (1989). *Staff development implications from a state wide assessment of classroom teachers' testing skills and practice*. (ED312309). Retrieved from: <http://eric.ed.gov/ERICWebPortal>.
- McDaniel, E. (1994). *Understanding educational measurement*. Madison: Brown and Benchmark Publishers.
- McMillan, J. H., & Schumacher, S. (2006). *Research methods in education. A conceptual introduction* (5<sup>th</sup> ed.). New York: Longman.
- Mehrens, W. A. & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. New York: Harcourt Brace College Publishers.
- Mehrens, W. A. & Lehmann, I. J. (2001). *Measurement and evaluation in education and psychology* (4<sup>th</sup> ed.). New York: Holt, Rinehart and Winston Inc.

- Mehrens, W. A. & Lehmann, I. J. (2009). National tests and local curriculum: Match or mismatch. *Educational Measurement: Issues and Practices*, 3(3), 9-15.
- Messick, S. (2003). The psychology of educational measurement. *Journal of Educational Measurement*, 21, 215-237.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement*, Volum 3 (pp. 3-103). New York: American Council on Education, Macmillan.
- Miller, A. L, McIntire, A. S. & Lovler, L. R. (2011). *Foundation of psychological testing* (3<sup>rd</sup> ed.). Washington DC: Sage Publications Inc.
- Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62, 229–258.
- Nitko, A. J. (2001). *Educational tests and measurements* (3<sup>rd</sup> ed.). Upper Saddle River, New Jersey: Prentice-Hall, Inc.
- Northern Canadian Protocol for Collaboration in Education (2006). *Rethinking classroom assessment with purpose in mind*. Retrieved from <http://digitalcollection.gov.mb.ca/awweb/pdfopener?sms=1&did=12503>
- .Oduro, O. G. (2008). *Testing practices of Senior Secondary School teachers in the Ashanti Region*. Unpublished master's thesis. University of Cape Coast, Cape Coast, Ghana.
- Oescher, J. & Kirby P. C. (1998). *Assessing teacher-made tests in secondary mathematics and science classrooms*. (ED322169). Retrieve from <http://eric.ed.gov/ERICWebPortal/home>. on line.

- Oliver, P. (2010). *The student's guide to research ethics*. Berkshire: Open University Press.6, Retrieved from <http://smj.sma.org.sga9.pdf>.
- Osuola, E. C. (2001). *Introduction to research methodology* (3<sup>rd</sup> ed.). Onitsha, Nigeria: Africana F.E.P Publishers Ltd.
- Pallant, J. (2007). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows* (Version 10). Australia: Allen and Unwin.
- Pecku, N. K. (2000). *Formal assessment in the classroom: The Ghana Education Service termly assessment plan*. Paper presented to the Quality Improvement in Primary Schools (QUIPS) Project. Funded by the USAID, April 2000.
- Perry, N. E., & Meisels, S. J. (1996). *How accurate are teacher judgements of students' academic performance?* (Working Paper No. 96-08). Washington, DC: National Center for Educational Statistics.
- Polit, D. F., & Beck, C. T. (2004). *Nursing research: Principles and methods*. Philadelphia, PA: Lippincott Williams & Wilkins.
- Polit, D. F., & Hungler, B. P. (1999). *Nursing research: Principles and methods* (6<sup>th</sup> ed.). Philadelphia: Lippincott.
- Quagrain, A. K. (1992). *Teacher-competence in the use of essay type tests: A study of the secondary schools in the Western Region of Ghana*. Unpublished thesis. University of Cape Coast, Cape coast Ghana.
- Robson, C. (2002). *Real world research: A resource for social scientists and practitioners-researchers*. Oxford: Blackwell Publishers Ltd.
- Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education*, 17 (1), 1–

- Rudner, L. M. & Schafer, W. D. (Eds.). (2002). *What teachers need to know about assessment*. Washington, DC: National Education Association.
- Salvia, J., & Ysseldyke, J. E. (2001). *Assessment in special and remedial education* (3<sup>rd</sup> ed.). Boston: Houghton Mifflin.
- Sarantakos, S. (2005). *Social research* (3<sup>rd</sup> ed.). Melbourne: MacMillan Education.
- Seidman, I. H. (2006). *Interviewing as qualitative research: A guide for researchers in education and the social Sciences*. Columbia: Teachers College Press.
- Silker, R. T. (2003). *Teachers and tests*. London: Basil Blackwell
- Stern, W. (1928). *Intelligence of children and youth and its research methods* (4<sup>th</sup> ed.). Leipzig. [Citation Time(s):2].
- Stiggins, R., & Bridgeford, N. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22 (4), 271-286.
- Stiggins, R. J., & Conklin, N. F. (1992). *In teacher's hands: Investigating the practices of classroom assessment*. Albany: State University of New York Press.
- Stiggins, R. (1999). Assessment without victims. Retrieved from <http://www.nsd.org>. on line.
- Stiggins, R., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory into Practice*, 44(1), 11–18.
- Tamakloe, E. K. (2006). Principles and methods of teaching. Accra: Lack Mask Ltd. To close achievement gaps. *Theory into Practice*, 44(1), 19–26.

- Tamakloe, E. K., Atta, E. T. & Amedahe, F. K. (1996). *Principles and methods of teaching*. Cantoment., Accra: Black Mask.
- Tamakloe, E. K., Atta, E. T., & Amedahe, F. K. (2006). *Principles and methods of teaching* (2<sup>nd</sup> ed.). Accra: Black Mask Ltd. (p. 181).
- Tinkelman, S. N. (1971). Planning the objective test. In R. L. Thorndike (Ed.), *Educational measurement* (2<sup>nd</sup> ed.) (pp. 46-80). Washington, DC: American Council on Education.
- Tom, K., & Gary, D. B. (2003). *Educational testing and measurement: Classroom application and practice*. Hoboken, NJ: John Willey and Sons Inc.
- Trochim, W. M. K. (2006). *Introduction to validity: Social research methods*, Retrieved from [www.socialresearchmethods.net/kb/introval.php](http://www.socialresearchmethods.net/kb/introval.php).
- Trochim, W. (2000). *The research methods knowledge*. Cincinnati, OH: Atomic Dog Publishing.
- Wiliam, D. (1993). Validity, dependability and reliability in national curriculum assessment. *The Curriculum Journal*, 4(3), 335-350.
- Williams, J. M. (2008). Writing quality teacher-made tests: A handbook for teachers. *ERIC Document Reproduction Service*, No. 349 726.
- Wood, J. G. (2007). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.). *Review of research in education* (pp. 11-16). Washington, D.C.: American Educational Research Association.



**APPENDICES**

**APPENDIX A**

**UNIVERSITY OF CAPE COAST**

**COLLEGE OF EDUCATION STUDIES**

**DEPARTMENT OF EDUCATION AND PSYCHOLOGY**

**QUESTIONNAIRE FOR TEACHERS**

*Dear Respondent*

The purpose of this study is to investigate some underlying assessment practices (Test Construction, Administration and Scoring) of teachers in the Junior High Schools in Cape Coast Metropolis.

Kindly provide information on the items where applicable. Please respond to the statements as truthfully and honestly as you can. Please be assured of confidentiality and anonymity of any information you provide. This exercise is not to victimize respondents for the information given.

Thank you in advance for your cooperation.

**DIRECTIONS:** Please tick [] where appropriate in the spaces provided.

**SECTION A**

**Demographic Data of Respondent**

1. Highest Qualification: Diploma [  ] Bachelor of Education (BED) [  ] BA/BSc [  ] Masters [  ] Others [.....] Please specify
2. Class of Teaching: Form 1 [  ] Form 2 [  ] Form 3 [  ] Tick as many that are applicable.
3. Years of teaching Experience: 1-5 [  ] 6-10 [  ] 11-15 [  ] 16 and above [  ]

4. Subject(s) taught.....

## SECTION B

### CONSTRUCTION OF ACHIEVEMENT TEST

Please read the test formats below and indicate with a tick [✓] how often you use the following test format in assessing your students in the class. Indicate the extent to which the statement is Always- A, Very Often- VO, Often- O and Not Often- NO

Test Format	A	VO	O	NO
5. How often do you construct essay test item for your students?				
6. How often do you construct multiple choice test item for your students?				
7. How often do you construct short answer / fill in the blank test item for your students?				
8. How often do you construct matching blank test item for your students?				
9. How often do you construct True & False test items for your students?				

**SECTION C**

**TEACHERS PRACTICE IN CONSTRUCTION OF ACHIEVEMENT**

**TEST**

Please read the statement below and indicate with a tick */√/* your level of practice in test construction for your students in the class. Indicate the extent to which the statement is Always-A, Very Often- VO, Often- O and Not Often- NO,

Statement	A	VO	O	NO
<b>When constructing my test items I:</b>				
10. state the purpose of which the test will be used				
11. specify the construct to be measured				
12. use a test specification table				
13. match learning outcomes to the items				
14. construct items only when it is time to assess students				
15. set questions from past questions papers				
16. use questions from pupils text books				
17. ask any other colleagues to help me construct items				
18. ask other colleagues in the subject area to review the items by reading it over				
19. prepare marking scheme after the students have answered the questions				

20. consider meaning of wording against different ethnic background of the students				
21. do consider students language proficiency				
22. consider variation of students with respect to physical disabilities when writing the items				
23. consider the time individuals will spend on a question				
24. try solving the question myself to determine the time require for the questions				
25. provide clear and simple instructions on how test is to be answered				
26. evaluate items given to the students to answer				
27. write items at least two weeks before the date of testing				
28. write more items than needed before I review and select from them				
29. follow the principles of test construction for each format				

## SECTION D

### TEACHERS PRACTICE IN ADMINISTRATION OF TEST ITEMS

Please read the statements below and indicate with a tick  your level of practice in test administration for your students in the class. Indicate the extent to which the statement is Always-A, Very Often- VO, Often- O and Not Often- NO,

Statement	A	VO	O	NO
<b>In administering my test items, I:</b>				
30. make my students aware of the rules and regulations covering the conduct of the test				
31. ensure that there is adequate ventilation and lighting expected in the testing room				
32. make provision for extra answer sheets and writing materials during examination time				
33. make sure pupils start the test promptly and stop on time				
34. give more instructions during the time the students are taking the test				

35. inform students in advance the topics and the areas that the test will cover				
36. I prepare the classroom a day before the test is taking				
37. make sure tests are giving immediately before or after a long vacation, holidays or other important events				
38. inform the students about the test format (Essay-type/Objective type tests)				
39. make provision for emergencies during the time the test is taken				
40. proof read all my test items to eliminate ambiguities and unclear statement before administering the test				
41. use “Do not disturb examination in progress sign” at the entrance of the examination classroom				

## SECTION E

### TEACHERS PRACTICE IN SCORING OF TEST ITEMS

Please read the statements below and indicate with a tick  $/\surd/$  your level of practice in test scoring for your students in the class. Indicate the extent to which the statement is Always-A, Very Often- VO, Often- O, and Not Often- NO

Statements	A	VO	O	NO
<b>With regard to scoring, I:</b>				
42. promptly mark my students papers just after the test is taken				
43. usually prepare a form of scoring guide, either an analytic scoring rubric or a holistic scoring rubric				
44. make sure test takers are kept as anonymous as possible. That is different forms of identification are used instead of names				
45. grade the responses item by item and not script by script after the tests are taken				
46. keep scores of previously graded items out of sight when evaluating the rest of the items				

47. periodically rescore previously scored papers				
48. make sure the scripts are shuffled before starting to score				
49. score the essay test when I am physically sound, mentally alert and in an environment with very little or no distraction				
50. constantly follow the scoring guide as I score test items				
51. am influenced by the first few papers read when scoring tests items				
52. score a particular question or item on all papers at one sitting				
53. provide comments and errors corrected on the scripts for class tests to facilitate learning				
54. give extra marks to students besides the subject matter based on Handwriting, Gender, Grammatical expression, Length of students essay etc.				



**APPENDIX B**

**ETHICAL CLEARANCE FORM**

**UNIVERSITY OF CAPE COAST**  
COLLEGE OF EDUCATION STUDIES  
**ETHICAL REVIEW BOARD**



UNIVERSITY POST OFFICE  
CAPE COAST, GHANA

Our Ref: CES-ERB/ucc.edu/17/10

Your Ref: .....

Date: 14-09-17 .....

Chairman, CES-ERB  
Prof. J. A. Omotosho  
jomotosh@ucc.edu.gh  
0243784739

Dear Sir/Madam,

**ETHICAL REQUIREMENTS CLEARANCE FOR RESEARCH STUDY**

Vice-Chairman, CES-ERB  
Prof. K. Edjah  
kedjah@ucc.edu.gh  
0244742357

The bearer, Mr. Genest Ofoei Sasu ..... Reg. No ED/MEP/15/0002 is an M.Phil /Ph.D student in the Department of Education and Psychology..... College of Education Studies, University of Cape Coast, Cape Coast, Ghana. He/She wishes to undertake a research study on the topic An evaluation of testing practices of Junior High School teachers in the Cape Coast Metropolis.....

The Ethical Review Board (ERB) of the College of Education Studies (CES) has assessed the proposal submitted by the bearer. The said proposal satisfies the College's ethical requirements for the conduct of the study.

In view of the above, the researcher has been cleared and given approval to commence his/her study. The ERB would be grateful if you would give him/her the necessary assistance that may be needed to facilitate the conduct of the said research.

Secretary, CES-ERB  
Dr. (Mrs.) L. D. Forde  
lforde@ucc.edu.gh  
0244786680

Thank you.  
Yours sincerely,

Dr. (Mrs.) Linda Dzama Forde  
(Secretary, CES-ERB)

## APPENDIX C

### INTRODUCTORY LETTER

**UNIVERSITY OF CAPE COAST**  
COLLEGE OF EDUCATION STUDIES  
**DEPARTMENT OF EDUCATIONAL FOUNDATIONS**

Telephone: 233-3321-32440/4 & 32480/3  
Direct: 03321-36037  
Fax: 03321-30184  
Telex: 2552, UCC, GH.  
Telegram & Cables: University, Cape Coast  
Email: [edufound@ucc.edu.gh](mailto:edufound@ucc.edu.gh)



UNIVERSITY POST OFFICE  
CAPE COAST, GHANA

Our Ref:

Your Ref:

12<sup>th</sup> April, 2017

TO WHOM IT MAY CONCERN

**LETTER OF INTRODUCTION**  
**MR. ERNEST OFORI SASU**

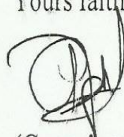
We confirm that the above-mentioned name is an M.Phil Educational Measurement and Evaluation student at the Department of Education and Psychology, UCC.

As part of the Programme requirement he is writing theses entitled "Evaluation of Junior High School Teachers on their testing practices in Cape-Coast Metropolis".

We would be very grateful if you could assist him with any information he may need for his research. All information retrieved would be treated confidentially.

Thank you.

Yours faithfully,



(Georgina Nyantakyiwaa Thompson)  
Principal Administrative Assistant  
For: Head

## APPENDIX D

### GES PERMISSION NOTE

# GHANA EDUCATION SERVICE

*In case of reply the  
Number and date of this  
Letter should be quoted*



REPUBLIC OF GHANA

METROPOLITAN EDUCATION DIRECTORATE  
P. O. BOX 164  
CAPE COAST

Tel. 03321-32514/33405  
Fax 03321-32199  
Email: capecoastmeo@yahoo.com  
My Ref. No GES/MD/SF.1/VOL. 4/100

11<sup>th</sup> May, 2017

**HEADTEACHERS CONCERNED  
CAPE COAST METROPOLIS**

#### RE: PERMISSION TO CONDUCT RESEARCH

We write to inform you that the Metro Education Directorate has granted permission to Mr. Ernest Ofori Sasu an M.Phil Educational Measurement and Evaluation student at the Department of Education and Psychology University of Cape Coast to conduct a study in your school.

He is conducting a research on "Evaluation of JHS Teachers on their testing practices in Cape Coast Metropolis.

You should however ensure that his study will not interfere with normal teaching and learning activities.

Please, accord him the necessary assistance to make his study successful.

Thank you.

STEPHEN RICHARD AMOAH  
METRO DIRECTOR OF EDUCATION  
CAPE COAST

HEADTEACHER  
KWEGYIR AGGREY BASIC SCHOOL  
P. O. BOX AD 1027  
CAPE COAST

cc  
Mr. Ernest Ofori Sasu

## APPENDIX E

### INFORM CONSENT FORM

#### TESTING PRACTICES OF JUNIOR HIGH SCHOOL TEACHERS IN THE CAPECOAST METROPOLIS

Please read this consent document carefully before you decide to participate in this study. The purpose of this study is to examine testing practices of Junior High School teachers in Cape Coast Metropolis. Please candidly respond to the questionnaire items regarding testing practices of teachers. This questionnaire will take you 20 minutes to complete. There are no foreseeable risks associated with this study.

**Confidentiality:** Your identity will be kept confidential to the extent provided by law. Your information will be assigned a code number.

**Voluntary participation:** Your participation in this study is completely voluntary. There is no penalty for not participating.

**Right to withdraw from the study:** You have the right to withdraw from the study at any time without consequence.

**Agreement:** I have read the procedure described above. I voluntarily agree to participate in the procedure and I have received a copy of this description.

Name \_\_\_\_\_

Signature \_\_\_\_\_

Date: \_\_\_\_\_

**APPENDIX F**  
**RELIABILITY TEST**

**Case Processing Summary**

	N	%
Valid	299	99.7
Excluded <sup>a</sup>	1	.3
Total	300	100.0

a. Listwise deletion based on all variables in the procedure.

**Reliability Statistics**

Cronbach's Alpha	N of Items
.873	50