UNIVERSITY OF CAPE COAST

MULTIPLE-CHOICE CONSTRUCTION COMPETENCIES AND ITEMS' QUALITY: EVIDENCE FROM SELECTED SENIOR HIGH SCHOOL SUBJECT TEACHERS IN KWAHU-SOUTH DISTRICT

PROSPER KISSI

2020

UNIVERSITY OF CAPE COAST

MULTIPLE-CHOICE CONSTRUCTION COMPETENCIES AND ITEMS' QUALITY: EVIDENCE FROM SELECTED SENIOR HIGH SCHOOL SUBJECT TEACHERS IN KWAHU-SOUTH DISTRICT

BY

PROSPER KISSI

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 for the award of Master of Philosophy degree in Measurement and Evaluation

JULY 2020

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: ……………………….……………………………….……………………

Supervisor's Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines of 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 purpose of the study was to explore the relationship between the multiple-choice test construction competencies of senior high school teachers in the Kwahu-South District and the quality of multiple-choice test items they construct, as well as the effect of years of teaching on such relationship. To examine the relationship among the variables under investigation, the quantitative approach is employed using correlational research design. The total number of teachers that constituted the population for the study was 157. However, with the use of purposive sampling technique, the study covered only 47 participants (n = 47) out of the 157 teachers. Questionnaire and document examination were used as the main data collection instruments. The overall reliability coefficient of the 20-item questionnaire using Cronbach's alpha was .75. The data collected was analysed using means and standard deviations, frequency count, percentages, and Pearson Product-Moment Correlation Coefficient. Finding from the study revealed no significant relationship between the teachers' test construction competencies and the quality of the multiple-choice test items. This implied that the presence of problem items influenced the quality of the multiple-choice tests in a manner that it showed no significant relationship with the teachers' self-reported appreciable or high levels of multiple-choice test construction competencies. Thus, it was recommended that school authorities and classroom teachers pay critical attention to factors that reduce the quality of multiple-choice items and put up measures that will improve the reliability and validity of results obtained from assessments that involve the use of multiple-choice test items.

KEYWORDS

Assessment competencies

Quality of multiple-choice items

Reliability and validity

Test construction competence

Test length

Years of teaching

ACKNOWLEDGEMENTS

I extend my heartfelt thank you to Dr. Kenneth Asamoah-Gyimah my principal supervisor and Dr. Eric Anane my co-supervisor, who through their selfless patience have tirelessly and energetically devoted countless number of hours to the supervision of this work. My profound and sincere gratitude is also extended to Dr. Stephen Doh Fia, Mr. Edward Danso, Dr. Priscilla Commey Mintah, and Mrs. Georgina Nyantakyiwaa Thompson for their diverse and unflinching support which has helped me to realise my academic aspirations.

I also wish to show my appreciation to the leadership and the vetting committee of the Samuel and Emelia Brew-Butler Research Fund who deemed me worthy for support from their research fund.

I would like to express my heartfelt thanks to Ms. Ruth Annan Brew and Ms. Regina Mawusi Nugba for their gallant support and encouragement which contributed to the completion of this work. Benedicta Ama Yekua Etuaful, your remarkable assistance during the period of my data entry and analysis is duly acknowledged. Also, I thank all the teachers who graciously agreed to participate in this study.

Lastly, I thank my parents, my brothers, Eric Owusu Mensah Dankwa, Alexander Sono Acheampong, my sisters, Rita Agyeiwaa Acheampong and Regina Ohui for the interest they have taken in my studies and for supporting and encouraging me throughout my life.

DEDICATION

To my mother, Joyce Omari

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS/ACRONYMS

AFT	American Federation of Teachers
BECE	Basic Education Certificate Examination
CTT	Classical Test Theory
GES	Ghana Education Service
GPA	Grade Point Average
KED	Kwahu-East District
KSD	Kwahu-South District
NCME	National Council on Measurement in Education
NEA	National Education Association
PA	Parallel Analysis
PCA	Principal Component Analysis
PPMCC	Pearson Product-Moment Correlation Coefficient
SHS	Senior High School
SHSs	Senior High Schools
TTCCQ-MC	Teachers' Multiple-Choice Test Construction Competence Questionnaire
TCSI	Test Construction Skill Inventory
WASSCE	West African Senior School Certificate Examination

CHAPTER ONE

INTRODUCTION

In the field of education, teaching and learning will not be effective without adequate knowledge on how much and how well students have been able to acquire the necessary knowledge, skills and certain abilities after instructional period(s). Assessment which is an integral part of teaching and learning processes provides teachers, students and stakeholders with such knowledge. However, research studies (outside Ghana and within Ghana) have indicated that classroom teachers encounter difficulties in assessing students learning outcomes. Schools, teachers, students and the society at large will not be exempted from the effects of wrongful educational decisions when invalid interpretations and uses of assessment results are being made. In Ghana, though this problem about assessment of students' achievement has been explored and investigated within certain contexts, there is the need to investigate further issues relating to classroom assessment in the country, especially within certain contexts where the issue appears unexplored and not investigated.

Background to the Study

Imagine a situation where the individuals within a particular society do not possess adequate skills and knowledge for problem-solving and making effective decisions. Conceive the nature of a society where people are not able to read and write, communicate effectively, and lack basic mathematical abilities like adding, subtracting, dividing and multiplying. Within such a community, development becomes something difficult to achieve. Therefore, to foster development in a given country, education becomes inevitable. Education is an important measure when it comes to the development of human capital; it is also linked with an individuals' well-being and opportunities for better living (Battle & Lewis, 2002). It is generally believed that the quality of education available to citizens of any given country influences its development. Adane (2013) emphasised this by saying that the foundation for any solid development must begin with the development of human resources.

To develop human resources, educational goals are established based on the wants and needs of society (Butler, McColskey, & O'Sullivan, 2005; Kubiszyn & Borich, 2013; Nitko, 2001). These goals are then translated into learning outcomes. Specific instructional objectives are also set to meet the various learning outcomes. Instruction becomes the means by which students acquire the learnable bits of information under the necessary conditions (Kubiszyn & Borich, 2013; Nitko, 2001). Now, the question is: Since students' academic achievement and performance, in a way, is representative of the resources they possess as an essential asset for social and economic development, how is one sure that students possess the required knowledge, skills and abilities to fit into a particular society? This issue of interest endorses assessment as an essential and integral part of the teaching and learning process.

Assessment is defined as the formal process of collecting, analysing, and reporting standardised information about students' knowledge, skills, and abilities (Bunch, 2012). However, according to the American Federation of

Teachers, National Council on Measurement in Education, and National Education Association (AFT, NCME, & NEA, as cited in Nitko, 2001), assessment is not just reporting information gathered; it is also about using such information for making relevant decisions concerning students, curricula and programmes, and educational policy. The information, thus, obtained should be reliable and valid for its intended purpose(s) and use(s). Reliability refers to the consistency and stability of the assessment results (Nitko, 2001). Validity, on the other hand, refers to the soundness of the interpretation and use of assessment results (Nitko, 2001). By implication, any form of test constructed should portray these qualities whether it is a standardised or non-standardised/teacher-made.

According to Harris (2002), specifically, a standardised test is any measure that is useful in evaluating characteristics or skills of students, with specific procedures for administering and scoring the tests and interpreting the assessment results; they are usually developed by test construction professionals or experts. In the Ghanaian educational context, with summative assessment, examples of standardised tests at the basic level and secondary level are Basic Education Certificate Examination (BECE) and West African Senior School Certificate Examination (WASSCE) respectively.

On the other hand, non-standardised measures which are often referred to as teacher-made tests are constructed, administered, and scored by classroom teachers, and often consist of completion, true-false, matching, multiple-choice, and essay items (Kubiszyn & Borich, 2013). These teacher-made tests are often flexible, or variable, in terms of their administration and scoring procedures, and in the amount of attention given to their construction.

Different teachers may be more or less careful in constructing their tests, may allow more or less time for the test to be taken, and may be more or less stringent in grading the test (Kubiszyn & Borich, 2013).

Amedahe (2014) has indicated that at the basic and secondary levels, Ghana's educational system lacks standardised measures for assessing and monitoring pupils' or students' performance at the various grade levels. He further indicated that the only standardised assessment instrument at the basic level is the National Education Assessment (NEA) for Primary 3 and 6 in mathematics and English language administered to a sample of between 3-5% of the population every two years. Therefore, the Ghanaian educational system depends largely on teacher-made tests when it comes to classroom assessment (Amedahe, 2014; Asamoah-Gyimah, 2002). McMillan (2013) defined classroom assessment as:

> A broad and evolving conceptualization of a process that teachers and students use in collecting, evaluating and using evidence of student learning for a variety of purposes, including diagnosing student strengths and weaknesses, monitoring student progress towards meeting desired levels of proficiency, assigning grades, and providing feedback to parents. (p.4).

Classroom assessment which relies greatly on teacher-made tests plays an increasingly crucial role in the field of education. That is, teacher-made tests aid in pre-assessment (that is, the assessment of what student already know prior to teaching); formative assessment (which is the assessment of student performance incorporated into the act of teaching); and summative assessment (the assessment of student learning at the end of some instructional period) of pupils' or students' learning outcomes (Gareis & Grant, 2015), which, in turn, informs relevant educational decisions. Therefore, the essentiality for teachers to understand and utilise classroom assessments is greater than ever before (Guskey, 2003; Guskey & Jung, 2013). That is, teachers must be as proficient and competent in the area of assessment as they have traditionally been in the areas of curriculum and instruction (Gareis & Grant, 2015). This draws our attention to an important construct namely assessment competency.

What is assessment competency? Since assessment is an important aspect of the activities of teaching, the concept of assessment competence can be inferred from Adodo's (2013) definition of competency in teaching. Competency in teaching refers to the ability of a teacher to exhibit on the job skills and knowledge gained as a result of training (Adodo, 2013). Inferred, competency in assessment or assessment competency can be explained as the ability of a teacher to exhibit or apply knowledge and skills gained as a result of training in assessment.

Teachers' competencies in assessment are specified in standards for teacher competence in educational assessment of students. The standards express specific expectations for assessing knowledge or skills that teachers should possess in order to perform well in their evaluation effort or assessment of students (Ololube, 2008). The standards as developed by AFT, NCME, and NEA (as cited in Nitko, 2001) are that teachers should be skilled in:

1. selecting assessment procedures suitable for instructional decisions.
2. developing assessment techniques suitable for instructional decisions

3. administering, marking, and interpreting the results of both externally-produced and teacher-produced assessment procedures.
4. using results obtained from assessment when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5. developing valid student or pupil grading procedures which use student or pupil assessments.
6. communicating assessment results to students, parents, other educators and other lay audiences.
7. acknowledging illegal, unethical, and otherwise inappropriate assessment procedures and uses of assessment information.

Outside Ghana, based on observed classroom assessment nonconformities to the aforementioned standards (or other related standards) on teacher competence in assessment of students, a number of studies have been conducted. These studies investigated classroom teacher's assessment competency, assessment practices, the quality of the instruments they develop, and some factors that might account for differences in their assessment competencies, practices and effectiveness and quality of tests they construct (Agu, Onyekuba, & Anyichie, 2013; Alkharusi, 2011; Hamafyelto, Hamman-Tukur & Hamafyelto, 2015; Harpster, 1999; Kinyua & Okunya, 2014; Magno, 2003; Marso & Pigge, 1989; Ovat & Ofem, 2017; Tshabalala, Mapolisa, Gazimbe, & Ncube, 2015). For example, in Matabeleland North (Western Zimbabwe), Tshabalala, Mapolisa, Gazimbe and Ncube (2015) conducted a study which bears the title: Establishing the effectiveness of teacher-made tests in Nkayi district primary schools. In their study, they found that most of the teachers did know the standard procedures of constructing, marking, and grading tests.

From the standards on teachers' competencies in assessment, it is evident that teachers' competencies in the construction of instrument that will provide more valid and reliable assessment results are an aspect of what goes into the activities of classroom assessment. According to Hamafyelto et al. (2015), teachers' test construction competencies and the quality of the assessment tools developed are important to the achievement of teaching and learning goals. Nevertheless, what observations have been made in Ghana with regard to the standards that are descriptive of teachers' test construction competencies in assessment of students?

It has been unravelled in the Ghanaian educational settings that most classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction (Amedahe, 1989; Anhwere, 2009; Wiredu, 2013; Quaigrain, 1992; Sasu, 2017). According to Amedahe (1989), to a large extent, secondary school teachers in the Central Region did not follow the basic suggested principles of classroom test construction. Additionally, Quaigrain (1992) indicated that most Ghanaian teachers had inadequate skills for constructing essay type tests. Moreover, Teacher Training College tutors did not adhere to the basic principle of testing in the construction of classroom tests or teacher-made test (Anhwere, 2009). Wiredu (2013) also found that tutors in Nurses' Training Colleges in the Western and Central Regions of Ghana overlooked some basic principles in crafting test items. Also, it was found that Junior High School teachers in the Cape Coast Metropolis did not follow test construction principles to an appreciable level (Sasu, 2017). These findings are tentative answers related to some aspects of Ghanaian teachers' test construction competencies.

On the subject of studies on Ghanaian teachers' test construction competencies, Oduro-Okyireh (2008) has given contradicting evidence that teachers in Senior High Schools (SHSs) in the Ashanti Region of Ghana follow the principles of test construction. Looking at the different populations previous studies have examined and the mixed nature of findings on some aspects of test construction competencies, it is clear that these studies do not give holistic and consistent view of test construction competencies of teachers in the Ghanaian educational settings. Thus, in Ghana, previous studies related to teachers' test construction competencies arouse curiosity about test construction competencies of teachers in other educational settings. Hence, research is needed to investigate test construction competencies of other populations of teachers in Ghana. Senior High School (SHS) teachers in the Kwahu-South District (KSD, in the Eastern Region of Ghana) is one of these populations. KSD because, in descriptive terms, it is not certain the state of test construction competencies of SHS teachers in the area.

Ghanaian classroom trained and untrained teachers, from the basic level to the university level, construct, administer and score classroom achievement tests regardless of whether they have had training in measurement and evaluation or not (Anhwere, 2009). When classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction, the quality of the tests they construct is questionable. According to Chau (as cited in Hamafyelto et al., 2015), teacher's test construction competence is directly related to ensuring the quality of a test.

Poor test quality negatively affects the validity of assessment results (Amedahe & Asamoah-Gyimah, 2016).

From the aforementioned, by implication, when teacher-made tests are low in quality, school administrators and classroom teachers will not be able to provide support and educational opportunities that each student needs (Agu et al., 2013). In other words, lack of or low degree of validity of test results leads to undependable inferences about student learning (Amedahe & Asamoah-Gyimah, 2016; Gareis & Grant, 2015). Based on this, educational decisions such as selection of students for educational opportunities would be wrongfully made.

Questionnaire as a self-report measure has been a common instrument that has been used to investigate test construction competencies or practices of classroom teachers in Ghana (see Amedahe, 1989; Quaigrain 1992; Oduro-Okyireh, 2008; Anhwere, 2009; Wiredu, 2013; Sasu, 2017). Amedahe (1989), Quaigrain (1992), and Wiredu (2013) verified teachers' responses to questionnaire items by directly examining samples of tests developed by the teachers for constructional flaws. The direct examination provided some qualitative information with regard to the quality of the teacher-made tests. However, the study of Oduro-Okyireh (2008), Anhwere (2009), and Sasu (2017) involved no direct analysis of samples of teachers-made tests to verify teachers' responses to questionnaire items on test construction practices.

To go beyond just relying on responses of teachers on self-report measures, Oduro-Okyireh (2008) suggested research be conducted on the quality of teacher-made tests. The quality of tests developed is investigated using item analysis. Therefore, Oduro-Okyireh's suggestion implies a direct assessment of actual test made by classroom teachers. The direct assessment procedure will help to validate teachers' responses to any self-report measure used in the assessment of their practices or competencies. Ary, Jacobs, Sorensen and Razavieh (2010) have identified that direct observation of the behaviour of a random sample of respondents is a brilliant strategy to validate their responses to a self-report measure. There are two main approaches to item analysis –qualitative and quantitative approaches (Kubiszyn & Borich, 2013).

In Ghana, in terms of evaluating the quality of teacher-made test items, from previous studies, none of the item analysis approaches has empirically been employed to investigate the quality of test items constructed by classroom teachers in the KSD. For example, Quaigrain and Arhin (2017) conducted quantitative item analysis study which focused on item and test quality and explored the relationship between difficulty index and discrimination index with distractor efficiency. However, the study was conducted among first-year students pursuing Diploma in Education at Cape Coast Polytechnic. Accordingly, there was the need to also conduct an item analysis study among SHS teachers in the KSD in order to validate their responses to the self-report measure that was administered.

Quantitative item analysis is more feasible and useful for multiple-choice test items than essay items (Kubiszyn & Borich, 2013). Therefore, in this study, the quality of multiple-choice test items constructed by classroom teachers was investigated excluding any other item formats. The output from quantitative item analysis on the multiple-choice test items require some form of explanations; therefore, qualitative item analysis approach was also employed to identify certain errors associated with the construction of teacher-made multiple-choice items among SHS teachers in the KSD.

According to Chau (as cited in Hamafyelto et al., 2015), teacher's test construction competence is directly related to ensuring the quality of a test. Nevertheless, it appears no study has been conducted to quantitatively examine Chau's perspective by finding out the relationship between multiple-choice test construction competencies and the quality of the test items among SHS teachers in Ghana. Accordingly, the relationship between multiple-choice test construction competencies and the quality of the teacher-made test items among SHS teachers in the KSD was investigated.

According to Leedy and Ormrod (2013), it is vital to investigate variable(s) that might help explain the relationship between two variables under investigation. Years of teaching has been identified as a variable that may influence test construction competence and quality of test items (Amedahe, 1989; Marso & Pigge, 1989; Dosumu, 2002; Magno, 2003; Agu et al., 2013; Kinyua & Okunya, 2014). However, in Ghana, it appears previous studies have not looked at the effect of years of teaching on the relationship between test construction competence and quality of test items. Therefore, for this work, there was the need to investigate the effect of years of teaching on the relationship between multiple-choice test construction competencies and the quality of the items among SHS teachers in the KSD.

It has been observed in the literature that to improve the reliability and validity of assessment results, there is the need to lengthen the assessment procedures (Allen & Yen, 2002; Crocker & Algina, 2008; Nitko, 2001). This implies that one should put little confidence in a student's performance based on few multiple-choice items used in assessing student's achievement (Crocker & Algina, 2008). However, Crocker and Algina (2008) have stated that improving test quality by increasing test length works when all the items are representative of the domain sampled and have the same or similar level of appropriate difficulty and discrimination indices. This means that the greater the number of test items appropriate in difficulty and discrimination indices, the degree of errors in students observed scores reduces; hence, resulting in appreciable level of reliability coefficient.

Reliability coefficient is the squared correlation between observed scores and true scores (Furr & Bacharach, 2014). This suggests that reliability coefficient of the test will reduce when there are more problem items contributing to errors in observed scores. On the other hand, the presence of more good items reduces the errors; therefore, resulting in improved reliability coefficient (Crocker & Algina, 2008). Therefore, Crocker and Algina (2008) have stated that items must be well constructed and free of technical flaws that may cause examinees to respond on some basis not related to the content.

From the aforementioned, to achieve an appreciable level of reliability coefficient, the theory endorses that it is good for teachers to construct relatively adequate number of test items without problem items or which has relatively few problem items. Nevertheless, Downing (2003) has stated that teachers perceive test construction procedures as waste of time and non-motivating. Such a line of thinking can negatively influence their level of attention in ensuring that test items are not problem items or there are few of them. This calls for the need to investigate among classroom teachers what happens to: (a) the number of good items in a test when test length is either increasing or decreasing; (b) the number of problem items as test length is either increasing or decreasing.

Nonetheless, in Ghana, previous studies (examples: Amedahe, 1989; Oduro-Okyireh, 2008; Anhwere, 2009; Wiredu, 2013; Sasu, 2017) conducted in relation to test construction have not examined the relationship between test length and the number of good and problem items identified with tests constructed by classroom teachers. Thus, there is the need to examine the relationship between: (a) test length and the number of good items; and (b) test length and the number of problem items observed with items constructed by classroom teachers in the country.

On the topic of test construction competencies and practices, previous studies conducted in Ghana at the SHS level have covered English Language, Mathematics, History, Geography, and Religious Studies teachers. Amedahe's (1989) study was carried out among English Language, Mathematics and History teachers. To refute or confirm the tentative findings of his study, he recommended that an extensive research is needed to cover most of the subject teachers in the SHSs. So far, it appears only two studies (Oduro-Okyireh, 2008; Quaigrain, 1992) have investigated test construction competencies and practices of teachers at the SHS level. Quaigrain's (1992) study focused on History, Geography, Religious Studies and English Language teachers, while Oduro-Okyireh's (2008) study focused on English Language, Core Mathematics and Integrated Science teachers.

Most of the subjects taught at the SHS level have not been covered by previous works on test construction competencies of SHS teachers in Ghana. It is against this background that this present study was extended to cover SHS teachers in the KSD who teach Financial Accounting, Cost Accounting, Business Management and Economics. These subject teachers were selected in addition to English Language, Core Mathematics, and Integrated Science teachers in the KSD SHSs.

Statement of the Problem

Most of the studies (Amedahe, 1989; Anhwere, 2009; Wiredu, 2013; Quaigrain, 1992; Sasu, 2017) conducted about teachers' test construction competencies have revealed that in the Ghanaian educational system, classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction. However, contradicting evidence given by Oduro-Okyireh (2008) exists in the literature. This situation arouses curiosity about test construction competencies and characteristics or quality of teacher-made tests in other educational settings in the country.

Amedahe (1989) made a recommendation that extensive research is needed to explore teachers' test construction competencies and the quality of teacher-made tests in Ghana. Previous studies conducted in the country (examples: Amedahe, 1989; Oduro-Okyireh, 2008; Anhwere, 2009; Wiredu, 2013; Sasu, 2017) give the impression that the extent to which issues related to teachers' test construction competencies and quality of test items has been explored in Ghana is low. Following Amedahe's recommendation, empirical evidence is needed concerning the test construction competencies and quality of test items constructed by classroom teachers in the country, especially within educational contexts where these issues seem unexplored.

Exploring test construction competencies of classroom teachers is very essential. This is because Chau (as cited in Hamafyelto et al., 2015) have stated that teacher's test construction competence is directly related to ensuring the quality of a test or a test has good characteristics. Based on this premise, when classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction, it will result in crafting items with low quality, which, in turn, negatively affect the validity of assessment results (Amedahe & Asamoah-Gyimah, 2016).

According to Agu et al. (2013), when tests constructed by classroom teachers are low in quality, school administrators and teachers are not able to make available support and educational opportunities that each student needs. For instance, achievement test of good quality will serve the purpose of helping the teacher to know students who have mastered a given content and those who have not (Joshua, 2005; Kubiszyn & Borich, 2013; Nitko, 2001). Reliable information concerning students who have not gain mastery over certain content areas can help teachers and educators to provide educational support (for example, extra classes) so that students' achievement can be maximised (Nitko, 2001). However, if most of the items are not able to spell out the differences, students who have not gained mastery could be classified as part of students who achieved an appreciable level of knowledge with respect to instructional objectives. This will make them miss any form of educational support that could have improved their achievement.

Conclusions based on the literature concerning test construction competencies and quality of teacher-made tests (that is, Amedahe, 1989; Marso & Pigge, 1989; Dosumu, 2002; Magno, 2003; Agu et al., 2013; Kinyua & Okunya, 2014) give the impression that though test construction competencies might be related to the quality of test items, years of teaching might serve as a third variable that may influence such a relationship. Moreover, another issue that is related to test construction competence and test quality is the number of test items teachers are able to construct based on their test construction competencies, and how the number of items generated (test length) is related to the number of good items and problem items identified with the tests they construct.

Nevertheless, from previous works examined in the background to this study, in the Ghanaian educational settings, research conducted in the area of testing practices in secondary schools in the Central Region (Amedahe, 1989); Teacher-competence in the use of essay tests: A study of secondary schools in the Western Region (Quaigrain, 1992); Testing practices of SHS teachers in the Ashanti Region (Oduro-Okyireh, 2008); Assessment practices of Teacher Training College tutors (Anhwere, 2009); Assessment practices of tutors in the Nurses' Training Colleges in Western and Central Regions of Ghana (Wiredu, 2013); Testing practices of Junior High School teachers in the Cape Coast Metropolis (Sasu, 2017); and Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation among first-year students pursuing Diploma in Education at Cape Coast Polytechnic (Quaigrain & Arhin, 2017) appears to leave the following research gaps to be filled in this research:

1. There is the need to explore the relationship between multiple-choice test construction competencies and the quality of multiple-choice test items, as well as effect of years of teaching on such relationship among SHS teachers in the KSD.

2. There is also the need to explore the relationship between test length and the number of good items and problem items produced by SHS teachers in the KSD.

Purpose of the Study

The purpose of the study was to explore the relationship between the multiple-choice test construction competencies of senior high school teachers in the KSD and the quality of multiple-choice test items they construct, as well as the effect of years of teaching on such relationship.

Research Objectives

Specifically, the study sought to:

1. describe multiple-choice test construction competencies of teachers in assessing students learning outcomes at the senior high school level in the KSD;
2. establish the characteristics of the multiple-choice test items constructed by the teachers in the KSD based on the following criteria: difficulty index, and discrimination index;
3. establish the characteristics of the multiple-choice tests in terms of format and item construction errors associated with teacher-made multiple-choice tests in the KSD;
4. establish the characteristics of the multiple-choice tests by examining the relationship between test length and the number of good items and problem items produced by senior high school teachers in the KSD;
5. explore the relationship between multiple-choice test construction competencies of teachers and the quality of multiple-choice test items in the KSD;

6. explore the effect of teachers' years of teaching on the relationship between multiple-choice test construction competencies of teachers and the quality of multiple-choice test items in the KSD.

Research Questions

Research questions that were addressed in this study include:

1. What multiple-choice test construction competencies do teachers have in assessing students learning outcomes at the senior high school level in the Kwahu-South District?
2. What are the characteristics of the multiple-choice test items based on the following criteria: difficulty index, and discrimination index in the Kwahu-South District?
3. What are the types of error associated with teacher-made multiple-choice tests among senior high school teachers in the Kwahu-South District construct?

Research Hypotheses

Four research hypotheses were tested in this study:

Hypothesis 1

Ho : There is no statistically significant relationship between test length and the number of good items produced by senior high school teachers in the Kwahu-South District.

H_A : There is statistically significant relationship between test length and the number of good items produced by senior high school teachers in the Kwahu-South District.

Hypothesis 2

Ho : There is no statistically significant relationship between test length and the number of problem items found in the multiple-choice test items constructed by senior high school teachers in the Kwahu-South District.

H_A : There is statistically significant relationship between test length and the number of problem items found in the multiple-choice test items constructed by senior high school teachers in the Kwahu-South District.

Hypothesis 3

Ho : There is no statistically significant relationship between multiple-choice test construction competencies of teachers and the quality of multiple-choice test items in the Kwahu-South District.

H_A : There is statistically significant relationship between multiple-choice test construction competencies of teachers and the quality of multiple-choice test items in the Kwahu-South District.

Hypothesis 4

Ho : Teachers' years of teaching has no statistically significant effect on the relationship between multiple-choice test construction competencies of teachers and the quality of multiple-choice test items in the Kwahu-South District.

H_A : Teachers' years of teaching has statistically significant effect on the relationship between multiple-choice test construction competencies of teachers and the quality of multiple-choice test items in the Kwahu-South District.

Significance of the Study

It is envisaged that, empirical evidence obtained by establishing the characteristics of the multiple-choice tests and exploring the relationship between multiple-choice test construction competencies and the quality of multiple-choice test items, as well as the effect of their years of teaching on such relationship among the teachers will help stakeholders to put appropriate measures or educational programmes leading to sustainable or improved test construction competencies and quality of multiple-choice test items. For instance, the District Directorate of Ghana Education Service could organise training sessions or training programmes in test construction to help sustain or improve teachers' test construction competencies in the participating schools. This will help the teachers to generate well-constructed test items that will aid in ascertaining lapses in the students' acquisition of knowledge, so that the necessary educational support can be given to help students maximise their academic achievement.

Moreover, it is hoped that the study will complement studies that have already been undertaken elsewhere in this subject matter. That is, the study will add to the spectrum of knowledge for teachers' assessment competencies in Ghana.

The study shall also be of benefit in that the findings of this study would serve students, educationists, and experts in measurement as an important reference source for further studies. For instance, recommendations made in this study would be a good source of research problems for further studies on the concept of quality of teacher-made tests.

Delimitations

Assessment competencies encompass the activities of choosing, developing or constructing test, administering and scoring the test, interpreting and using assessment results for relevant educational decisions in an ethically and legally acceptable manner. However, this research is delimited to explore the aspect of competencies in constructing the test. Thus, the relationship between multiple-choice test construction competencies and the quality of the multiple-choice test items constructed by form 1, form 2 and form 3 Financial Accounting, Cost Accounting, Business Management, Economics, English Language, Integrated Science and Core Mathematics SHS teachers in the KSD, as well as the effect of their years of teaching on such relationship was investigated.

The KSD was selected for the study because the nature of the study in terms of data collection, made it necessary to consider and use my existing rapport with some of the headteachers and teachers in the district who could act as gatekeepers towards successful data collection. Researchers agree that positive relationships or rapport with gatekeepers are very essential towards selecting or gaining access to the site or study area, and potential research participants (Feldman, Bell, & Berger, 2003). According to Neuman (2014) most study areas have gatekeepers who are the people having formal or informal authority to control access to a study area. They are the people members in the study area obey, whether or not they have official titles. It is ethically judicious to call on gatekeepers who are willing to give access and assist with cooperation of potential respondents for a study (Neuman, 2014).

It is imperative to add that the SHS teachers in the district have similar characteristics with other population of teachers in the country in the sense that they construct test items, administer, score and interpret students' assessment results. Therefore, they served as good sub-population for exploring the variables of interest as evident in the research questions and hypotheses.

In addition, the study involved scanning of over 1500 students' responses to multiple-choice tests items, copies of end-of-semester teacher-made achievement tests, and marking schemes. Therefore, data collection was anticipated to last for about a period of two-months. To have cost effective accommodation, easy access to study area, research participants, space for scanning documents, successful data collection within the period, there was also the need to comparatively and conveniently use KSD as the study area.

The 'teachers' test construction competencies' was an important variable to be investigated by this research because when the test is faulty or full of format and constructional flaws, its quality in detecting individual differences would be questionable. Moreover, the decisions that will be taken per the information gathered through its administration will be invalid, less valid, or undependable.

It is evident that not all SHS subject teachers in the KSD were included in this study. Covering all SHS subject teachers in the KSD might be of interest, however, it was not feasible to provide the current state of their multiple-choice test construction competencies. That is, direct assessment of test items is time consuming using item analysis approach. Therefore, per the period given for the completion of this research, it was impractical to validate the responses of all the subject teachers to the questionnaires with respect to directly analysing samples of their multiple-choice test items.

Students were not included in the study. This is because concerning the purpose of the study, it was required that data be collected from the classroom teachers. Besides, the nature of analysis to answer research questions and hypotheses did not require any form of responses from students. This is justified on the grounds that standards for evaluating quality of the test items constructed by the classroom teachers required subject area experts and experts with background in educational measurement and evaluation or test construction. Nonetheless, future studies can include students who could give their perceptions concerning teacher-made tests when there is the need to.

Limitations

In this research, the correlational research design was used for the purpose of examining the degree of relationship between teachers' test construction competencies and the quality of the items they constructed, as well as examining the effect of years of teaching on such relationship. Thus, the results and research findings just give an idea of cause and effect and does not establish causation.

The total number of teachers that constituted the population for the study was 157. However, with the use of purposive sampling technique, the study covered only 47 participants (n = 47) out of the 157 teachers. Consequently, the conclusions based on the relatively small sample of teachers do not present holistic view of the test construction competencies of the entire population of teachers considered for the study.

Definition of Terms

Certain variables and terms used in this study have been defined to put them in context. The operational definition of such variables and terms are as follows.

Teachers' multiple-choice test construction competencies

This refers to how well teachers are able to employ their abilities in applying the principles of constructing multiple-choice test items when constructing multiple-choice tests. Operationally, it is defined as teacher's composite score on the 'Teachers' Multiple-Choice Test Construction Competence Questionnaire (TTCCQ-MC)'.

Characteristics of the multiple-choice test(s)

1. In terms of quantitative item analysis (that is, with the use of difficulty and discrimination indices), it refers to the description in terms of the number or proportions of good and poor (or problem) items based on the criteria set for evaluating or judging the quality of the test items constructed by the teachers.
2. It is also the description of the multiple-choice tests in terms of the relationship between test length and the number of good items and problem items.
3. With regard to qualitative item analysis, it is the types of error identified with the tests constructed by the teachers using the 'Multiple-Choice Test Error Analysis Checklist'.

Quality of multiple-choice test items/Proportion of good items/Multiple-choice test items' quality

This refers to the number of good items per the assessment criteria divided by the total number of items that qualified for quantitative item analysis (valid items). This has also been referred to as *items' quality* in the context of this study.

Proportion of poor (problem) items

It refers to the number of poor or problem items per the assessment criteria divided by the total number of items that qualified for quantitative item analysis (valid items).

Valid Items

Multiple-choice items that qualified for item analysis have been termed 'valid items' in the context of this study.

Invalid Items

Objective items other than multiple-choice items, and multiple-choice items that were scored for each student as bonus are termed as invalid items. Quantitative item analysis was not performed on these items.

Test Length

This is the total number of items that qualified for item analysis for each multiple-choice test constructed by the classroom teachers.

Years of teaching

This is the number of years classroom teachers have spent in the classroom performing their duties and such duties include constructing test items to assess students learning outcomes. It also involves teachers' experiences with constructing multiple-choice test items.

Organisation of the Study

The study is organised into five chapters. Chapter One covers the background to the study, statement of the problem, purpose of the study, research objectives, questions and hypotheses, significance of the study, delimitations, and limitations to the study. Definition of terms and organisation of the study also complements Chapter One.

Chapter Two is devoted to conceptual, theoretical and empirical underpinnings that contributed towards investigating the issues of teachers' test construction competencies and the quality of the multiple-choice test items they construct. Information, based on which the need for this study emerged, were gathered from abstracts, books, journals, the internet, and works people have done.

Chapter Three discusses how the study was conducted. It harmonises the methodological components of the study, which is presented in seven sections namely research design, study area, population, sampling procedure, research instrument, ethical considerations, data collection procedure, and how the data collected was processed and analysed.

Chapter Four presents results and findings from the study. The research findings in relation to teachers' test construction competencies are also discussed under this chapter.

Chapter Five, which is the final chapter, comprises the summary, conclusions and recommendations. Moreover, suggestions for further research are also presented under this chapter.

CHAPTER TWO

LITERATURE REVIEW

Introduction

This chapter is devoted to conceptual, theoretical and empirical underpinnings that contributed towards investigating the issues of teachers' test construction competencies and the quality of the multiple-choice test items they construct. Information, based on which the need for this study emerged, were gathered from abstracts, books, journals, the internet, and works people have done. With respect to the purpose of the study, the following thematic areas were reviewed:

A. Theoretical Review
   i. Classical Test Theory
B. Conceptual Review
   i. Test as a Tool of Measurement
   ii. Standardised Tests and Non-standardised Tests
   iii. Formative Assessment and Summative Assessment
   iv. Quality of Assessment Procedures
   v. Test Construction Process
   vi. Principles, Guidelines or Suggestions for Constructing and Improving the Quality of Multiple-Choice Tests
   vii. Assessment Competence versus Assessment Practice
   viii. Assessment Competence, Test Construction Competence and Multiple-Choice Test Construction Competence

C. Empirical Review

i. Test Construction Competence, Test Quality, and Years of Teaching

Theoretical Review

Classical Test Theory (Model)

Classical test theory (CTT) as a theory of measurement describes the conceptual basis of reliability and defines ways for estimating the reliability of psychological measures (Gulliksen, 1950; Magnusson, 1967). It is one of the most pressing and important issues from Charles Spearman's (British psychologist) interest in the concept of correlation. Spearman, from 1904 to 1913, published logic and mathematical arguments that test scores are weak estimations of human traits, and as a result the observed correlation between fallible test scores is lower than the correlation between their true objective values (Crocker & Algina, 2008). In continual efforts to explain the terms fallible measures and true objective values, Spearman set the basis for the classical true-score model (Crocker & Algina, 2008).

Remarkably, authors such as Guilford (1954), Gulliksen (1950), Magnusson (1967), and Lord and Novick (1968), have recapitulated and explained that the essence of Spearman's model was that any observed test score could be envisioned as the composite of two hypothetical components –a true score and a random error component. Mathematically, this is expressed in the form $X = T + E$, where X represents the observed test score; T, the individual's true score; and E, a random error component (Crocker & Algina, 2008).

Consequently, CTT is a simple mathematical model that describes how measurement errors can influence observed score (Allen & Yen, 2002). It is also known as classical true-score theory. The theory states that for every observed score, there is a true score, or true underlying ability, that can be observed accurately if there were no measurement errors (Allen & Yen, 2002). Observed score refers to value that are obtained from the measurement of some characteristic of an individual. A true score is a theoretical idea that refers to the average score taken over repeated independent testing with the same test or alternative forms. It is also the real or actual level of performance on the psychological attribute being measure by a test (Furr & Bacharach, 2014). Apart from the influence of true scores, probable factors that affect observed scores are described by the theory as errors of measurement (Furr & Bacharach, 2014). True scores and error scores are unobservable theoretical constructs while observed score is observable in nature (Nitko, 2001).

Assumptions of classical true-score theory

Assumptions made with respect to classical true-score theory have been outlined by Allen and Yen (2002) as the following:

1. Observed score (X) on a psychological measure is equal to the sum of true score (T) and error score (E). That is, X = T + E.
2. Ɛ(X) = T. This states that the expected value of X, Ɛ(X), which is also known as the population mean, is equal to "T". This assumption is the definition of T: T is the mean of the theoretical distribution of observed scores that will be found in repeated independent testing of the same person with the same test for infinite number of times.

3. ρET = 0. The symbol "ρ" represents relationship. Thus, "ρET = 0" is the assumption that there is no relationship between error scores and true score. This means that test takers with high true score do not have systematically more negative or positive measurement errors than test takers with low true score. This assumption will be violated if for example, on the administration of college entrance exams, students with low true scores copied answers from those with high true scores. This situation will create a negative correlation between true score and an error score.
4. When there are two test forms, test 1 and test 2, the CTT assumes that the error scores from test 1 (E1) and the error scores from test 2 (E2) are uncorrelated (that is, ρE1E2 = 0). That is, if a person has a negative error score in Test 1, he or she is not more likely to have a negative or positive error score in Test 2. This assumption is not reasonable if the observed scores are greatly affected by factors such as examinee's mood, fatigue, effects of the environment, or practice effect.
5. ρE1T2 = 0; this assumption states that, the error scores on one test (E1) are uncorrelated with the true scores on another test (T2). 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^I that satisfies assumption 1 through to assumption 5, and if, for every group of test takers T = T^I and variance of σ = σ^I then the test are called parallel test. For σ equal

to σ', the condition leading to an 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 to assumption 5, and if, for every group of test takers , $T_1 = T_2 + C_{12}$, where C is a constant, then the tests are labelled τ -equivalent tests.

Regarding the assumptions, the situation where there are no errors of measurement in observed scores, one can greatly and confidentially depend on the observed scores for relevant decisions. This is because, from repeated independent testing, all the observed scores reflect the true ability of the candidate who is assessed. Also, suppose that a core mathematics achievement test is administered to a group of students who differ in ability, and the obtained scores are without measurement errors, the teacher would place his or her confidence in the assessment results since the differences (variability) in the students' test scores accurately reflect the differences in their true levels of knowledge in mathematics.

Nevertheless, the existence of errors of measurement results in deviations of the observed scores from the true scores (Bhattacherjee, 2012), and this minimises one's confidence and dependability on the assessment results. How much confidence one can place in test results is a question of two main concepts on quality of assessment procedures: (a) reliability and (b) validity. Therefore, classical true-score theory provides understanding of factors (measurement errors) that influence observed scores reliability and validity. Examples of such factors are mistakes in scoring test items, fatigue, and guessing (Amedahe & Asamoah-Gyimah, 2016; Crocker & Algina, 2008).

To effectively control and reduce the impact of measurement errors (or improve the quality of assessment procedures), the few assumptions of the theory expands into procedures and principles for test construction and evaluation (Allen & Yen, 2002).

The relevance of the classical true-score theory to the study

The relevance of the classical true-score theory to the present study are as follows.

1. The theory helps to understand certain factors that influence the quality (or validity) of assessment results or students' observed scores apart from the influence of true scores (Furr & Bacharach, 2014). These factors can be categorised into four namely factors associated with the test, the test taker, the testing environment, and the scoring. For the purpose of the study, test-related factors were examined. To examine test-related factors that affect the quality of assessment results, the theory assumes all the other factors that reduces reliability are adequately controlled for (Crocker & Algina, 2008). Thus, test-related factors (characteristics) were examined with the assumption that all other factors that affect students' observed scores were adequately controlled for.
2. Errors associated with tests negatively affect the reliability and validity of the entire assessment results (Nitko, 2001). To help improve the quality of tests, some principles, guidelines, or suggestions have been given by researchers, professionals, and experts in educational assessment of students and psychological testing. Hence, the theory draws attention to certain guidelines that teachers should be able to

follow in order to improve upon the quality of their assessment procedures (Allen & Yen, 2002). In this study, the guidelines examined included the test construction process and principles for test construction. The principles informed the development of the research instrument (the questionnaire) for obtaining data on the multiple-choice test construction competencies of the classroom teachers. It also informed the construction of the checklist for item analysis.

3. The theory emphasises that the quality of test items depends on the test's ability to bring out individual differences on a construct of interest the teacher wishes to measure (Furr & Bacharach, 2014). According to Joshua (2005), good tests are constructed or developed. Therefore, to achieve test of good quality, it involves writing appropriate test items, appropriate test instructions and adequate competence in putting the items together in different ways to achieve anticipated format and purpose. This means that the test's ability to bring out such differences depends on the test construction competence of the classroom teacher. Therefore, the theory supports the need to investigate the relationship between test construction competencies and quality of test items prepared by the classroom teachers.
4. The theory also offers methods for analysing the quality of tests prepared by classroom teachers based on students observed scores (Crocker & Algina, 2008). These methods are quantitative item analysis and qualitative item analysis (Kubiszyn & Borich, 2013). Accordingly, quantitative item analysis were used in identifying the number of good items and problem items each test have using items'

difficulty and discrimination indices. Qualitative item analysis was used to report on the frequency of specific type of errors observed with the problem items using the checklist.

5. Based on the theory, to improve the reliability and validity of assessment results, there is the need to lengthen the assessment procedures (Allen & Yen, 2002; Crocker & Algina, 2008; Nitko, 2001). However, Crocker and Algina (2008) have stated that this assumption works well when all the items are representative of the domain sampled and have the same or similar level of appropriate difficulty and discrimination indices. This implies that the greater the number of test items appropriate in content, difficulty and discrimination indices, the degree of errors in students observed scores reduces; hence, resulting in appreciable level of reliability coefficient.

Reliability coefficient is the squared correlation between observed scores and true scores (Furr & Bacharach, 2014). This suggests that reliability coefficient of the test will reduce when there are more problem items contributing to errors in observed scores. On the other hand, the presence of more good items reduces errors; therefore, resulting in improved reliability coefficient (Crocker & Algina, 2008). Accordingly, Crocker and Algina (2008) have stated that items must be well crafted and free of technical flaws that may cause examinees to respond on some basis unrelated to the content.

From the aforesaid, to achieve an appreciable level of reliability coefficient, the theory endorses that it is good for teachers to construct a relatively adequate number of test items without problem items or which has relatively few problem items. Consequently, the theory brought out the need to investigate among classroom teachers what happens to: (a) the number of good items in a test when test length is either increasing or decreasing; (b) the number of problem items as test length either increases or decreases.

In conclusion, classical true-score theory has been described as a weak theory since it is easy to meet its set of assumptions (Allen & Yen, 2002). Yet, the application of this theory to investigate the quality of the test items was of peculiar interest because it helped to understand how teachers' competencies in applying the principles of test construction are related to test quality. It endorsed the use of quantitative methods of evaluating the quality of the test items based on test scores. Moreover, it offered qualitative item analysis for detecting the frequency of types of error identified with test items constructed by the classroom teachers. How methods of evaluating the quality of tests (quantitative and qualitative item analysis) were employed is discussed in the section of this chapter titled conceptual review.

Conceptual Review

Test as a Tool of Measurement

A test is a tool or systematic procedure for observing and describing one or more attributes of a student using either numerical scale or a classification scheme (Nitko, 2001). A test can be simply referred to as a measuring device or procedure (Cohen & Swerdlik, 2010). Just as a "ruler" is a helpful measuring instrument in the hands of the tailor, and it enhances the job performance of that tailor, test is similarly a useful measuring instrument in the hands of the teacher or school administrator, and it enables each of these professionals to do their work effectively (Joshua, 2005). For instance, test helps the teacher to assess learner's needs, report pupils' progress to parents, and monitor learning progress, among others. Its systematic procedure follows four major stages of testing namely: construction, administration, scoring and interpretation (Joshua, 2005). As a tool for measurement, care should be taken in its construction and development. In general terms, there are two main types of test: standardised tests and non-standardised tests (Kubiszyn & Borich, 2013).

Standardised and Non-standardised Tests

According to Harris (2002), precisely, a standardised test is any measure that is useful in evaluating characteristics or skills of students, with specific procedures for administering and scoring the tests and interpreting the assessment results; they are usually developed by test construction professionals or experts. In the Ghanaian educational context, with summative assessment, examples of standardised tests at the basic level and secondary level are Basic Education Certificate Examination (BECE) and West African Senior School Certificate Examination (WASSCE) respectively.

On the other hand, non-standardised measures which are often referred to as teacher-made tests are constructed, administered, and scored by classroom teachers, and often consists of completion, true-false, matching, multiple-choice, and essay items (Kubiszyn & Borich, 2013). These teacher-made tests are often flexible, or variable, in terms of their administration and scoring procedures, and in the amount of attention given to their construction. Different teachers may be more or less careful in constructing their tests, may allow more or less time for the test to be taken, and may be more or less stringent in grading the test (Kubiszyn & Borich, 2013).

Amedahe (2014) has indicated that at the basic and secondary levels, Ghana’s educational system lacks standardised measures for assessing and monitoring pupils’ or students’ performance at the various grade levels. He further indicated that the only standardised assessment instrument at the basic level is the National Education Assessment (NEA) for Primary 3 and 6 in mathematics and English language administered to a sample of between 3-5% of the population every two years. Therefore, the Ghanaian educational system depends largely on teacher-made tests when it comes to classroom assessment (Amedahe, 2014; Asamoah-Gyimah, 2002). Indeed, the demands of classroom assessment go well beyond readily available instruments to embrace teacher-made tests (Sanders & Vogel, 1993). McMillan (2013) defined classroom assessment as:

> A broad and evolving conceptualization of a process that teachers and students use in collecting, evaluating and using evidence of student learning for a variety of purposes, including diagnosing student strengths and weaknesses, monitoring student progress towards meeting desired levels of proficiency, assigning grades, and providing feedback to parents. (p.4).

Classroom assessment which relies greatly on teacher-made tests plays an increasingly crucial role in the field of education. That is, teacher-made tests aid in pre-assessment (which is the assessment of what student already know prior to teaching); formative assessment (the assessment of student performance incorporated into the act of teaching); and summative assessment (the assessment of student learning at the end of some instructional period) of pupils' or students' learning outcomes (Gareis & Grant, 2015). Concerning the purpose of this research, there was the need to look at formative and summative assessment in the classroom.

Formative and Summative Assessment in the Classroom

In the field of education, formative assessment is mostly employed to monitor the leaming progress of students during instructional sessions. It also helps to provide continuous feedback to students, identify areas that improvement is essential, and reinforce learning (Linn & Gronlund, as cited in Elharrar, 2006). The areas of improvement include knowledge, skills, and attitudes (Choi, Nam & Lee, 2001). Assessments which are formative in nature are generally performed on a continuous basis with the purpose of arriving at what should be done in order to improve students' achievement in the near future (Gattullo, 2000). For instance, relevant educational information obtained from formative assessment of students can inform educators to establish educational programmes that will help improve overall teaching and maximisation of students' achievement.

According to Elharrar (2006) summative assessment is usually aimed at certifying a student's mastery of objectives and to gauge the level of acquisition of a specific leaming objective or curriculum goal. It is largely used to determine at one point in time, or after a set number of performances, how much a student knows and can do (Callahan, 2006). The main purpose of this form of assessment is to obtain information on what students know and understand and is usually used to assign grades (Kubiszyn & Borich, 2013). This type of assessment is often made up of traditional testing techniques. This is due to the fact that it usually consists of paper-and-pencil assessment techniques in which student information is gathered through administered end of term multiple-choice test (Gattullo, 2000).

Quality of Assessment Procedures

Quality of assessment procedures is of great concern when it comes to the assessment of student learning. Ghanaian classroom trained and untrained teachers, from the basic level to the university level, construct, administer and score classroom achievement tests regardless of whether they have had training in measurement and evaluation or not (Anhwere, 2009). When classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction, the quality of the tests they construct is questionable. Why? According to Chau (as cited in Hamafyelto et al., 2015), teacher’s test construction competence is directly related to ensuring the quality of a test. Poor test quality negatively affects the validity of assessment results (Amedahe & Asamoah-Gyimah, 2016). From the aforesaid, by implication, when teacher-made tests are low in quality, school administrators and teachers will not be able to make available support and educational opportunities that each student needs (Agu et al., 2013). In other words, lack of or low degree of validity of test results leads to undependable inferences about student learning (Amedahe & Asamoah-Gyimah, 2016; Gareis & Grant, 2015) based on which educational decisions such as promotion and selection of students for educational opportunities would be wrongfully made.

To avoid or minimise the negative effects of assessment procedures which are low in quality, the onus rests on classroom teachers to ensure the quality of the assessment procedures they employ. In terms of quality of assessment procedures, the important question is asked: How reliable and valid are the assessment results?

Validity of assessment results

Validity is the most fundamental and significant quality in the development, interpretation and use of educational assessment procedures (Furr & Bacharach, 2014). It is an abstract concept (Furr & Bacharach, 2014) which refers to the appropriateness or soundness of the interpretations and use of students' results obtained on an assessment procedure (Nitko, 2001). Validity is an abstract concept because it cannot be directly observed. Instead of directly observing validity, people depend on evidences that are indicative of its presence. To validate the interpretations and uses of scores obtained by students on a particular test, classroom teachers must provide evidence that interpretations and uses of the results are appropriate. According to Furr and Bacharach (2014), there are mainly three types of interrelated validity evidences presented in the 1985 standards for educational and psychological testing. They are: construct-related validity evidence; content-related validity evidence; and criterion-related validity evidence.

Construct-related validity evidence

Construct-related validity evidence looks at whether an individual's performance to be measured could be treated as legitimate indicator of the psychological construct or capability the classroom teacher hopes to assess (Furr & Bacharach, 2014). In the classroom situation, the constructs of greatest significance to teachers are those of learned knowledge, which cannot be observed directly: A construct of what a student possess or have achieved that no one can see. For instance, student's achievement in core mathematics and financial accounting at the end of a term cannot be directly observed (Furr & Bacharach, 2014). Therefore, to establish construct-related evidence of validity, teachers must establish that the visible student behaviours they choose to observe are appropriate indicators of the students' knowledge they wish to assess (Nitko, 2001; Furr & Bacharach, 2014).

Content-related validity evidence

This establishes how well the actual content of questions, tasks, observations, or other elements of a test corresponds to the student performance that is to be observed (Allen & Yen, 2002; Oosterhof, 2003). Content-related evidence of validity is often established while an assessment is being planned. It involves systematic analysis of what the test is intended to measure. Poor planning or lack of planning by the teacher may result in a test that does not incorporate targeted behaviours. There are two main types of content validity namely face validity and logical validity (Allen & Yen, 2002).

Face validity

Face validity is achieved when an individual examines a test and concludes that the test measures the relevant trait of interest (Allen & Yen, 2002). The person making this examination can be anyone from an expert to an examinee. If people disagree, face validity is in question. Face validity may be sufficient to justify the use of some tests (Allen & Yen, 2002; Morrow, Mood, Disch, & Kang, 2016). The classroom exam, when carefully prepared, has some face validity. For example, an arithmetic test, on the "face" of it, measures arithmetic performance. Face validity may be essential for some tests because of their intended use (Allen & Yen, 2002).

Logical validity

Logical or sampling validity is a more sophisticated version of face validity (Allen & Yen, 2002). It includes the careful description of the domain of behaviours to be measured by a test and the logical design of items to cover all the important areas of this domain. Logical validity is mainly useful in the construction and designing of achievement tests (Allen & Yen, 2002). Because content validity is based on subjective judgments, the determination of this type of validity is more subjective to error than are other types of validity (Allen & Yen, 2002). Nevertheless, in general terms, establishing content validity is the first concern in the construction and designing of tests, and items are written to satisfy content-validity requirements (Allen & Yen, 2002; Morrow et al., 2016). Through statistical item analysis technique, the test can be revised and improved to guarantee that other aspects of good measurements are achieved. Usually, the mere fact that test has content validity is not a sufficient justification for its use. Before it is used, the test should have proven effectiveness, such as criterion-related validity (Allen & Yen, 2002).

Criterion-related validity evidence

This indicates how well a student's performance on test correlates with her or his performance on relevant criterion measures external to the test (Crocker & Algina, 2008). To establish this validity evidence, a test is administered to a group of individuals and their test scores are compared to a criterion measure, or to a standard, that reflects the particular variable of interest. The criterion can be those reflecting academic achievement (for example, grape point average, GPA), previously developed tests and instructor's ratings (Domino & Domino, 2006). There are two types of criterion-related validity and they are predictive validity and concurrent validity (Allen & Yen, 2002; Crocker & Algina, 2008).

Predictive validity

Predictive validity involves using tests scores to guess or forecast about future behaviour (Allen & Yen, 2002). For example, let us assume we have a standardised test (such as WASSCE) that we wish to validate by examining how well it will predict GPA at the college level. In an ideal world, the test would be administered to unselected sample of students, let them all enter college education, wait for 3 years, obtain each of the student's cumulative GPA, and correlate the test scores with the GPA. This process of validation is called predictive validity (Domino & Domino, 2006).

Concurrent validity

Concurrent validity refers to the degree to which test scores and criterion measurements made at the time the test was given are related (Crocker & Algina, 2008). Sometimes, it is apparently difficult finding an unselected sample, convincing school officials to admit all of them, and waiting 3 years in the name of obtaining predictive validity evidence (Domino & Domino, 2006). Nevertheless, when due to time factor, predictive validity evidence is difficult to achieve, it might make sense to collect both the test scores and the criterion data at the same time, and the scores correlated for concurrent validity (Domino & Domino, 2006). For example, all form 3 students of a mechanics' institute can be administered a mechanical aptitude test and have instructors to individually rate each student on their mechanical aptitude and both scores correlated. This is called concurrent validity; for both the test scores and the criterion scores are collected within the same period of time (Domino & Domino, 2006).

Reliability of assessment results

Reliability, which is a group characteristic, refers to the consistency of assessment scores over time; it is indicative of the precision with which a trait is measured (Amedahe & Asamoah-Gyimah, 2016). Most often, behavioural scientists treat reliability as if it is an all-or-none issue. For example, someone might ask whether a particular achievement test is reliable, and there is the likelihood for a teacher to answer 'Yes' or 'No' (Furr & Bacharach, 2014). Such a response would portray reliability as something being present or absent. It is more appropriate to describe the reliability of a test as being very low, low, moderate, high or very high rather than thinking of it as something which is there or not (Furr & Bacharach, 2014; Nitko, 2001).

Reliability which is judged to be on a continuum of less reliable to more reliable is a necessary but not sufficient condition for validity evidence (Nitko, 2001). For instance, suppose a teacher administered achievement test to a student on two different occasions, and he or she scored 2 out of 10 on the first administration and 2 out of 10 on the second administration. It can be said that the performance of the student is stable over the two administrations (that is, high in reliability). If the test items were based on what the student was taught in class, are free of format and constructional flaws, the consistency of the results becomes necessary condition towards the teacher's decision of describing the student as a low achieving student.

On the other hand, the moment the teacher is questioned about whether the environment in which the student took the test was conducive on both occasions, whether the test items reflected what was taught, this implies that the reliability (stability) in the student's assessment results is not sufficient for the teacher to classify him or her as below average student. Therefore, the teacher should be willing to provide these other validity evidences before the consistency in the student's assessment results could be accepted as sufficient for interpreting the student's achievement.

According to Furr and Bacharach (2014), just as a psychological attribute such as test anxiety is an unobserved feature of an individual, reliability is an unobserved characteristic of test scores –thus, a theoretical notion. Additionally, just as one estimates an individual's level of test anxiety, likewise test's reliability. Given certain assumptions of CTT, it is possible to calculate numerical values (indices) that estimate the degree of test's reliability (Furr & Bacharach, 2014).

The general way to obtain reliability indices is for a teacher to administer a test to a group of students one or more times and obtain the scores (Furr & Bacharach, 2014). The teacher, therefore, can correlate the scores from the two administrations to obtain the reliability coefficient for the test (Nitko, 2001). The reliability coefficient helps to know whether the relative standing of the students in the group changes from one administration to the next. There are several types of reliability coefficients discussed by Nitko. They include test-retest reliability, Spearman-Brown (split-halves) reliability, Kuder-Richardson and coefficient alpha reliability (Nitko, 2001) among others.

When a teacher is interested in the stability of scores over a period on a fixed sample of assessment tasks, the test-retest reliability coefficient can be estimated (Nitko, 2001). However, when the teacher is interested to estimate the equivalence of a test using information about the internal consistency of students' responses, Spearman-Brown (split-halves) reliability, Kuder-Richardson or coefficient alpha reliability can be used (Nitko, 2001). Spearman-Brown (split-halves) reliability can be used by the classroom teacher, when he or she wants to estimate the equivalence of a test by dividing test into two equivalent halves (namely A and B) and correlating the scores on A with B (Nitko, 2001). On the other hand, the teacher can also employ Kuder-Richardson reliability to estimate the equivalence of an assessment instrument when the test items are dichotomously scored as 0 or 1 (Crocker & Algina, 2008).

Nitko (2001) has indicated that there are two main procedures for estimating Kuder-Richardson reliability, which include Kuder-Richardson formula 20 (KR20) and Kuder-Richardson formula 21 (KR21). KR20 uses data on the proportion of students answering each item correctly and the standard deviation of the total scores while KR21 is computationally simpler version of KR20 that uses only the mean and standard deviation of the total scores (Crocker & Algina, 2008). Where test items are not dichotomously scored (for instance, a scale of 1 to 4), a more general version of KR20 known as coefficient alpha is appropriate to be used (Nitko, 2001).

Factors that simultaneously reduces validity and reliability

Some factors simultaneously affect the reliability and validity of assessment results. These factors can be categorised into four namely factors associated with the test, the test taker, the testing environment, and the scoring.

Test-related factors

According to Amedahe and Asamoah-Gyimah (2016), “a test is usually a composite of single items” (p. 79). It, therefore, takes up the features of the individual items it contains, and that any weakness in the individual items from which the total score is obtained would be reflected in the total scores as errors. The circumstance where the errors are introduced into the total scores reduces its reliability and validity. In fact, any deviation made by classroom teachers from general test construction principles and specific principles for the construction of multiple-choice tests becomes test-related errors that reduce the quality of teacher-made multiple-choice test. The following are some examples of test-related factors that reduce reliability and validity of assessment results:

Test layout and item format errors

Weaknesses in test layout and format can take the form of poorly spaced items, the use of font size which students find difficult to see and read (Kubiszyn & Borich, 2013). These factors would generally create discomfort and pose difficulty for the student with regard to what exactly is either being measured or what to do. When this happens, the nature of the weakness in test layout and format tends to lower the extent to which student’s performance can be relied upon (Kubiszyn & Borich, 2013).

Improper arrangement of test items according to difficulty level can also affect the validity of the results (Amedahe & Asamoah-Gyimah, 2016). Test items are characteristically organised in order of difficulty, with the easiest items first. This is essential for motivational purposes among other things (Brown, 2004). When difficult items are positioned early in the test, they may affect students in such a way that they spend too much time on them and this would, in turn, prevent them from reaching items they could have easily answered. Also, such arrangement may frustrate the students and consequently affect the reliability and validity of their assessment results (Amedahe & Asamoah-Gyimah, 2016).

Another test layout and format-related factor that affects quality of assessment procedures is an identifiable pattern of answers. This factor applies to objective type test items (Amedahe & Asamoah-Gyimah, 2016; Nitko, 2001). When the best or correct answers in a test are placed in some systematic pattern (for example, A, C, B, A, C, B), it will enable the students to guess the answer to some items after completing some sessions of the test. Guessing correct answers base on pattern that emerges will not help portray the students' actual ability on the achievement test (Nitko, 2001). Therefore, any interpretation and usage of obtained scores may have low reliability and validity.

Item constructional errors

The ambiguity of test items, wrong use of punctuations, wrong spelling, poor wording are all examples of grammatical errors that can reduce the quality of the test when constructing test items (Kubiszyn & Borich, 2013; Morrow, Jackson, Disch, & Mood, 2000; Nitko, 2001). For example, the ambiguity of test items may lead to differences in how an item is interpreted and may give rise to guessing which reduces reliability and validity (Amedahe & Asamoah-Gyimah, 2016). Wrong usage of punctuation, wrong spelling, poor wording also have the potency to generally create discomfort and pose difficulty for the student with regard to what exactly is either being measured or what to do.

Unclear directions are also one of the test-related errors. At all times, there is the need for classroom teachers to provide clear directions as to how testees are expected to answer test items. This will help them to respond meaningfully to a set of test items (Amedahe & Asamoah-Gyimah, 2016; Joshua, 2005). Nevertheless, if test directions are not clearly stated as to how learners are to respond to the test items or record their responses, it will tend to decrease the reliability and validity of the observed scores. This is because testees may get confused over how to respond and in what manner to record their answers, and such confusion may affect their performance by not answering or recording their responses as expected (Nitko, 2001). Thus, teachers must provide clear directions to their students any time they construct test items.

Clues to answers are another test-related error which reduces reliability and validity of assessment results. In an ideal world, an examinee will respond to a multiple-choice question incorrectly only if he or she does not know the answer and correctly if he or she knows. However, the presence of clues in multiple-choice items adversely influences reliability and validity of assessment results (Morrow et al., 2000).

All testees are not equally good at recognising clues; therefore, the effects of clues leading to the correct answer are not as predictable as those emanating from chance (Morrow et al., 2000). The only way to deal with the problem is to remove the clues from the items. Some clues are relatively obvious; others are delicate (Morrow et al., 2000). For example, it is usually easy to identify the use of a keyword in both the stem and the correct response or a keyed response that is the only one that grammatically agrees with the stem (for example, stem calls for a plural answer and all but one of the alternatives are singular). Clang associations (that is, words that sound as if they belong together –for instance, up and down, shoes and socks) are frequently relatively difficult for the test constructor to recognise but provide immediate clues to the test takers (Morrow et al., 2000).

Test difficulty has likewise been identified as test-related factor that affect the quality of assessment results. The difficulty of a selection type test item is defined in terms of the percentage of students that has answered the particular item correctly (Amedahe & Asamoah-Gyimah, 2016; Kubiszyn & Borich, 2013). For the selection type of items, when a test is difficult, students may be induced to cheat, and guess the answers (Amedahe & Asamoah-Gyimah, 2016). This results in the introduction of measurement errors into the observed scores (Crocker & Algina, 2008).

Another test-related error is inadequate time limits assigned to the test. Content delivered to students in class is time bound. For instance, contents which are more technical and difficult require more time to deliver (Morrow et al., 2000; Morrow et al., 2016). Similarly, since test items are dependent on given content, it requires an adequate time limit for students to also demonstrate their knowledge. Test takers need to be given adequate time within which they will complete a given test taking into consideration the content demands of the various test items. This is a significant factor in a power test (that is, a test which measures what a student knows or his or her ability to do something rather than measure the speed of the student in completing a task) (Amedahe & Asamoah-Gyimah, 2016).

Though speed test applies in certain contexts, most of achievement tests carried out in the classroom are power assessment and therefore should reduce the effects of speed on student performance (Morrow et al., 2000). If the effect of speed is not controlled, most test takers may not finish the test before the allotted time will expire. When this happens, it could be interpreted as test takers were not offered the opportunity to demonstrate their ability in terms of their knowledge on the subject matter content (Morrow et al., 2000). On the contrary, some test takers may complete test items in a haste (without meaningful thought to the test items) because of lack of adequate time. According to Morrow et al. (2000), this may lead to poor performance on the test. In either case, the reliability and validity of observed scores will tend to be lowered.

Test length

It has been observed in the literature that the greater the number of test items that enter into the formulation of a test score, the more reliable and valid that score will be (Crocker & Algina, 2008; Nitko, 2001). This implies that test length is one of the test-related factors that affect reliability and validity. When classroom teachers construct few multiple-choice items for the assessment of student learning based on a given content area, the degree to which one can put confidence in a student's performance is little. Thus, to improve the reliability and validity of assessment results, there is the need to lengthen the assessment procedures (Nitko, 2001).

Teacher-related factors that lead to errors in the test

Teacher factors that introduce errors in the test include lack of adequate knowledge about students' characteristics, and poor test construction skills.

Lack of adequate knowledge about students' characteristics

When constructing test items, it is advisable to write test items in a language that is at the level of their students and therefore the sentence structure should not be too complex for the student's level (Amedahe & Asamoah-gyimah, 2016; Nitko, 2001). However, teachers who do not have adequate knowledge about their students may use language which is above their students' level of understanding. When the vocabulary and sentence structure are complicated and very difficult for the students taking the test, it will result in the assessment procedure measuring student's comprehension ability rather than the student's achievement in a subject matter (Nitko, 2001). In this case, the interpretation and use of the assessment results may have low reliability and validity.

Inadequate or poor test construction competence

Though a teacher might possess adequate knowledge of his or her students, however it is observed that some classroom teachers do not possess adequate skills to assess student learning outcomes (Amedahe, 1989, Anhwere, 2009; Rivera; 2007; Sasu, 2017). That is, some teachers have difficulty in applying or following test construction principles such as avoiding the use of items that are ambiguous and clues that have the potential to lead students to right answers. Consequently, this minimises the extent to which one can depend on assessment results produced from the test they construct.

Student-related factors

Factors such as low level of motivation, emotional disturbance, and over anxiety are probable test taker factors that can simultaneously influence their observed scores on a particular test (Amedahe & Asamoah-Gyimah, 2016; Cohen & Swerdlik, 2010). These factors are inherent in students and tend to interfere with their performance during a test. Thus, the factors tend to restrict and modify students' responses in the assessment situation which in turn distort the results (Amedahe & Asamoah-Gyimah, 2016; Nitko, 2001). Once there is distortion of the results, its interpretation and use will not be dependable.

Another student factor that tends to simultaneously reduce reliability and validity is guessing. Ideally, test takers will respond to multiple-choice items correctly only if they have knowledge of the right answer and incorrectly if they do not have such knowledge. Nevertheless, a test taker may blindly guess the right answer to a question, and there is no way the examiner can determine from the response given whether the response is reflective of knowledge acquired or the student's luck in guessing (Morrow et al., 2000). Therefore, when this happens, the extent to which the assessment results can be relied on and trusted is lowered.

Testing environment or conditions-related factors

Factors related to the testing environment or testing conditions that influence reliability and validity of assessment results include improper invigilation, poor sitting arrangements, poor lightening system and disruptive noise during testing (Joshua, 2005; Kubiszyn & Borich, 2013; Nitko, 2001). This is because the factors tend to affect individual students performance differently and most often negatively (Amedahe & Asamoah-Gyimah, 2016). For instance, where the outlined environmental factors are present, it can lead to different levels of frustration among the test takers. When students pay attention or spend time to ease their frustration instead of fully utilising the allotted time for the completion of the test, it might hinder their ability of putting up their best.

Scoring-related factors

Reliability and validity of assessment results is also lowered when teachers are inconsistent in scoring the responses of their students (Kubiszyn & Borich, 2013). This might happen as a result of favouring some students over other students by being generous to some and very hard on others, overlooking marking scheme, and not counting well the number of items students had correct on a given test (Amedahe & Asamoah-Gyimah, 2016).

Test Construction Process

In constructing a particular test for the assessment of students' learning, the classroom teacher must go through certain process defined in the literature by assessment or measurement and evaluation experts for test construction. The process helps to ensure that test items constructed are of good quality to produce more reliable and valid assessment results based on which educational decisions would be made. Amedahe and Asamoah-Gyimah (2016), Joshua (2005), Cohen and Swerdlik (2010), Crocker and Algina (2008), and Izard (2005) have discussed stages for test construction. Based on general observation, six stages for construction of teacher-made tests can be outlined as follows.

A. Test conceptualisation and planning the test
B. Writing and initial review of the items
C. Assembling the test
D. Qualitative evaluation of the test
E. Tryout, administration and quantitative evaluation of the test
F. Revision of the test

Test conceptualisation and planning the test

At this stage, careful considerations by the teacher should cover the purpose of the test, content area students are to be assessed, characteristics of students to be tested, test construction, administration, scoring, and interpretation. Here, careful thought should be given to measures that will help him or her to improve on the validity and reliability of the assessment results.

One of the ways of improving the quality of the test is ensuring content validity. This can be achieved through the development of test specification table (Joshua, 2005; Nitko, 2001); therefore, this stage permits the teacher to develop the test specification table (or test blueprint) and also decide on the item format that will be appropriate for assessing students' learning outcomes. The test blueprint is a two-dimensional table that matches course content to levels of instructional objectives (Amedahe & Asamoah-Gyimah, 2016; Joshua, 2005). Also, careful consideration is giving to how the test items will be scored, analysed, interpreted and reported to ensure an appreciable degree of reliability and validity.

Writing and initial review of the items

After the test conceptualisation and development of the test plan (test specification table), the next stage is to write the test items. Therefore, classroom teachers should be competent in writing the items. Even where they lack such competencies, Rivera (2007) believes that classroom teachers can master the writing of test items through practice.

In writing test items, teachers informed by their test blueprint should choose an item format or combination of formats for assessment of instructional objectives. Item format refers to variables such as form, plan, structure, arrangement, and layout of individual test items (Cohen & Swerdlik, 2010). Broadly, there are two main item formats namely objective type item format and essay type item format (Nitko, 2001; Joshua, 2005). The objective type item format comprises true-false items, matching items, multiple-choice items, and short-answer items. Essay formats consist of extended and restricted response items (Amedahe & Asamoah-Gyimah, 2016). In Ghana, apparently, at the SHS level, end-of-semester test items should be made up of multiple-choice items and essay items.

As this research is devoted to the variables teachers' multiple-choice test construction competencies and the quality of the multiple-choice test items they develop for the summative assessment of students' learning outcomes, there was the need to review literature on multiple-choice item format based on the following thematic areas: What is a multiple-choice item?; Advantages and Disadvantages of using multiple-choice item format.

What is a multiple-choice item?

Multiple-choice item is an item which is made up of one or more introductory sentences followed by a list of two or more suggested responses (Nitko, 2001). The student is required to choose the correct answer from among the responses the teacher gives (Nitko, 2001). The part of the item that asks the question is called the stem. Instead of asking a question, it may set the task a student must perform or state the problem a student must solve. The list of suggested responses to the stem is called options. The options are also known as alternatives, responses or choices (Morrow et al., 2000; Nitko, 2001). Usually, only one of the options is the correct or best answer to the question or problem the teacher pose. This is called the keyed answer, keyed alternative, or simply the key. The remaining incorrect options are called distractors or foils (Joshua, 2005; Nitko, 2001).

Advantages and disadvantages of using multiple-choice item format

Advantages and disadvantages associated with the use of multiple-choice test items have been discussed comprehensively in various textbooks, articles and journals on educational measurement and evaluation, psychological testing, among others. The following advantages and disadvantages of using multiple-choice items in assessment of students' achievement were discussed by Amedahe and Asamoah-Gyimah (2016), Nitko (2001), Kubiszyn and Borich (2013), Oosterhof (2003), and Cohen and Swerdlik (2010):

Advantages of using multiple-choice items

a. Multiple-choice questions have considerable versatility in measuring objectives from knowledge to the evaluation level.

b. Since writing is minimised, a substantial amount of course material can be sampled in a relatively short time.

c. Scoring is highly objective, requiring only a count of the number of correct responses.

d. Multiple-choice items can be crafted so that students must discriminate among options that vary in degree of correctness. This allows students to select the best alternative and avoids the absolute judgements found in true-false tests.

e. Since there are multiple options, effects of guessing are minimised.

f. Multiple-choice items are amenable to item analysis, which permits a determination of which items are ambiguous or too difficult.

Disadvantages of using multiple-choice items

a. Multiple-choice questions can be time-consuming to write.

b. If not carefully written, multiple-choice questions can sometimes have more than one defensible correct answer.

c. The items are somewhat susceptible to guessing.

d. Multiple-choice items often must indirectly measure targeted behaviours.

To ensure that the assessment task neither prevent nor inhibit a student’s ability to demonstrate attainment of the learning target, the care should be taken to follow the guidelines for constructing multiple-choice test items. For instance, to avoid ambiguous and imprecise items, inappropriate and unfamiliar vocabulary, and poorly worded directions, after the first draft of the items, the items should be reviewed and edited. This process will help to engage in the first phase of qualitatively analysing the drafted initial pool of items before assembling the test. Moreover, the marking scheme should be prepared in conjunction with drafting the items (Etsey, as cited in Amedahe & Asamoah-Gyimah, 2016).

Assembling the test

Under this stage, the final draft made is packaged and produced in a neat and legible form. During the packaging, the items should be adequately spaced for easy reading (Kubiszyn & Borich, 2013). When items are crowded together, a student may inadvertently perceive a word, phrase, or line from a preceding or following item as part of the item in question. Naturally, this interferes with a student's capacity to demonstrate his or her true ability (Kubiszyn & Borich, 2013). Here, instructions should be clearly indicated on the test. For instance, how the students are required to answer the items should be indicated.

During the reproduction phase of assembling the test, the copies should be inspected for legibility, and omission of some pages stapling the multipage test (Kubiszyn & Borich, 2013). Actually, the test should be reproduced in a form that no examinee would be disadvantaged in any way as a result of wrong spellings, omitted part of some questions, poor printing and photocopy and other similar factors. In the course, measures should also be employed to ensure security of the test (Kubiszyn & Borich, 2013; Nitko, 2001).

Qualitative evaluation of the test

This method is used to review the items for test construction errors that might have creeped into the printed copy (Kubiszyn & Borich, 2013). Again, it is required for assessing the worth of the test before it is produced in large numbers to be administered (Amedahe & Asamoah-Gyimah, 2016). Hence, qualitative evaluation (or item analysis) is a non-numerical method for analysing test items not employing student responses, but considering content validity, clarity, practicality, efficiency and fairness (Amedahe & Asamoah-Gyimah, 2016). Therefore, after crafting test items and initial qualitative item analysis has been done, and a copy of the test printed out, another qualitative evaluation of the test is required.

Content validity, as one of the qualitative evaluation criteria, answers the questions: Are the items representative sample of the instructional objectives covered in class? Does the test genuinely reflect the level of difficulty of materials covered in class? If the answer is 'Yes', then content-related validity evidence is established (Amedahe & Asamoah-Gyimah, 2016). Clarity as another measure of evaluating the worth of the test refers to how the items are constructed and phrased while simultaneously judging them against the ability levels of the students. That is, the test material should be clear to students as to what is being measured and what they are required to do in attending to the questions (Nitko, 2001).

Practicality is concerned with the adequacy of the necessary materials and appropriateness of time allocated for the completion of the test (Brown, 2004). Efficiency of a test seeks information as to whether the way the test is presented is the best to assess the desired knowledge, skill, or attitude of examines in relation to instructional objectives (Amedahe & Asamoah-Gyimah, 2016). Conversely, fairness refers to the freedom of a test from any kind of bias. The test should be judged as appropriate for all qualified examinees irrespective of race, religion, gender, or age. The test should not disadvantage any examinee, or group of examinees, on any basis other than the examinee's lack of the knowledge and skills the test is intended to measure (Nitko, 2001).

Test tryout, administration and quantitative evaluation of the test

Quantitative evaluation (or item analysis) is a numerical method for analysing test items employing student response alternatives or options (Kubiszyn & Borich, 2013). Before one would be able to conduct quantitative item analysis, the test should be administered to a sample with similar characteristics as the actual group who will be taking the final test (Shillingburg, 2016). This is called test tryout. According to Cohen and Swerdlik (2010), for classroom teachers, test tryout (pilot work) need not to be part of the process of developing their tests for classroom use. However, the classroom teacher can engage in quantitative evaluation of test items after test has been administered. The technique will enable them to assess the quality or utility of the items. It does so by identifying distractors or response options that are not doing what they are supposed to be doing. Quantitative evaluation of test items is ideally suited for examining the usefulness of multiple-choice formats (Kubiszyn & Borich, 2013).

Quantitative Evaluation based on classical true-score theory

According to Hambleton and Jones (1993), based on the CTT, quantitative evaluation of test items includes: (a) determining sample-specific item parameters by using simple mathematical methods and moderate sample sizes, and (b) selecting items based on statistical criteria. In CTT, standard item analysis techniques encompass an assessment of item difficulty and discrimination indices and item distractors (Hambleton & Jones, 1993).

Item difficulty index (p/p-value/p-index)

It represents the proportion of students who answered the item correctly (Joshua, 2005; Kubiszyn & Borich, 2013). It could also be defined as the percentage of students who got the item right, or answered correctly each test item (Amedahe & Asamoah-Gyimah, 2016). Difficulty indices vary from "0" for a very difficult item (nobody got it right) to '1" for a very easy item (everybody got it correct). Therefore, the higher the difficulty index of an item, the easier the item, and vice versa (Joshua, 2005). It is calculated by dividing number of students who answer an item correctly by the total number of examinees who attempted the item; and the formula is:

$$p = \frac{\text{Total number of students who got the answer correct}}{\text{Total number of students who attempted the item}}$$

Allen and Yen (1979) have recommended that a good item should have a p-value ranging from .30 to .70; a difficult item should have a p-value below .30 and an item with a p-value above .70 should be considered as easy (see Table 1). With respect to this suggestion, analysis on item difficulty was provided to help indicate effective items, more difficult items and easy items.

Table 1 –*Guideline for Using Difficulty Index*

Difficulty index	Item Evaluation
Above .70	Easy
.30 to .70	Moderate
Below .30	Difficult

Source: Allen and Yen (1979)

Item discrimination (D)

Item discrimination refers to the degree or extent to which an item differentiates between high and low ability test takers (Brown, 2004). Discrimination indices (D-values) differ from –1.00 to +1.00. The higher the index, the better the item (Joshua, 2005). A negative discriminating index means that the greater proportion of lower group answered an item correctly than the proportion of upper group. On the contrary, when the proportion in upper group who answered the item correctly is greater than proportion in lower group who got the item right, the D-index becomes a positive value. The D-value becomes zero (no discrimination), when the proportion in the upper group who got the item correct is equal to the proportion in the lower group who got the item correct (Kubiszyn & Borich, 2013). This value can be calculated by the formula:

$$D = \frac{\begin{array}{c}\text{Number who got item}\\ \text{correct in upper group}\end{array} - \begin{array}{c}\text{Number who got item}\\ \text{correct in lower group}\end{array}}{\begin{array}{c}\text{Number of students in either group (Where group sizes}\\ \text{are equal)}\end{array}}$$

Based on practical experience, Ebel and Frisbie (1991) offered guideline for interpretation of D-values when the groups are established with total test score as the criterion. The guideline is presented in Table 2.

Table 2 –*Guideline for Using the Discrimination Index*

Index of Discrimination	Item Evaluation
D ≥ .40	Excellent discrimination
.30 ≤ D ≤ .39	Good discrimination
.20 ≤ D ≤ .29	Acceptable discrimination
.10 ≤ D ≤ .19	Low discrimination
D < .10	Poor discrimination

Source: Ebel and Frisbie (1991)

Sample size and item statistics (discrimination index and difficulty index)

Because item statistics depend to a great extent on the characteristics of the examinee sample used in the analysis, an important concern of test developers applying CTT is that the examinee sample should be representative of the overall population for whom the test is intended (Hambleton & Jones, 1993). Heterogeneous samples will, generally, result in higher estimates of item discrimination indices, whereas item difficulty estimates rise and fall with high-ability and low-ability groups, respectively. Despite the inherent difficulty of obtaining a representative sample, an advantage of this approach to item analysis is that item statistics can be accurately calibrated on examinee samples of modest size (Hambleton & Jones, 1993).

Item distractors

Students' performance is dependent on how distractors are designed (Dufresne, Leonard, & Gerace, as cited in Quaigrain & Arhin, 2017). According to Cohen and Swerdlik (2010), the quality of each alternative within a multiple-choice item can be readily assessed with inference to the comparative performance of upper and lower scorers. Gronlund and Linn (1990) observed that low-scoring students, who have not grasped the subject content, should choose the distractors more often, whereas, high scorers should reject them more often while choosing the correct option. If students consistently fail to choose certain multiple-choice options, it may be that those options are perhaps implausible and, therefore, of little use as foils in multiple-choice items.

It should be emphasised that when it comes to distractor analysis, the pattern of responses on the distractors for examinees in the upper group are very informative. For instance, according to Nitko (2001), items are miskeyed when the majority of students in the upper group tends to choose one particular incorrect response. Options become ambiguous when students in the upper group are unable to distinguish between the keyed answer and one or more of the foils (Kubiszyn & Borich, 2013). Ambiguity occurs when an item is poorly written or students lack knowledge on the content assessed. Blind guessing occurs where two or more alternatives are approximately and equally plausible to majority of students in the upper group (Nitko, 2001). The effectiveness of a distractor can be measured by using 'option distraction index (or distractor index)'. It is given by the formula:

$$\text{Option Distraction Index} = \frac{\begin{array}{c}\text{Number who chose}\\ \text{the option in lower}\\ \text{group}\end{array} - \begin{array}{c}\text{Number who chose the}\\ \text{option in upper group}\end{array}}{\begin{array}{c}\text{Number of students in either group (Where group}\\ \text{sizes are equal)}\end{array}}$$

Item selection

Items are selected based on two item characteristics: item difficulty and item discrimination (Crocker & Algina, 2008; Hambleton & Jones, 1993; Nitko, 2001). The choice of item difficulty level desired is usually driven by the purpose of the test and the anticipated ability distribution of the group for whom the test is intended. For instance, the case where the purpose of a test is to choose a small group of high-ability examinees for the award of a scholarship. Under this circumstance, items that are usually selected are quite difficult for the population at large (Nitko, 2001).

It is imperative to note that most norm-referenced achievement tests are commonly designed to differentiate examinees with regard to their competence in the measured areas (Nitko, 2001). That is, the test is designed to yield a broad range of scores maximising discriminations among all examinees taking the test. When a test is designed for this purpose, items are generally chosen to have a medium level and narrow range of difficulty (Crocker & Algina, 2008; Hambleton & Jones, 1993).

Revision of the test

Detection of poor items (at least for norm-referenced tests) is quite straightforward and is basically achieved through careful study of item statistics. A poor item is identified by an item difficulty value that is too high or too low, or low discrimination index (Hambleton & Jones, 1993). It is appropriate to point out that classical item analysis procedures, together with an analysis of distractors, have the potential to provide the test developer with invaluable information concerning constructional flaws such as grammatical cues, implausible distractors, double negatives which creeped into the final version of a test used in field testing or main administration to intended students (Hambleton & Jones, 1993). This information will inform the test constructor to qualitatively reexamine test items even by considering problem items identified by students after taking the test and revise the items accordingly.

Principles, Guidelines or Suggestions for Constructing and Improving the Quality of Multiple-Choice Tests

Errors associated with multiple-choice tests negatively affect the reliability and validity of the entire assessment results. To help improve the quality of the multiple-choice test, some principles, guidelines or suggestion have been given by researchers, professionals, and experts in educational assessment of students and psychological testing. In constructing multiple-choice test items, it is quintessential to follow the general principles of test construction and specific item format test construction principles. The outlined general test construction principles and specific principles for the construction of multiple-choice test are organised as indicated by Etsey (as cited in Amedahe & Asamoah-Gyimah, 2016), Kubiszyn and Borich (2013), Joshua (2005), and Nitko (2001).

General principles for test construction

a. Begin item writing far enough in advance that you will have time to revise them.
b. Align the content of the test with your instructional objectives.
c. Include items or questions with varying difficulty level.
d. Match test items to the vocabulary level of the students.
e. Be sure that item deals with an important aspect of the content area.
f. Write or prepare more items than actually needed.
g. Be sure that the problem posed is clear and unambiguous.
h. Be sure that each item is independent of all other items. That is, the answer to one item should not be required as a condition for answering the next item. A hint to one answer should not be embedded in another item.
i. Be sure the item has one correct or best answer on which all experts would agree.

j. Prevent unintended clues to the answer in the statement or question. Grammatical inconsistencies such as ‘a’ or ‘an’ give clues to the correct answer to those students who are not well prepared.

k. Give specific instructions on the test. For example, instructions should be given as to how students are required to answer the questions.

l. Give the appropriate time limit for completion of test.

m. Appropriately assemble the test items. For example, use font size that students can see and read, properly space the items, and arrange test items according to difficulty level (that is, from low to high), number the items one after the other without an interruption, and appropriately assign page numbers.

n. Use appropriate number of items to test students’ achievement.

o. Review items for constructional errors.

p. Evaluate the test items for clarity, practicality, efficiency, and fairness.

Specific principles for constructing multiple-choice items

a. Present the stem as a direct question.

b. Present a definite, explicit and singular question or problem in the stem.

c. Eliminate excessive verbiage or irrelevant information from the stem.

d. Include in the stem any word(s) that might otherwise be repeated in each alternative.

e. Use negatively stated stems carefully (by underlying and/or capitalising or bolding the negative word in the stem.

f. Make alternatives grammatically parallel with each other and consistent with the stem.

g. Make alternatives mutually exclusive or independent of each other.

h. Avoid the use of "none of the above" as an option when an item is of the best answer type.

i. Avoid the use of "all of the above" as part of the options to the stem of an item.

j. Make alternatives approximately equal in length.

k. Present alternatives in logical order (for example, chronological, most to least, alphabetical) when possible.

l. Keep all parts of an item (stem and its options) on the same page.

m. Arrange the alternatives in a vertical manner.

n. Use plausible distractors/options/ alternatives.

Items for the development of the research instrument titled, 'Teachers' Multiple-Choice Test Construction Competence Questionnaire (TTCCQ-MC)' (**see Appendix C**) were objectively, fairly, and comprehensively derived based on the above general test construction principles and specific principles for the construction of multiple-choice tests. The principles also informed the development of the 'Multiple-Choice Test Error Analysis Checklist' **(Appendix K)** for the assessment of format and constructional flaws made by classroom teachers in constructing the multiple-choice tests.

Assessment Competence and Assessment Practice

Competence is the ability of an individual that comprises aspects of knowledge, skills and work attitude matched with standards that have been set (Gilley & Steven, 1989; Lucy, 2014). Competence can also be referred to as the ability to perform the role or task of incorporating knowledge, skills, attitudes and personal values, and the ability to build knowledge and skills dependent on learning and experience (Maba, 2017). From both definitions the following are evident about competence:

a. It is an acquired ability which cannot be seen but can be demonstrated. In both definitions, the phrases: 'Ability… matched with standards that have been set', and 'Ability to perform and to build knowledge and skills dependent on learning and experience' portray the construct as ability. 'Ability matched with standards that have been set' implies that the individual in question has been made aware of certain relevant standards through education (whether formal or informal); therefore, she or he is expected to possess such ability for acting or performing certain roles in accordance with what she or he has acquired.

b. Moreover, from the definition of Maba, new experiences can lead to the integration and modification of the acquired ability or what has already been learnt.

c. The acquired ability has components which are integrated, and the components are knowledge, skills, attitudes, values.

Dependent on the definitional analysis on what is meant by competence, assessment competence can be described as an acquired, modifiable, unobservable but demonstrable ability which is an integration of an individual's knowledge, skills, attitudes and values in/on assessment. As assessment is an important aspect of the activities of teaching, the concept of assessment competence can also be inferred from Adodo's (2013) definition of competency in teaching. Competency in teaching refers to the ability of a teacher to exhibit on the job skills and knowledge gained as a result of training (Adodo, 2013). Inferred, competency in assessment or assessment competency can be defined as the ability of a teacher to exhibit or apply knowledge and skills gained as a result of training in assessment.

Assessment practice, on the other hand, can be defined as set of activities carried out by the teacher in relation to gathering, analysing and interpreting information on student learning, and making relevant educational decisions concerning the student and the instructional process. Though teachers engage in assessment activities, assessment competence answers the question: how well do classroom teachers employ their ability (which is an integration of their knowledge, skills, attitudes and values in/on assessment) to successfully carry out those activities to match expected standards or to ensure improvement in their assessment activities? Since assessment competence in itself cannot directly be observed, such construct can be inferred from what teachers do in terms of how well they go about their assessment practices.

Assessment Competence, Test Construction Competence and Multiple-Choice Test Construction Competence

Assessment results obtained are used in making relevant educational decisions about students, teachers, curricula and programmes, and educational policy. Therefore, the essentiality for teachers to understand and utilise classroom assessments is greater than ever before (Guskey, 2003; Guskey & Jung, 2013). That is, teachers must be as proficient and competent in the area of assessment as they have traditionally been in the areas of curriculum and instruction (Gareis & Grant, 2015). Nitko (2001) has also emphasised teacher's competence in the area of students' assessment by stating that because assessment activities should focus on information one needs to make a particular educational decision, one has to become competent in selecting and using assessments.

Teacher's competence in assessment is specified in standards for teacher competence in educational assessment of students. The standards express specific expectations for obtaining relevant information on knowledge or skills that teachers should possess to perform well in their evaluation effort or assessment of students (Ololube, 2008). The first step taken by the AFT, NCME, and NEA to develop these standards for teacher competence in student assessment is a major step in the right direction of improving the quality of student assessment (Sanders & Vogel, 1993).The standards as developed by AFT, NCME, and NEA (as cited in Nitko, 2001) are the following: Teachers should be skilled in:

1. selecting assessment procedures suitable for instructional decisions.
2. developing assessment techniques suitable for instructional decisions.
3. administering, marking, and interpreting the results of both externally-produced and teacher-produced assessment procedures.
4. using results obtained from assessment when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5. developing valid student or pupil grading procedures which use student or pupil assessments.
6. communicating assessment results to students, parents, other educators and other lay audiences.
7. acknowledging illegal, unethical, and otherwise inappropriate assessment procedures and uses of assessment information.

From the standards, assessment competencies encompass the activities of choosing, developing or constructing, administering and scoring the test, interpreting and using assessment results for relevant educational decisions in an ethically and legally acceptable manner. A close look at the assessment competencies, one would recognise that teachers' assessment competencies involve a standard concerning teachers' ability to develop and construct a test. This standard represents teachers' test construction competencies.

Test construction competence: Ability in detecting individual differences

Measurement is based on simple but crucial assumption that psychological differences exist and can be detected through well-designed measurement process (Furr & Bacharach, 2014). The well-designed measurement process is a question of the quality of the test constructed to detect individual differences on a given psychological construct such as achievement in mathematics. Therefore, constructing a test of good quality is entirely dependent on an individual's ability to quantify the differences among people (Furr & Bacharach, 2014). For example, in the educational settings, the onus rests on the teacher's ability to construct a measuring instrument that would help them detect students who have gained mastery in a given content area and those who have not.

From the aforesaid, teacher's test construction competence becomes an essential ability that is required in detecting individual differences in achievement. Therefore, in this study, it is assumed that regardless of the ability of students, classroom teachers should possess a minimum competence of crafting items with at least acceptable difficulty indices and discrimination indices.

Allen and Yen (1979) have recommended that a good item should have a p-value (difficulty index) ranging from .30 to .70. Thus test items with indices outside this acceptable range would be considered as problem items. Kubiszyn and Borich (2013) point that in selecting items, some experts insist that the discrimination index should be at least .30, while others believe that as long as discrimination index has a positive value, the items discrimination ability is adequate. Naturally, one needs items that have high discrimination values; nevertheless, Kubiszyn and Borich have recommended that one can seriously consider any item with a positive discrimination index for norm-referenced test(s). For the criteria suggested by Allen and Yen, and Kubiszyn and Borich recommendation of at least positive discrimination index for norm-reference tests, the following criteria were used in determining the characteristics of the teacher-made test items:

1. An item is of good quality if it is within the range of .30 to .70 and has a positive discrimination index.
2. An item is a problem item if it is within the range of .30 to .70 but has zero discrimination index.
3. An item is a problem item if it is within the range of .30 to .70 but has a negative discrimination index.
4. An item is a problem item if it falls outside the range of .30 to .70 but has positive discrimination index.
5. An item is a problem item if it falls outside the range of .30 to .70 and has zero discrimination index.
6. An item is a problem item if it fall outside the range of .30 to .70 and has negative discrimination index.

According to Chau (as cited in Hamafyelto et al., 2015) teacher's level of test construction competence is one of the elements that directly influence the quality of his/her test questions. The quality of the test items could help identify students' weaknesses in a given content area so that appropriate measures are put up to enhance teaching and learning (Nitko, 2001). McMillan (2000) has stated that what is most essential about assessment is to understand how general, fundamental assessment principles and ideas can be used to enhance students' learning and teacher effectiveness. Thus, test construction competence as one of the assessment competence standards calls on teachers to be skilled in following certain principles in constructing assessment instrument or methods appropriate for instructional decisions. According to AFT, NCME, and NEA (as cited in Nitko, 2001), classroom teachers who possess this competence will have the conceptual and application skills that follow:

1. They will be skilled in planning the gathering of information that helps the decisions they will make.
2. They will choose the technique which is more suitable to the intent of the teacher's instruction.
3. They will be acquainted with and adhere to appropriate principles for developing and using assessment methods or techniques in their teaching, avoiding common mistakes in student assessment.
4. They meeting this criteria will also be skilled in using student data to analyse the quality of each assessment technique they use.

Under selecting the technique which are appropriate to the intends of the teacher's instruction, teachers have to select an item format or formats that help to effectively assess students on the instructional objectives covered at the end of the instructional period(s). In Ghana, per the SHS teaching syllabi, item formats used in constructing end-of-semester teacher-made examination questions or tests are the multiple-choice test item format and the essay format.

In this study, though teachers' test construction competencies should have covered both item formats, only teachers' competencies in constructing multiple-choice items as part of end-of-semester examination test items were investigated; therefore, the term teachers' multiple-choice test construction competencies. Delimiting teachers' test construction competencies to multiple-choice items is justified because if this study was to investigate the relationship between teachers' test construction competencies (in terms of both multiple-choice and essay items) and the quality of the test items using quantitative evaluation (or item analysis) approach, conducting quantitative item analysis for the essay test items was not practicable.

Empirical Review

Here, findings from previous research conducted outside Ghana and in Ghana in relation to test construction competence, test quality, and years of teaching are discussed. Issues of assessment competencies and practices of classroom teachers have served as the backdrop for research on the quality of teacher-made test across countries (for example, United States of America (USA) and in Africa). The studies indicated that classroom teachers are faced with some challenges in applying basic principles in their testing practices. In Ghana, it is evident from previous works that there is the need to investigate further issues that have been observed concerning assessment competencies or practices, especially among populations where the issues seem not investigated or unexplored.

Test Construction Competence, Test Quality and Years of Teaching

Empirically, the quality (or characteristics) of teacher-made tests has been said to be related to factors such as test construction competence or proficiency and years of teaching experience. Moreover, it is also evidential that test construction competence is influenced by years of teaching. The empirical review as discussed here reveals such a relationship among test construction competence, test quality, and years of teaching.

Studies Conducted outside Ghana

Rivera (2007) conducted a study titled, 'Test item construction and validation: Developing a state-wide assessment for agricultural science education'. The study was carried out in the New York State to develop, validate, and field test separate banks of test items for the animal systems and plant systems content areas. One of the specific objectives of the study was to draft test items and validate the items through experts' judgement and item analysis. Therefore, secondary school teachers of agricultural education were engaged in crafting test items. The quality of the items was assessed through experts' judgement. From general observation, Rivera concluded that the teachers lacked the skills in generating well-constructed items.

According to Agu et al. (2013), in Nigeria, the quality of classroom achievement tests faced criticisms for lack of proper psychometric properties. The issue bothered on teachers' possession or non-possession of competencies in constructing test items. Consequently, they conducted a study titled, 'Measuring teachers' competencies in constructing classroom-based tests in

Nigerian secondary schools: Need for a test construction skill inventory'. In their study, they developed and validated a Test Construction Skill Inventory (TCSI) for assessing the secondary school teachers' competencies in constructing classroom-based tests. The TCSI was also found to be reliable with a coefficient of .73. Therefore, the TCSI was recommended as an important measure for determining the secondary school teachers' test construction skills in Anambra State, Nigeria.

Further, Hamafyelto et al. (2015) carried out research to assess the relationship between commerce teachers' competence in test construction and test quality in Borno State, Nigeria. As part of the research objectives, the areas of competence of Borno State senior secondary schools teachers of commerce in constructing multiple-choice and essay questions were assessed. From the study's results, the examination questions were described as having low content validity (test quality): Most of the items were concentrated on lower levels of the cognitive domain (that is, remembering and understanding). This meant that Borno State senior secondary schools teachers were not competent in constructing examination questions. Accordingly, they recommended workshops and seminars be organised for the teachers to improve their competence in test construction.

In Matabeleland North (Western Zimbabwe), Tshabalala et al. (2015) conducted a study which sought to establish the effectiveness of teacher-made tests in primary schools. The study was meant to expose barriers that hinder the use of teacher-made tests so that practical suggestions could be found to improve the situation regarding classroom testing. They identified a lack of technical know-how to construct proper tests as one of the challenges faced by teachers in their attempts to construct and give teacher-made tests to pupils. Further, information from the study showed that most of the respondents did not consider validity and reliability when constructing and marking teacher-made tests.

Moreover, for hypothesis tested in terms of years of teaching and test construction competencies, Agu et al. (2013) observed that there was a significant difference in the mean ratings of more experienced and less experienced teachers. This difference observed was an indication that the TCSI is sensitive to years of experience. Dosumu (2002) observed that the more experienced a teacher is, the more he begins to understand and appreciate some important test construction skills. The implication of this is that the TCSI could be administered on the teachers bearing in mind their years of experience as an independent variable.

Kinyua and Okunya (2014) also conducted a study to investigate quality of teacher-made tests. The study was specifically carried out in Nyahururu District of Laikipia County in Kenya to establish factors that influence the validity and reliability of teacher made tests. Years of teaching was one of such factors. The findings of the study revealed that teachers with more experience prepared tests which were more valid and reliable (that is, of good quality). This supports the findings of Magno's (2003) study that highly experienced teachers prepared examinations with high validity and reliability.

Studies Conducted in Ghana

It has also been unravelled in the Ghanaian educational settings that most classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction (Amedahe, 1989; Anhwere, 2009; Wiredu, 2013; Quaigrain, 1992; Sasu, 2017). According to Amedahe (1989), to a large extent, secondary school teachers in the Central Region did not follow the basic suggested principles of classroom test construction. Additionally, Quaigrain (1992) indicated that most Ghanaian teachers had inadequate skills for constructing essay type tests. Moreover, Teacher Training College tutors did not adhere to the basic principle of testing in the construction of classroom tests or teacher-made test (Anhwere, 2009). Wiredu (2013) also found that tutors in Nurses' Training Colleges in the Western and Central Regions of Ghana overlooked some basic principles in crafting test items. Also, it was found that Junior High School teachers in the Cape Coast Metropolis did not follow test construction principles to an appreciable level (Sasu, 2017). These findings are tentative answers related to some aspects of Ghanaian teachers' test construction competencies.

In relation to the aforementioned studies on Ghanaian teachers' test construction competencies, Oduro-Okyireh (2008) has given contradicting evidence that teachers in SHSs in the Ashante Region of Ghana follow the principles of test construction. Looking at the different populations previous studies have examined and the mixed nature of findings on some aspects of test construction competence, it is clear that previous studies do not give holistic and consistent view of test construction competencies of teachers in the Ghanaian educational settings. Thus, previous studies related to teachers' test construction competencies arouse curiosity about test construction competencies of teachers in other educational settings in Ghana. Consequently, research is needed to investigate test construction competencies of other populations of teachers in Ghana. SHS teachers in the KSD (in the

Eastern Region of Ghana) is one of these populations. KSD because, in descriptive terms, it is not certain the state of test construction competencies of SHS teachers in the area.

Ghanaian classroom trained and untrained teachers, from the basic level to the university level, construct, administer and score classroom achievement tests regardless of whether they have had training in measurement and evaluation or not (Anhwere, 2009). When classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction, the quality of the tests they construct is questionable. According to Chau (as cited in Hamafyelto et al., 2015), teacher's test construction competence is directly related to ensuring the quality of a test. Poor test quality negatively affects the validity of assessment results (Amedahe & Asamoah-Gyimah, 2016).

From the aforesaid, by implication, when teacher-made tests are low in quality, school administrators and teachers will not be able to make available support and educational opportunities that each student needs (Agu et al., 2013). In other words, lack of or low degree of validity of test results leads to undependable inferences about student learning (Amedahe & Asamoah-Gyimah, 2016; Gareis & Grant, 2015). Based on this, educational decisions such as selection of students for educational opportunities would be wrongfully made.

Questionnaire as a self-report measure has been a common instrument which has been used to investigate test construction competencies or practices of classroom teachers in Ghana (Amedahe, 1989; Quagrain 1992; Oduro-Okyireh, 2008; Anhwere, 2009; Wiredu, 2013; Sasu, 2017). Amedahe (1989), Quagrain (1992), and Wiredu (2013) verified teachers' responses to questionnaire items by directly examining samples of tests developed by the teachers for constructional flaws. The direct examination provided some qualitative information with regard to the quality of the teacher-made tests. However, the study of Oduro-Okyireh (2008), Anhwere (2009), and Sasu (2017) involved no direct analysis of samples of teachers-made tests to verify teachers' responses to questionnaire items on test construction practices.

To go beyond just relying on responses of teachers on self-report measures, Oduro-Okyireh (2008) suggested research be conducted on the quality of teacher-made tests. The quality of tests developed is investigated using item analysis. Therefore, Oduro-Okyireh's suggestion implies a direct assessment of actual test made by classroom teachers. The direct assessment procedure will help to validate teachers' responses to any self-report measure used in the assessment of their practices or competencies. Ary et al. (2010) have identified that direct observation of the behaviour of a random sample of respondents is a brilliant strategy to validate their responses to a self-report measure. There are two main approaches to item analysis –qualitative and quantitative approaches (Kubiszyn & Borich, 2013).

In Ghana, in terms of evaluating the quality of teacher-made test items, from previous studies, none of the item analysis approaches has empirically been employed to investigate the quality of test items constructed by classroom teachers in the KSD. For example, Quaigrain and Arhin (2017) conducted quantitative item analysis study which focused on item and test quality and explored the relationship between difficulty index and discrimination index with distractor efficiency. However, the study was conducted among first-year students pursuing Diploma in Education at Cape Coast Polytechnic. Accordingly, there was the need to also conduct item analysis study among SHS teachers in the KSD to validate their responses to the self-report measure that was administered.

Quantitative item analysis is more feasible and useful for multiple-choice test items than essay items (Kubiszyn & Borich, 2013). Therefore, in this study, the quality of multiple-choice test items constructed by classroom teachers was investigated excluding any other item formats. The output from quantitative item analysis on the multiple-choice test items require some form of explanations; therefore, qualitative item analysis approach was also employed to identify certain errors associated with the construction of teacher-made multiple-choice tests among SHS teachers in the KSD.

According to Chau (as cited in Hamafyelto et al., 2015), teacher’s test construction competence is directly related to ensuring the quality of a test. Nevertheless, it appears no study has been conducted to quantitatively examine Chau’s perspective by finding out the relationship between multiple-choice test construction competencies and the quality of the test items among SHS teachers in Ghana. Thus, in this study, the relationship between multiple-choice test construction competencies and the quality of the teacher-made test items among SHS teachers in the KSD was investigated.

From the perspective of Leedy and Ormrod (2013), it is vital to investigate variable(s) that might help explain the relationship between two variables under investigation. Years of teaching has been identified as a variable that may influence test construction competence and quality of test items (Amedahe, 1989; Marso & Pigge, 1989; Dosumu, 2002; Magno, 2003;

Agu et al., 2013; Kinyua & Okunya, 2014). However, in Ghana, it appears previous studies have not looked at the effect of years of teaching on the relationship between test construction competence and quality of test items. Therefore, in this work, there was the need to investigate the effect of years of teaching on the relationship between multiple-choice test construction competencies and the quality of the items among SHS teachers in the KSD.

It has been observed in the literature that to improve the reliability and validity of assessment results, there is the need to lengthen the assessment procedures (Allen & Yen, 2002; Crocker & Algina, 2008; Nitko, 2001). This implies that one should put little confidence in a student's performance based on few multiple-choice items used in assessing the student's achievement (Crocker & Algina, 2008). However, Crocker and Algina (2008) have stated that improving test quality by increasing test length works well when all the items are representative of the domain sampled and have the same or similar level of appropriate difficulty and discrimination indices. This means that the greater the number of test items appropriate in difficulty and discrimination indices, the degree of errors in students observed scores reduces; hence, resulting in an appreciable level of reliability coefficient.

Reliability coefficient is the squared correlation between observed scores and true scores (Furr & Bacharach, 2014). This implies that reliability coefficient of the test will reduce when there are more problem items contributing to errors in observed scores. On the other hand, the presence of more good items reduces the errors; therefore, resulting in improved reliability coefficient (Crocker & Algina, 2008). Accordingly, Crocker and Algina

(2008) have stated that items must be well written and free of technical flaws that may cause examinees to respond on some basis unrelated to the content.

From the aforesaid, to achieve an appreciable level of reliability coefficient, the theory endorses that it is good for teachers to construct a relatively adequate number of test items without problem items or which has relatively few problem items. Nevertheless, Downing (2003) has stated that teachers perceive test construction procedures as waste of time and non-motivating. Such a line of thinking can negatively influence their level of attention in ensuring that test items are not problem items or there are few of them. This suggests the need to investigate among classroom teachers what happens to: (a) the number of good items in a test when test length is either increasing or decreasing; (b) the number of problem items as test length is either increasing or decreasing.

However, in Ghana, previous studies (examples: Amedahe, 1989; Oduro-Okyireh, 2008; Anhwere, 2009; Wiredu, 2013; Sasu, 2017) conducted in relation to test construction have not examined the relationship between test length and the number of good and problem items identified with tests constructed by classroom teachers. Hence, there is the need to examine the relationship between: (a) test length and the number of good items; and (b) test length and the number of problem items observed with items constructed by classroom teachers in the country.

With regard to test construction competencies and practices, previous studies conducted in Ghana at the SHS level have covered English Language, Mathematics, History, Geography, and Religious Studies teachers. Amedahe's (1989) study was carried out among English Language, Mathematics and

History teachers. To refute or confirm the tentative findings of his study, he recommended that an extensive research is needed to cover most of the subject teachers in the SHSs. Nonetheless, since then, it appears only two studies (Oduro-Okyireh, 2008; Quaigrain, 1992) have been conducted investigating test construction competencies and practices of teachers at the SHS level. Quaigrain's (1992) study focused on History, Geography, Religious Studies and English Language teachers, while Oduro-Okyireh's (2008) study focused on English Language, Core Mathematics and Integrated Science teachers.

Most of the subjects taught at the SHS level have not been covered by previous works on test construction competencies of SHS teachers in Ghana. It is against this background that this present study was extended to cover SHS teachers in the KSD who teach Financial Accounting, Cost Accounting, Business Management and Economics. These subject teachers were selected in addition to English Language, Core Mathematics, and Integrated Science teachers in the KSD SHSs.

From the aforesaid, it is evident that not all SHS teachers in the KSD were included in this study. Covering all SHS subject teachers in the KSD might be of interest, however, it was not feasible to provide the current state of their multiple-choice test construction competencies. That is, direct assessment of test items is time consuming using item analysis approach. Therefore, per the period given for the completion of this research work, it was impractical to validate the responses of all the subject teachers to the questionnaires with respect to directly analysing samples of their multiple-choice test items.

Conceptual Framework

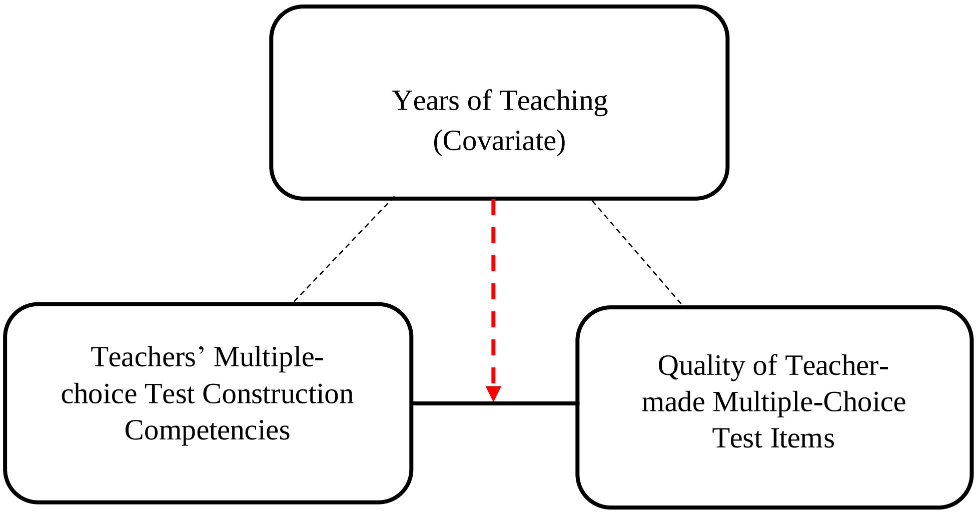

Years of Teaching
(Covariate)
Teachers' Multiple-choice Test Construction Competencies
Quality of Teacher-made Multiple-Choice Test Items

Figure 1: Conceptual Framework on the Relationship between Teachers' Multiple-Choice Test Construction Competencies and Quality of Teacher-Made Multiple-Choice Test Items, as well as the Influence of Years of Teaching on such Relationship

Source: Researcher's own framework (2019)

Chapter Summary

Investigating research problems and issues observed in the area of test construction competencies of classroom teachers is of concern not only in Ghana but among other countries outside Ghana. Thus, issues pertaining to test construction competencies of classroom teachers have served as the backdrop for research on the quality of teacher-made tests across countries (for example, the United States of America (USA) and in Africa). The studies indicated that classroom teachers are faced with some challenges in applying basic principles in their testing practices. In Ghana, it is evident from previous works that there is the need to investigate further issues that have been observed with respect to assessment competencies or practices, especially among populations where the issues seem not investigated or unexplored.

Observations made with regard to the literature concerning test construction competence and quality of teacher-made tests (that is, Amedahe, 1989; Marso & Pigge, 1989; Dosumu, 2002; Magno, 2003; Agu et al., 2013; Kinyua & Okunya, 2014) give the impression that though test construction competencies might be related to the quality of test items, years of teaching might serve as a third variable that may influence such a relationship. Besides, another issue that is related to test construction competence and test quality is the number of test items teachers can construct based on their test construction competencies, and how the number of items generated (test length) is related to the number of good items and problem items identified with tests constructed by classroom teachers.

Previous studies conducted in Ghana (examples: Amedahe, 1989; Oduro-Okyireh, 2008; Anhwere, 2009; Wiredu, 2013; Sasu, 2017) suggest that the extent to which issues related to teachers' test construction competencies and quality of test items have been explored in the country is low. Nevertheless, exploring issues related to test construction competencies of classroom teachers is very essential. This is because Chau (as cited in Hamafyelto et al., 2015) has stated that a teacher's test construction competence is directly related to ensuring the quality of a test he or she develops. Based on this premise, when classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction, it will result in crafting items with low quality, which, in turn, negatively affect the validity of assessment results (Amedahe & Asamoah-Gyimah, 2016).

According to Agu et al. (2013), when tests constructed by classroom teachers are low in quality, school administrators and teachers are not able to make available support and educational opportunities that each student needs. When students miss relevant educational support, they will not be able to maximise achievement, which is the main purpose of education. Therefore, in the Ghanaian educational settings, there is the need to investigate the relationship between the multiple-choice test construction competencies of SHS teachers in the Kwahu-South District and the quality of test items they construct, as well as the effect of years of teaching on such relationship.

Empirical evidence obtained as a result of examining the relationship among the variables will help teachers and other educators to put appropriate measures or educational programmes leading to sustainable or improved test construction competencies. These competencies will help ascertain lapses in students' acquisition of knowledge through the use of good test items, so that the necessary support can be given to students to help them maximise their academic achievement.

CHAPTER THREE

RESEARCH METHODS

Introduction

The chapter discusses how the study was conducted. It is presented in seven sections namely research design, study area, population, sampling procedure, research instrument, data collection procedure, and the how data collected was analysed.

Research Design

Research design is the overall plan for gathering data to address the research question. It is also the precise data analysis procedures or techniques that the investigator intends to use (Fraenkel, Wallen, & Hyun, 2012). To examine the relationship among the variables under investigation, the quantitative approach was employed using correlational research design.

The main purpose of a correlational research design is to clarify one's understanding of significant occurrences by identifying relationships among variables (Fraenkel et al., 2012). In other words, a correlational research design helps to search for and examine the degree to which one or more relationships of some kind exist. The design calls for no manipulation or intervention on the part of the investigator other than administering the instrument(s) relevant towards the collection of the data desired (Fraenkel et al., 2012).

It must be emphasised that studies that involve examining the correlation between or among variables do not, in and of themselves, establish cause and effect (Fraenkel et al., 2012). However, correlation studies can give the idea of cause and effect. So, variables found not to be related or only slightly related (that is, when correlations below .20 are obtained) would be dropped from further consideration, while those found to be more highly related (that is, when correlations beyond –.40 or +.40 are obtained) would often bring to the investigators attention to further the research, using an experimental design, to understand whether the relationships are indeed causal (Fraenkel & Wallen, 2009).

Correlational research design is sometimes referred to as a form of descriptive research because it describes an existing relationship between variables –thus, also labelled as descriptive correlation design (Fraenkel et al., 2012). However, how it describes relationship is to a certain extent different from the descriptions identified in other types of studies (such as survey research) (Fraenkel et al., 2012). Generally, one could carry out this type of research to look for and to describe relationships that may exist among naturally occurring phenomena, without making the effort in any way to alter these phenomena (Fraenkel & Wallen, 2009).

Polit and Beck (2004) have explained that, in carrying out a study, selecting a good research design should be guided by whether the design does the best possible job of providing reliable answers to the research question. Accordingly, although the discovery of a correlational relationship does not establish a causal connection, in this research, this design will help to collect data to answer the research questions, and test the hypotheses geared towards identifying the current state or nature of the variables to be investigated in the KSD.

Study Area

The study area selected for this research is the KSD in the Eastern Region of Ghana. The district shares common boundaries with Kwahu East to the North, Asante-Akim South to the West, Kwahu West Municipality and East Akim District to the South and Fanteakwa District to the East. Precisely, it lies between latitudes 6° 35" N and 6° 45"N and longitude 0° 55" W and 0° 20"W. The total land size of KSD is 602km². There are four SHSs in the district namely Mpraeso Senior High School (MPASS), St. Pauls Senior High School (PASCO), Bepong Senior High School (BESCO), and Kwahu Ridge Senior High School (KRISTEC). These schools are found in the Mpraeso, Asakraka, Bepong, and Obo townships within the KSD respectively.

Population

A population is defined as the entire group of persons about which an individual wants information (Moore, 2001). It also refers to the entire group of individuals to whom the findings of a study apply (Ary et al., 2010). In a study, the researcher defines the specific population of interest to him or her (Ary et al., 2010). Consequently, the population of interest for this research work is defined as all form 1, form 2 and form 3 class teachers in the following subject areas: Financial Accounting, Cost Accounting, Business Management, Economics, English Language, Integrated Science and Core Mathematics SHS teachers in the KSD. The total number of teachers that constituted the population for the study was 157. The distribution of teachers in the respective SHSs in the KSD is presented in Table 3.

Table 3 –*Distribution of Teachers in the Respective SHSs and Subject Areas*

Subject Areas	SHSs in the KSD				
	KRISTEC	MPASS	BESCO	PASCO	Total
Business Management	3	2	2	3	10
Financial accounting	3	2	2	2	9
Cost Accounting	2	2	1	1	6
Economics	6	4	7	5	22
Core Mathematics	14	11	5	6	36
Integrated Science	11	17	7	10	45
English Language	12	6	5	6	29
Total	51	44	29	33	157

Source: Field data (2019)

Sampling Procedure

Research participants

Purposive sampling technique was used to arrive at a sample of 47 teachers (n = 47) out of the population of 157. The term purposive sampling is often used to denote a systematic strategy of selecting or recruiting specific types of people fitting a given criteria that are important to the research questions or purpose of the study (Barker, Pistrang, & Elliot, 2002; Howitt & Cramer, 2011). Cohen, Manion, and Morrison (2000) have indicated that it involves the activity of handpicking the cases to be included in the sample that is satisfactory to specific needs or purpose. Using this technique implies that some members of the population defined for a study will be excluded and others included (Clark & Creswell, 2015; Cohen et al., 2000). Based on the aforesaid, the following inclusion and exclusion criteria were used to select participant most appropriate to meet the purpose of the study:

Inclusion and exclusion criteria

The nature of the study required that all participants were willing to:

1. respond to a questionnaire;
2. make available all of the following documents: (a) copies of their latest end-of-semester self-constructed and administered multiple-choice test and (b) its marking scheme; and (c) students' responses on the administered end-of-semester multiple-choice test items.

Therefore, two teachers, out of the total population, were excluded from the study because of their unwillingness to participate. Moreover, other teachers excluded from the study were not able to provide all of the following documents over the data collection period: (a) copies of their latest end-of-semester self-constructed and administered multiple-choice test and (b) its marking scheme; and (c) students' responses to the administered end-of-semester multiple-choice items.

This type of sampling technique was suitable for the research because it helped to arrive at a sample most appropriate to learn about the central phenomenon. According to Fraenkel et al. (2012), drawing conclusion about a population after studying a purposive sample is never totally suitable, since researchers can never be sure that their sample is perfectly representative of the population. However, whenever purposive sampling is used, generalisation is made more plausible if data are presented to show that the sample is representative of the intended population on at least some relevant variables (Fraenkel et al., 2012).

Sampling of tests and problem items for qualitative items analysis

To examine the format and constructional flaws identified with the teacher-made tests, qualitative evaluation was purposively performed for Business Management tests and Core mathematics tests. Performing qualitative evaluation in the other subject areas could have revealed more specific problems in all the subject areas that contributed to unacceptable difficulty and discrimination indices. However, such general evaluation was not feasible in terms of easy access to subject area experts in English, Financial Accounting, Economics, Cost Accounting, and Integrated Science to help in qualitatively examining test items for constructional flaws such as ambiguities, more than one answer, and clues to correct answers.

Based on the quantitative item analysis, 82 and 155 items were identified as problem items for Business Management and Core Mathematics tests respectively. In all, there were 237 (that is, 82 + 155) problem items across the tests. To obtain a representative sample of the problem items, Krejcie and Morgan (1970) sample size determination criteria was used. Consequently, 144 sample of the problem items were appropriate to be qualitatively examined. Further, to get a fair representation of the problem items for each test, stratified (proportionate) sampling technique was used together with simple random sampling technique to select the items for qualitative evaluation. The number of problem items and items that were examined for each test is presented in Table 4.

Table 4 –*Sample Distribution of Items for Qualitative Evaluation*

Subject Area	Test type	Number of Problem items	Sample examined qualitative
Business	Test 1	16	10
Management	Test 2	16	10
	Test 3	30	18
	Test 4	20	12
Core	Test 5	11	7
Mathematics	Test 6	15	9
	Test 7	13	8
	Test 8	25	15
	Test 9	31	19
	Test 10	15	9
	Test 11	22	13
	Test 12	23	14
Total		237	144

Source: Field data (2019)

Data Collection Instruments

Teachers' multiple-choice test construction competence questionnaire

A 20-item instrument (see **Appendix I**) titled: Teachers' Multiple-Choice Test Construction Competence Questionnaire (TTCCQ-MC) was used to measure teachers' multiple-choice test construction competencies. I developed this instrument based on comprehensive literature review on test construction competence. The instrument is made up of two sections namely 'Section A' and 'Section B'. Section A is made up of items that help to obtain information on teachers' demographic variables. Section B is made up of items that help to measure teachers' multiple-choice test construction competencies. The scale of measurement that is used for the items under Section B is 4-point Likert-type scale on a continuum of strongly disagree (SD), disagree (D), agree (A) and strongly agree (SA).

A questionnaire is a self-report measuring device in which each respondent provide written responses to set of questions or mark items that indicate their responses (Ary et al., 2010; Johnson & Christensen, 2004). Normally, the primary purpose of a questionnaire is to describe the distributions of variables (such as motivation) in a specified group (for example, grade six pupils) (Ary et al., 2010).

A questionnaire is mostly used in quantitative studies. Its usage is very convenient whenever the sample size for a given study is large enough to make it uneconomical to rely on other methods such as interviews and observation (Amedahe & Asamoah-Gyimah, 2014; Osuola, 2001). Another benefit associated with the use of questionnaire is that its administration is easy and it takes relatively less time for participants to provide their responses to the set of items (Osuola, 2001). Nevertheless, Ary et al. (2010) have stated some likely problems that may affect the validity of a questionnaire. The following are some of the problems:

1. Most of the times, participants report what they believe or perceive is true but is not.
2. Participants may provide untrue responses that are more socially acceptable than what is happening in reality.
3. Participants may give answers that they perceive the investigator wants to hear.

Questionnaire was chosen because the study was quantitative in nature. Also, the use of questionnaire allows for broad geographical sampling and it can be used to cover a large sample as well (Osuola, 2001; Amedahe, 2002). Considering the large number of teachers estimated for the sample size, using a technique other than a questionnaire would make access to such large number of teachers difficult.

Content validity of TTCCQ-MC

In devising a set of items to measure the test construction competencies of the research participants, the content validity of the research instrument was established by making sure that it objectively, fairly and comprehensively covered the domain that it purports to cover. The 23-item TTCCQ-MC was examined by my supervisors who are experts in test construction for its content appropriateness and clarity. After my supervisors' examination of the instrument, the corrections were effected. For example, item 10 which was stated as '*when constructing multiple-choice test, I make options mutually exclusive*' was reconstructed as '*when constructing multiple-choice item, I make options independent of each other*' so that respondent could understand without much difficulty. The final draft of the 23-item TTCCQ-MC used for field testing is presented in **Appendix D.**

Field testing instrument to establish the construct validity of TTCCQ-MC

Having devised the 23-item questionnaire to measure the test construction competencies of the research participants, the 23-item TTCCQ-MC was field tested using a sample of 130 Financial Accounting, Cost Accounting, Business Management, Economics, English Language, Integrated Science and Core Mathematics SHS teachers in the Kwahu-East District (KED). These teachers were selected for the field testing of the instrument because their characteristics (such crafting test items to assess students' achievement, gender, major subject area they teach, major class they teach, and their highest qualification) were similar to that of the actual population for the study. After the data collection, principal component analysis (PCA) was conducted on the items with orthogonal rotation (varimax). The main aim of conducting the PCA was to reduce the 23 items that measures test construction competence into more manageable variables or potential themes and to establish the validity of the items.

The PCA analysis established the validity of the questionnaire, and the sub-themes within it. Three factors were detected from the PCA: competence in assembling the items, competence in achieving content validity, and competence in handling the items' alternatives. Reliability analysis was also conducted to confirm whether the questions included within those factors were answered consistently by the research participants. This helped to examine the reliability of the three factors.

Testing the assumptions for PCA

In testing the assumptions for PCA, the determinant of the correlation matrix as an indicator of multicollinearity was .005, which was substantially greater than the minimum recommended value of .00001. This meant that multi-collinearity was not a problem in conducting PCA. The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis, KMO = .64 and all KMO values for individual items were > .50, which was above the acceptable limit of .50 (Field, 2018). This meant that the sample size was adequate for PCA. The Bartlett's test of sphericity was significant (χ^2 (253) = 644.421, p < .001). This indicated that correlations between items were good for PCA. The results on determinant value KMO and Bartlett's test are presented in Table 5.

Table 5 –*Determinant, KMO and Bartlett’s Test*

Description		Statistic
Determinant		.005
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.637
Bartlett's Test of Sphericity:	Approx. Chi-Square	644.421
	df	253
	Sig.	.000

Source: Field data (2019)

Factor extraction based on PCA

After satisfying the assumptions for PCA, an initial analysis was run to obtain eigenvalue for each component in the data. Eight components had eigenvalues over Kaiser’s criterion of 1 and in combination explained 61.90% of the total variance. The scree plot (see Figure 2) showed point of inflexion that would justify retaining three components. Given the sample size of 130, and 23 items, the Kaiser’s criterion on eight components, and convergence of the scree plot on three components, parallel analysis (PA) was conducted in addition to examine the appropriate number of components to maintain. Hayton, Allen and Scarpello (2004) have point out that PA helps to identify the meaningful number of emerging factors from the set of items which are to be maintain.

The results from the PA (see **Appendix E**) indicated that only three components should be maintained in the final analysis since the random eigenvalues for the first three factors (that is, 1.8500, 1.7123, 1.5948) from the PA is less than the initial eigenvalues for the first three factors (3.706, 2.483, 1.887) from the PCA (see **Appendix F**). According to Hayton et al. (2004) components corresponding to actual eigenvalues that are greater than the random eigenvalues from the PA should be retained. Therefore, the meaningful factors to be maintained based on the PA corresponds to the suggested number of factors to be retained based on visual examination of the scree plot.

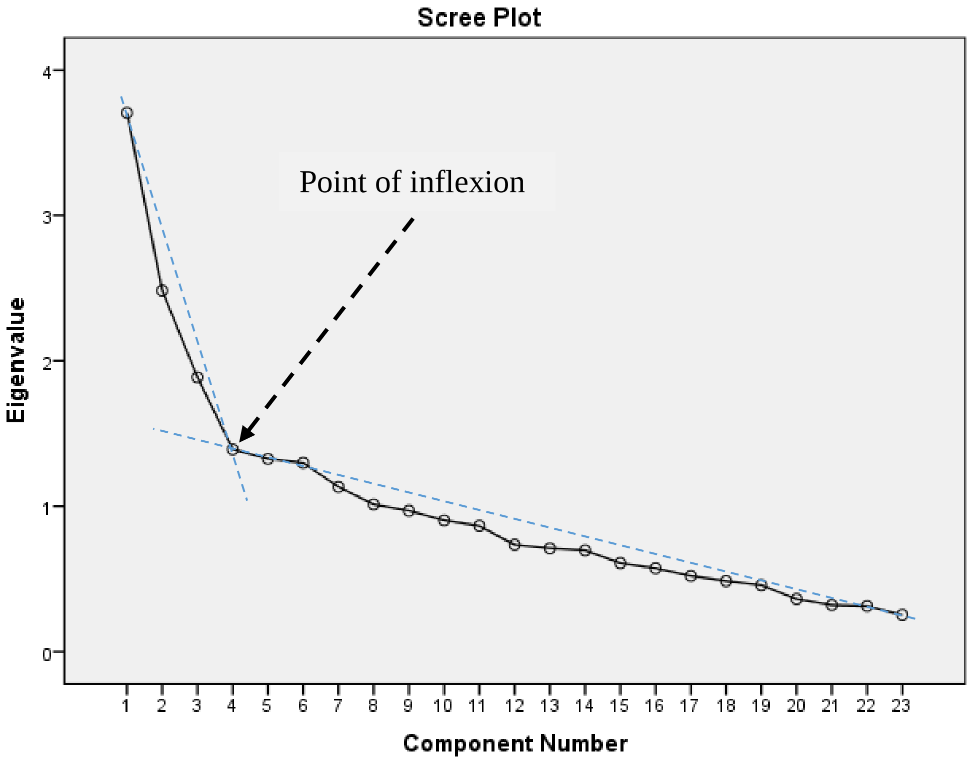


Figure 2 –Scree Plot for Factor Extraction
Source: Field data (2019)

Factor rotation in PCA and interpretation

The main method of factor rotation used is the varimax orthogonal rotation, using absolute cut-off point of .40 for factor loadings. The main reason behind suppressing loadings below .40 was centred on Stevens' (2002) recommendation that this cut-off point was suitable for interpretative purposes (that is, loadings greater than .40 represent substantive values). With this method, the three factors that were produced explained approximately 35.11% of variance in the data: Factor 1 (seven items, 12.01% of explained variance), Factor 2 (seven items, with 11.84% of explained variance), and Factor 3 (seven items, 11.27% of explained variance). To make sense out of the extracted factors, factor 1 could be interpreted as Competence in achieving content validity, factor 2, competence in assembling the items and factor 3, competence in handling the items' alternatives. **Appendix F** shows the percentage of variance explained by each factor, the eigenvalues associated with each of the factors, and the factor loadings after rotation.

Reliability analysis: Internal consistency of the TTCCQ-MC

Reliability analysis is often seen as a logical follow-on from factor analysis. It helps to examine whether the research instrument possesses internal consistency per the responses given by the participants (Mayers, 2013). Therefore, by using Cronbach's alpha, reliability analysis was conducted to assess the extent to which the items within each sub-theme or factor elicit consistent responses. Cronbach's alpha was appropriate because the Likert scale used in measuring teachers' test construction competencies meant that the items had wide range of scoring weights. When test items are not dichotomously scored (for instance, using a scale of 1 to 4 is used), a more general version of KR20 known as Cronbach's alpha is the appropriate reliability technique to be used (Amedahe & Asamoah-Gyimah, 2016; Crocker & Algina, 2008; Nitko, 2001).

Reliability of TTCCQ-MC after pilot testing

Factor 1 (Competence in achieving content validity) showed high internal consistency with α of .68. Cronbach's alpha would not benefit from the removal of any item. This level of internal consistency was also seen for Factor 2 (competence in assembling the items; α = .69). The internal consistency for the Factor 3 (competence in handling the items' alternatives; α = .50) could be described as low. However, the α value for Factor 3 could be improved by the removal of item 11 from the group of questions being measured. Consequently, the alpha value based on the six items (that is, after the removal of item 11) for factor 3 is .65. The overall coefficient of reliability of the instrument after removal of item 11 is .72. Reliability statistics for the factors are presented in Table 6.

Table 6 –*Reliability Statistics*

Factors	Cronbach's Alpha	Number of Items
Factor 1	.681	7
Factor 2	.692	7
Factor 3 (item 11 inclusive)	.497	7
Factor 3 (item 11 removed)	.647	6

Source: Field data (2019)

Rerun of PCA to check factor structure upon removal of item 11

The assumptions for PCA were still met for the remaining 22 items (the 23 items minus item 11) with determinant value of .006, KMO = .64, and Bartlett’s test of sphericity (χ^2 (231) = 622.299, p < .001), which can be described as significant. Results on determinant value, KMO and Bartlett’s test after removal of item 11 are presented in Table 7.

Table 7 –*Determinant, KMO and Bartlett’s Test after Removal of Item 11*

Description		Statistic
Determinant		.006
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.642
Bartlett's Test of Sphericity:	Approx. Chi-Square	622.299
	df	231
	Sig.	.000

Source: Field data (2019)

Further, from visual examination of the scree plot (Figure 3) and parallel analysis **(Appendix G)**, the number of factors did not change. That is, three factors were still extracted with the same set of items falling under each of them, except for factor 3 which now is made up of six items due to the removal of item 11.

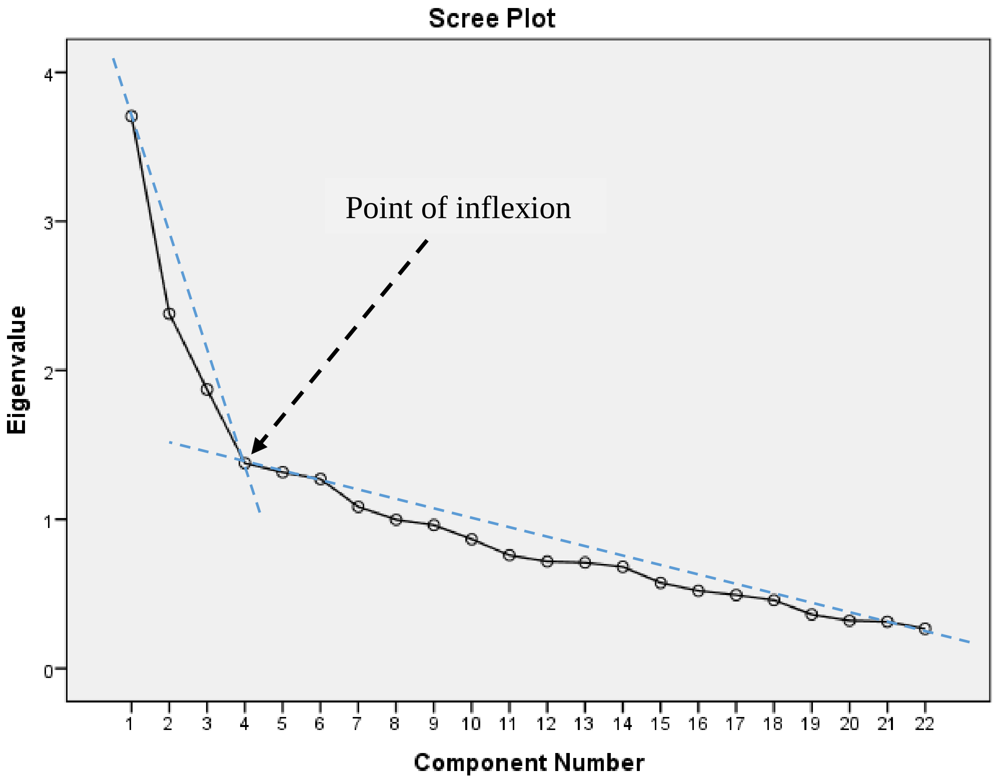


Figure 3 –Scree Plot for Factor Extraction (after removal of Item 11)
Source: Field data (2019)

However, the result from the factor loadings after rotation indicates that the set of items that previously constituted factor 2 is now labelled factor 1(competence in assembling the items; α = .69), the set of items that established factor 1 is now factor 2 (competence in achieving content validity; α = .68). The set of items that constituted Factor 3 remained factor 3 (competence in handling the items' alternatives; α = .65).

With the varimax orthogonal rotation, the three factors that were produced explained approximately 36.17% of variance in the data (as compared to the initial explained variance of 35.11%): Competence in assembling the items (Factor 1, seven items, with 12.34%), Competence in achieving content validity (Factor 2, seven items, 12.23%), and Competence in handling the items' alternatives (Factor 3, six items, 11.60%). **Appendix H** shows the percentage of variance explained by each factor, the eigenvalues associated with each of the factors, and the factor loadings after rotation after removal of item 11.

Decision on the final items for the TTCCQ-MC

The results from PCA and reliability analysis endorsed 20-item TTCCQ-MC (**Appendix I**). That is, considering the absolute cut off value of .40 for factor loadings, item 6 and item 8 did not load on any of the factors since their loadings were below the cut off value. However, the rest of the remaining 20 items had factor loadings greater than the cut off value of .40. Therefore, the final 20 items, with overall reliability coefficient of .72 for pilot testing, were considered valid for measuring the construct: Teachers' multiple-choice test construction competence. Based on the final TTCCQ-MC, item 1 to item 7 represent factor 1 (competence in assembling the items), item 8 to item 14 constitute factor 2 (competence in achieving content validity), and item 15 to item 20 make up factor 3 (competence in handling the items' alternatives).

Objectives of the three components

1. **Competence in assembling test items:** This helped to assess how well the teachers are able to organise test items and format the test to permit students to demonstrate their ability on a given content.
2. **Competence in achieving content validity:** This helped to describe how well the teachers are able to employ their ability in measuring what students have achieved in relation to instructional objectives.

3. **Competence in handling items' alternatives:** This helped to identify how well the teachers are able to craft alternatives in order to discriminate students with respect to a particular content achieved at the end of instructional session(s).

Reliability of TTCCQ-MC for the main data

Reliability analysis conducted on the main data gave evidence that the overall reliability coefficient for the 20-item questionnaire using Cronbach's alpha (α) was .75. This overall α value for the main data is consistent with the overall coefficient of reliability for the pilot testing which is .72. Factor 1 (competence in assembling the items) showed moderate internal consistency with α of .57. This level of internal consistency was also seen for Factor 2 (competence in achieving content validity; α = .56). Factor 3 (competence in handling the items' alternatives showed high internal consistency with α of .63). Cronbach's alpha would not benefit from the removal of any item.

Document examination

Students' responses on multiple-choice test items; Marking schemes; and Copies of latest end-of-semester administered teacher-made test

Document examination in this research covered students' responses on multiple-choice test items for end-of-semester administered teacher-made tests, copies of their marking schemes and end-of-semester teacher-made test. Using the marking schemes and students' responses on multiple-choice test items administered by the research participants, quantitative item analysis was performed to assess the characteristics of the multiple-choice test items for each of the classroom teachers. The assessment criteria used in assessing the characteristics of the items are based on the following item analysis descriptive statistics indices: difficulty index (p-value), and discrimination index (DI).

Criteria used for assessing the characteristics of the test items in quantitative item analysis

Based on the literature reviewed, the criteria suggested by Allen and Yen (1979) in terms of acceptable difficulty indices ranging from .30 to .70 and Kubiszyn and Borich (2013) recommendation of at least positive discrimination index for norm-reference tests, the following criteria were used in determining the characteristics of the teacher-made test items:

1. An item is judged as a good item if it is within the range of .30 to .70 and has a positive discrimination index.
2. An item is a problem item if it is within the range of .30 to .70 but has zero discrimination index.
3. An item is a problem item if it is within the range of .30 to .70 but has a negative discrimination index.
4. An item is a problem item if it falls outside the range of .30 to .70 but has positive discrimination index.
5. An item is a problem item if it falls outside the range of .30 to .70 and has zero discrimination index.
6. An item is a problem item if it fall outside the range of .30 to .70 and has negative discrimination index.

Criteria used for assessing the characteristics of a test in qualitative item analysis

In addition, with regard to qualitative evaluation of the teacher-made tests for format and constructional flaws, the participants' end-of-semester administered Business Management and Core Mathematics multiple-choice tests were assessed for errors using the 'Multiple-Choice Test Error Analysis Checklist' (**Appendix K**).

Ethical Considerations

It is vital to consider ethical principles to be involved for the success of every research. Thus, in this research, ethical issues that were taken into consideration are: (a) examination of whether psychological harm could come to any of the respondents as a result of their participation in the research, (b) confidentiality, (c) anonymity and (d) informed consent for voluntary participation.

Before the collection of data from respondents, to ensure that the study does not leave them with any psychological harm, they were not coerced by any means to participate, the purpose, objectives and significance of the study were explained to them. Various questions for clarifications were addressed. In addition, they were assured that all provided information will be treated strictly as confidential and anonymous. All those willing to participate were given consent forms (**Appendix C**) to sign as evidence to their understanding of the terms of the study and agreement to voluntarily participate, and that represents their informed consent. Informed consent is an agreement made by research participants indicating their willingness to take part in a study after acquiring knowledge about the procedures involved in the research (Neuman, 2014).

With respect to confidentiality, the respondents were told that information provided (such as their names, personal contacts) for the purpose of follow-ups shall not be made known in reporting the research. According to

Neuman (2014), concerning confidentiality, data should be reported in a way that does not associate specific persons to responses provided. Anonymity was also ensured through the use of codes such as 'ST ELF3, ST CMF1' instead of indicating their names on the research instrument. Oliver (2010) has expressed that anonymity is an essential issue in research as it gives the respondents the opportunity to have their identity concealed.

Data Collection Procedures

A copy of the research proposal was sent to the Institutional Review Board (IRB) for ethical clearance (**Appendix A**). When the ethical clearance was given, I visited the participating schools. Copies of introductory letter (**Appendix B**) taken from the Department of Education and Psychology were given to heads in the selected schools to seek their consent, permission and co-operation. Questionnaires were then distributed to the 47 research participants and they were asked to fill the instrument themselves. There was 100 percent return rate. The necessary documents (students' responses on multiple-choice test items, copies of the latest or available administered end-of-semester teacher-made multiple-choice test and their marking schemes) for the 2018/2019 academic year were also obtained from them. The main data collection period lasted for a period of one month and three weeks (that is, from February, 2019 to April, 2019).

Data Processing and Analysis

The data collected in this study was edited, coded and statistically analysed with descriptive statistical tools based on the research questions and hypotheses. Univariate descriptive statistical tools such as frequency, percentage, mean, standard deviation are used when analysis is to be performed on a single variable (Huck, 2012; Mayers, 2013). Demographic data, research questions 1, 2, and 3 were analysed using univariate descriptive statistical tools.

The bivariate statistical technique was used to address research hypotheses 1, 2, and 3. The multivariate technique was considered to address research hypothesis 4. According to Healey (2012), bivariate statistical tools are used when one is interested in examining the relationship between two variables while multivariate statistical tool is appropriate when one is interested to investigate the relationship among three or more variables. Specific decisions on statistical tools considered for the analysis of demographic data, research question 1, research question 2, research question 3, research hypothesis 1, research hypothesis 2, research hypothesis 3, and research hypothesis 4 are as follows.

Demographic variables

1. Gender is a qualitative variable with two categories namely male and female, and falls under the nominal scale of measurement. Therefore, univariate statistical tools such frequencies and percentages were used to describe it.
2. ‘Subject you teach’ is a nominal variable with seven categories namely Business Management, Cost Accounting, Financial Accounting, Economics, English language, Core Mathematics, and Integrated Science. Therefore, the use of frequencies and percentages to analyse it was appropriate.

3. 'Class you teach' is specified into three categories namely Form 1, Form 2, and Form 3 and falls under the nominal scale of measurement. Therefore, results were presented using frequencies and percentages.
4. 'Your Highest Qualification' is also a nominal variable with four categories. Hence, frequencies and percentages were used to analyse it.
5. 'Number of Years of Teaching' is quantitative variable which falls under the ratio scale of measurement. However, for easy interpretation, it was categorised, and result described and interpreted using univariate statistical tools such as frequencies and percentages.

Research Question 1

Research question 1 sought to describe the multiple-choice test construction competencies of the teachers in assessing students' learning outcomes at the SHS level in the KSD. The participants were instructed to carefully read each of the statements and decide how the statement applies to them by ticking a box based on the following guide: Strongly Disagree (SD), Disagree (D), Agree (A), and Strongly Agree (SA).

The scoring of items based on the four-point Likert scale of measurement are strongly agree = 4, agree = 3, disagree = 2 and strongly disagree = 1. Respondents were asked to indicate their levels of agreement or disagreement with statements concerning their competencies in constructing multiple-choice tests. A composite score based on all the items constitutes their test construction competence which is an ability that reflects their knowledge, skills, and attitudes and values towards test construction. It helps to understand whether the teacher employed his or her multiple-choice test construction competencies to an appreciable level or not. From the nature of the research question, data obtained falls under the interval scale of measurement. As a result, mean and standard deviation were used in analysing the research question.

With respect to the use of mean, criterion score (CS) of 2.50 using item means were established to determine their level of agreement or disagreement towards test construction competencies. An item mean score of 2.50 (that is, [1+2+3+4]/4 = 2.50) or above indicates teachers' positive attitudes , while a mean below 2.50 indicates teachers' negative attitudes which is embedded in each indicator of how well they employ their competencies in constructing multiple-choice test. Per the instructions given in terms of responding to the 20 items, positive attitudes imply an appreciable level of competencies, while negative attitudes imply an unappreciable level of competencies.

Research Question 2

In addressing this research question, quantitative item analysis was conducted for each of the end-of-semester multiple-choice test items provided by the participants. After obtaining the difficulty and the discrimination indices for each set of items constructed by classroom teachers, the number of items that met both acceptable criteria for discrimination index and difficulty index were judged as good or acceptable items. Items that did not meet the set criteria were judged as poor or problem items. The number of good items and the number of problem items fall under the ratio scale of measurement. As a result, means, and standard deviations were used to analyse the research question.

Research Question 3

Qualitative item analysis approach (that is, the direct examination of the tests) was employed to report on the common format and constructional flaws associated with the tests by the use of checklist (**Appendix K**). In this study, 'common format and constructional flaws' is a categorical variable, therefore, univariate statistical tools such as frequency count was reported.

Research Hypothesis 1

There are two sets of data: data on test length and the number of good items that are both measured on the ratio scale. The nature of the research hypothesis required that the relationship between the two variables should be established. From the preliminary analysis on the use of the bivariate correlational test, the distribution of the data on each of the variables was normal, the variables were linearly related, and homoscedasticity was assumed for the variables. These assumptions which were met made it most appropriate to use the Pearson Product-Moment Correlation Coefficient (PPMCC) to find answer to the research hypothesis. According to Mayers (2013), Pearson Product-Moment Correlation is employed when one wants to represent the relationship between two variables that are both measured on at least interval scale. This statistical tool was used because there was the need to establish the association between the variables in the hypothesis, and moreover, the data collected on each variable is measured on the ratio scale.

Research Hypothesis 2

There are two sets of data: data on test length and the number of problem items that are both measured on the ratio scale. The research hypothesis required that the relationship between the two variables be established. From the preliminary analysis on the use of the bivariate parametric correlational test, the results showed that the distribution of the data on each of the variables was normal, the variables were linearly related, and homoscedasticity was assumed for the variables. These assumptions were not violated. Therefore, the most appropriate statistical tool used to find answer to the research hypothesis is the Pearson Product-Moment Correlation Coefficient. Mayers (2013) has indicated that Pearson Product-Moment Correlation is employed when one wants to represent the relationship between two variables that are both measured on at least an interval scale. This statistical tool was used because there was the need to establish the relationship between the variables in the hypothesis, and moreover, the data on each variable is under the ratio scale of measurement.

Research Hypothesis 3

There are two sets of data: data on teachers' multiple-choice test construction competencies and the quality of the multiple-choice items that were measured on interval and ratio scales respectively. The nature of the research hypothesis required that the relationship between the two variables be established. From the preliminary analysis on the use of the bivariate correlational test, the distribution of the data on each of the variables was normal, the variables were linearly related, and homoscedasticity was assumed for the variables. These assumptions which were met made it most appropriate to use the Pearson Product-Moment Correlation Coefficient to find answer to the research hypothesis. Pearson Product-Moment Correlation is employed when one wants to represent the relationship between two variables that are measured on interval and ratio scales respectively (Mayers, 2013). This statistical tool was used because there was the need to establish the relationship between the variables in the hypothesis; moreover, the data collected on each variable falls under either interval or ratio scale of measurement.

Research Hypothesis 4

There are three sets of data: data on teachers' test construction competencies, quality of multiple-choice items, and years of teaching. The research hypothesis required that the influence of years of teaching on the relationship (if any) between teachers' test construction competencies and quality of the multiple-choice items be established. From the preliminary analysis on the assumptions of partial correlational test, the distribution of the data on each of the variables was normal, and homoscedasticity was assumed for the variables. Furthermore, 'teachers' test construction competencies' was linearly related to quality of multiple-choice items. 'Years of teaching' was also related to quality of multiple-choice items. However, the assumption of a linear relationship between teachers' test construction competencies and years of teaching was violated –there was zero correlation between the variables.

The violation of the one of the assumptions for partial correlation, meant that variables that demonstrated linear relationships and fulfilled the assumptions of normality and homoscedasticity should rather be investigated using Pearson Product-Moment Correlation Coefficient as the most appropriate statistical tool. Accordingly, there was the need to investigate the following:

1. The relationship between teachers' multiple-choice test construction competencies and the quality of multiple-choice items.

2. The relationship between years of teaching and the quality of multiple-choice items.

Nevertheless, the relationship between teachers' multiple-choice test construction competencies and the quality of multiple-choice items had already been established in research hypothesis 3. Therefore, it was quintessential to also explore the hypothesis that there is significant relationship between years of teaching and the quality of multiple-choice items using PPMCC. Mayers (2013) points out that the Pearson Product-Moment Correlation Coefficient (r) is used when one wants to represent the relationship between two variables that are both measured on at least an interval scale. This statistical tool was used because there was the need to establish the relationship between the variables in the hypothesis, and the data collected on each variable falls under either interval or ratio scale of measurement. The result with respect to this hypothesis is presented in Table 29 under the Chapter Four to this study.

Chapter Summary

In order to examine the relationship among the variables under investigation, the quantitative approach was employed using correlational research design. It must be emphasised that studies that involve examining correlation between or among variables do not, in and of themselves, establish cause and effect (Fraenkel et al., 2012).

Questionnaire and document examination were used as the main data collection instruments. A 20-item instrument titled: Teachers' Multiple-Choice Test Construction Competence Questionnaire (TTCCQ-MC) was used to measure teachers' multiple-choice test construction competencies. Benefits associated with the use of questionnaire is that its administration is easy and it takes relatively less time for participants to provide their responses to the set of items (Osuola, 2001). Nevertheless, Ary et al. (2010) have stated some likely problems that may affect the validity of a questionnaire. The following are some of the problems: (a) Most of the times, participants report what they believe or perceive is true but is not; (b) Participants may provide untrue responses that are more socially acceptable than what is happening in reality; and (c) Participants may give answers that they perceive the investigator wants to hear. The document examination covered students' responses on multiple-choice test items for end-of-semester administered teacher-made tests, copies of their marking schemes and end-of-semester teacher-made test.

The total number of teachers that constituted the population for the study was 157. However, with the use of purposive sampling technique, the study covered only 47 participants (n = 47) out of the 157 teachers. Therefore, the conclusions which was based on the relatively small sample of teachers do not present holistic view of the test construction competencies of the entire population of teachers considered for the study.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

This chapter presents the analyses of the data gathered from the field in relation to teachers' test construction competencies. The study adopted the correlational research design. In order to collect data for analysis, find answers to the research questions and test the hypotheses, questionnaires were distributed to the 47 research participants (21.30% (10) from MPASS, 19.15% (9) from KRISTEC, 27.66% (13) from BESCO and 31.92% (15) from PASCO). There was 100% return rate of the questionnaire. The results based on data collected are presented in two sections (Section A and B). 'Section A' deals with demographic information of the participants. 'Section B', on the other hand, is made up of the results of the main data and discussion of findings pertaining to the research questions and hypotheses driving this study.

Section A: Analysis of the Demographic Variables

Data gathered on the respondents' characteristics covered their gender, subject areas they teach, classes they teach, their highest educational qualification, and the number of years they have taught the SHS level. To provide answers to the demographic variables as specified on the questionnaire, the participants were instructed to tick [√] the appropriate response where required. The results are presented in Tables 8, 9, 10, 11, and 12.

Gender of Research Participants

Result on gender of respondents is presented in Table 8.

Table 8 –*Gender of Respondents*

Categories	Frequency	Percent
Male	43	91.49
Female	4	8.51
Total	47	100.00

Source: Field data (2019)

From Table 8, it can be observed that 91.49% of individuals in the sample are males and 8.51% females. This means that the number of males who participated in the study was more than females.

Subject Areas that Research Participants Teach

Result on subject area is presented in Table 9.

Table 9 –*Subject Area*

Categories	Frequency	Percent
Business management	4	8.51
Core mathematics	8	17.02
Cost Accounting	2	4.25
Economics	9	19.15
English Language	9	19.15
Financial Accounting	5	10.64
Integrated Science	10	21.28
Total	47	100.00

Source: Field data (2019)

From Table 9, 21.28% of the teachers teach Integrated Science and 4.25% were Cost Accounting teachers. Per the sample data, minority of the participants were Cost Accounting teachers and majority of them were Integrated Science teachers. Cumulatively, more Core subject teachers (17.02% + 19.15% + 21.28% = 57.45%) participated in the study as compared to the Business subject teachers (8.51% + 4.25% + 19.15% + 10.64% = 42.55%).

Classes that the Participants Teach

Result on the classes that the participants teach is illustrated in Table 10.

Table 10 –*Classes that Teachers Teach*

Categories	Frequency	Percent
Form 1	22	46.81
Form 2	7	14.89
Form 3	18	38.30
Total	47	100.00

Source: Field data (2019)

The data in Table 10 indicates that a frequency count of 22 representing 46.81% of the research participants were Form 1 teachers, 14.89% representing Form 2 teachers. This means that majority of the teachers were Form 1 teachers, and minority of them were Form 2 teachers.

Teacher Highest Qualification

Result on highest qualification of teachers is presented in Table 11.

Table 11 –*Teacher Qualification (Highest Qualification of Teachers)*

Categories	Frequency	Percent	Cumulative Percent
Master of philosophy	2	4.26	4.26
Master of education	2	4.26	8.52
First degree with education	32	68.08	76.60
First degree without education	11	23.40	100.00
Total	47	100.00	

Source: Field data (2019)

From Table 11, it can be seen that 68.08% had first degree with education, 23.40% had first degree without education, 4.26% of the participants had master of philosophy, and 4.26% had completed master of education programme. It is evident from the result that most of the participants were first degree holders with background in education. In the pursuance of first degree, master of education, and master of philosophy, one is introduced to courses related to educational assessment of students' learning outcomes. Therefore, from the cumulative percent, most of the participants (76.60%) should possess basic competence in assessment of students.

Years of Teaching

Data collected on years of teaching falls under the ratio scale of measurement. However, for the use of frequencies in reporting, it has been categorised for easy interpretation. Descriptive statistics for years of teaching are presented in Table 12.

Table 12 –*Descriptive Statistics on Teachers' Years of Teaching*

Categories	Frequency	Percent	Cumulative Percent
1 to 6 years	21	44.68	44.68
7 to 12 years	16	34.04	78.72
13 to 18 years	8	17.02	95.74
19 to 24 years	2	4.26	100.00
Total	47	100.00	

Source: Field data (2019)

As shown in Table 12, 21 representing 44.68% of the respondents had taught for 1 to 6 years, while 2 (4.26%) of the respondents had taught for 19 to 24 years. This means that most of the teachers had 1 to 5 years of teaching as opposed to 19 to 24 years. Further interpretation showed that cumulatively, majority of the participants had taught for 1 to 12 years (78.72%) while few had taught for 13 to 24 years (17.02% + 4.26% = 21.28%).

Section B: Analysis of the Main Data

Research Question 1: What multiple-choice test construction competencies do teachers have in assessing students learning outcomes at the senior high school level in the Kwahu-South District?

Research question 1 sought to help describe multiple-choice test construction competencies of teachers in assessing students' learning outcomes at the SHS level in the KSD. Table 13 discloses result based on analysed data for research question 1.

Table 13 –*Result on Teachers' Multiple-Choice Test Construction Competencies*

Q/N	Items	N	Mean (M)	Std. Deviation (SD)
7	number the test items one after the other	47	3.72	.45
5	give specific instructions on the test	47	3.72	.45
1	properly space the test items for easy reading	47	3.62	.49
11	give appropriate time for completion of test	47	3.62	.49
2	review test items for construction errors	47	3.60	.50
4	use appropriate number of test items	47	3.51	.51
8	make sure each item deals with an important aspect of content area	47	3.51	.51
13	include questions of varying difficulty	47	3.43	.54
10	prepare marking scheme while constructing the items	47	3.40	.61
6	appropriately assign page numbers to the test	47	3.34	.56
14	match items to vocabulary level of the students	47	3.32	.59
9	match test items to instructional objectives (intended outcomes of the appropriate difficulty level)	47	3.30	.55
12	pose clear and unambiguous items	47	3.21	.78
18	make options independent of each other	47	3.19	.58
20	make the options grammatically consistent with the stem	47	3.19	.61
19	avoid the use of "none of the above" as an option when an item is of the best answer type	47	3.13	.77
3	keep all parts of an item (stem and its options) on the same page	47	3.04	.66
17	present options in some logical order (e.g., chronological, most to least, alphabetical) when possible	47	2.87	.77
16	include in the stem any word(s) that might otherwise be repeated in each option	47	2.68	.73
15	make options approximately equal in length	47	2.64	.87
	MM / MSD		3.30	.60

Source: Field data (2019)

Key: MM = Mean of means; MSD = Mean of standard (std.) deviations

As it could be seen from Table 13, the overall mean (MM = 3.30, MSD = .60) was greater than the cut-off mean (M = 2.50). Moreover, the mean of standard deviations, and standard deviations ranging from .45 to .87 imply that the responses of the teachers on each item are very close to each other. Thus, the result gave ample evidence to believe that generally most of the participants possessed an appreciable level of competencies in constructing multiple-choice items in the KSD. In other words, most of the research participants possessed an appreciable level of competencies in achieving content validity, handling items' alternatives and assembling test items.

In terms of competence in achieving content validity, the items (item 8: make sure each item deals with an important aspect of content area, item 9: match test items to instructional objectives, item 10: prepare marking scheme while constructing the items, item 11: give appropriate time for completion of test, item 12: pose clear and unambiguous items, item 13: include questions of varying difficulty, and item 14: match items to vocabulary level of the students) have mean values (ranging from 3.21 to 3.62) greater than the criterion score (CS) 2.50. This implies that the teachers are able to employ their competencies, to an appreciable level, in measuring what students have achieved in relation to instructional objectives.

In relation to the competence in handling items' alternatives, the items (item 15: make options approximately equal in length, item 16: include in the stem any word(s) that might otherwise be repeated in each option, item 17: present options in some logical order (for example, chronological, most to least, alphabetical) when possible, item 18: make options independent of each other, item 19: avoid the use of "none of the above" as an option when an item is of the best answer type, and item 20: make the options grammatically consistent with the stem) have mean values (2.64 to 3.19) above the CS of 2.50. This could be interpreted as, to an appreciable level, the teachers are able to craft items' alternatives that discriminate students with respect to a particular content achieved at the end of instructional session(s).

Concerning competence in assembling the test, the items (item 1: properly space the test items for easy reading, item 2: review test items for construction errors, item 3: keep all parts of an item (stem and its options) on the same page, item 4: use appropriate number of test items, item 5: give specific instructions on the test, item 6: appropriately assign page numbers to the test, item 7: number the test items one after the other) have mean values (ranging from 3.04 to 3.72) that are above the CS of 2.50. This implies that the teachers, to an appreciable level, are able to organise test items and format the test to permit students to demonstrate their ability on a given content.

To throw further light on the interpretation of teachers' test construction competencies, the components as identified have been ranked based on their respective mean of means. The result is presented in Table 14.

Table 14 –*Ranks of Teachers' Multiple-choice Test Construction Competencies*

Component	N	Mean of means(MM)	Mean of Std. Deviation(MSD)	Ranks(R)
Competence in assembling test items	7	3.51	.24	1st
Competence in achieving Content validity	7	3.40	.14	2nd
Competence in handling items' Alternatives	6	2.95	.25	3rd

Source: Field data (2019)

From Table 14, it can be said that the research participants found it very easy to exhibit competence in assembling test items (MM = 3.51, MSD =

.24, R = 1st), and easy to demonstrate competence in achieving content validity (MM = 3.40, MSD = .14, R = 2nd). However, they found it quite difficult to demonstrate competence in handling the items' alternative (MM = 2.95, MSD = .25, R = 3rd).

Research Question 2: What are the characteristics of the multiple-choice test items based on the following criteria: difficulty index, and discrimination index in the Kwahu-South District?

Research Question 2 sought to establish the characteristics of the multiple-choice items based on the following criteria: difficulty index, and discrimination index for assessing the characteristics of the items. Based on quantitative items analysis statistics, items that met both acceptable criteria for discrimination index and difficulty index were judged as good items. Items that did not meet the set criteria were judged as problem items. Result on the characteristics of the multiple-choice items developed by the research participants is presented in Table 15.

Table 15 –*Characteristics of the Multiple-Choice Items Developed by the Research Participants*

Description	Items constructed by the teachers	Valid Items for Item analysis	Problem Items	Good Items
Sum (total)	2325.00	2306.00	1107.00	1199.00
Mean	49.47	49.06	23.55	25.51
Std. Deviation	14.15	14.14	8.98	8.51

Source: Field data (2019)

From Table 15, out of the total number of 2325 items, 2306 were deemed valid for item analysis. This means that 19 items were excluded from items analysis. With respect to the set criteria for assessing the characteristics of the items, out of the total 2306 items, 1199 items were described as good items. Out of the total, 1107 items were identified as problem items. This means that most of the test items constructed by the research participants are described as good items per their respective difficulty and discrimination indices. However, further analysis using ‘MedCalc’s Comparison of means calculator’ (see **Appendix J**) suggests that the average value for number of good items produced by the classroom teachers (M = 25.51, SD = 8.51) was not statistically greater than the average value for number of problem items produced (M = 23.55, SD = 8.98), t (92) = 0.03, p = .28, 2-tailed. Accordingly, it can be said that with respect to test characteristics, in general, the test items for assessing students’ achievement lacked a suitable level of psychometric properties. This is attributable to the fact that the total number of good items produced by the teachers was not statistically different from the total number of problem items.

Table 16 presents result on problem items based on unacceptable difficulty indices which are less than .30, difficulty indices which are greater than .70 and discrimination indices which are less than or equal to zero.

Table 16 –*Summary on Items Based on Unacceptable Difficulty and Discrimination Indices*

Description	Sum	Mean	Std. Deviation
Difficulty indices less than .30	664	14.13	6.98
Difficulty Indices greater than .70	295	6.28	4.56
Discrimination indices less than or equal to .00	395	8.40	5.74

Source: Field data (2019)

As indicated in Table 16, out of the total number of 2306 valid items (see Table 15) for item analysis, 664 had difficulty indices less than .30 (difficult items), 295 had difficulty indices greater than .70 (easy items). This means that most of the items were difficult.

Further, in sum, the unacceptable number of items according to Allen and Yen's (1979) item evaluation criteria for item difficulty is 959 (that is, 664 + 295). On the other hand, out of the 2306 valid items, 395 items had unacceptable discrimination indices less than or equal to zero based on Kubiszyn and Borich (2013) recommendation that one can seriously consider any item with a positive discrimination index for norm-referenced test(s). This means that most of the items had unacceptable difficulty indices as compared to the discrimination indices.

Research Question 3: What are the types of error associated with teacher-made multiple-choice tests senior high school teachers in the Kwahu-South District construct?

The literature reviewed on the use of quantitative item analysis in assessing items' characteristics revealed that the presence of problem items calls for qualitative evaluation of the multiple-choice test. Thus, Research Question 3 was established to help identify multiple-choice format and item constructional errors associated with teacher-made multiple-choice tests in the KSD. In addressing this research question, the participants' end-of-semester administered Business Management (BM) and Core Mathematics (CM) multiple-choice tests were assessed for errors using the 'Multiple-Choice Test Error Analysis Checklist' (**Appendix K**). In all, 12 achievement tests (BM, 4; CM, 8) were qualitatively examined. The result is presented in Table 17.

Table 17 –*Format and Constructional Errors Identified with the Business Management and Core Mathematics Tests*

Type of errors	BM	CM	Total
Test format errors	Freq.	Freq.	Freq.
Alternatives not presented in some logical order	4/4	3/8	7/12
Detectable pattern of correct answer	3/4	8/8	11/12
Horizontal arrangement of options	3/4	4/8	7/12
Options of items appearing in different columns/pages	3/4	5/8	8/12
Page numbers not assigned	4/4	6/8	10/12
Poor arrangement of items/spacing of test items	2/4	4/8	6/12
Use of font size difficult to see and read	0/4	6/8	6/12
Item construction errors	Freq.	Freq.	Freq.
Ambiguous items/More than one correct answer	2/4	5/8	7/12
Central theme, task or problem not presented in the stem	3/4	0/8	3/12
Clues to the correct answer	4/4	5/8	9/12
Heterogeneous options	2/4	0/8	2/12
Grammatical, punctuation, and spelling	4/4	0/8	4/12
Implausible distractors	2/4	6/8	8/12
Instructional related issues (No/ Incomplete instruction)	4/4	5/8	9/12
Cluing and linking items	1/4	0/8	1/12
No answer	1/4`	3/8	4/12
Wrong key to item	1/4	4/8	5/12
Not emphasising (e.g. bolding, underlying or capitalising) negative word in the stem	3/4	0/8	3/12
Time for completion of items not indicated on the test	4/4	5/8	9/12
Wrong answer	1/4	4/8	5/12

Source: Field data (2019)
Key: Freq. = Frequency; / = out of

As it can be seen from Table 17, with specific reference to format errors, 11 out of 12 tests (BM, 3 out of 4; CM, 8 out of 8) were identified to have detectable pattern of correct answers. Also, 6 out of the 12 tests were observed with items with font size which some of the students could find it more difficulty to see and read (BM, 0 out of 4; CM, 6 out of 8). Therefore, it could be said that most of the tests were identified with the problem of detectable pattern of correct answer as compared to the use of font size that students could find difficult to see and read.

In order to examine constructional flaws associated with the tests, problem items were qualitatively examined. From Table 17, each of the following errors was observed with the problem items across 9 out of the 12 tests: (a) clues to the correct answer, (b) instruction related issues (no and/or incomplete instruction), and (c) time for completion of items not indicated on the test. These observed errors are followed by other errors such as the use of implausible distractors (that is, 8 out of 12 tests) and ambiguous items/more than one correct answer (that is, 7 out of 12 tests). On the contrary, 1 out of the 12 tests was identified with cluing and linking items (that is, BM, 1 out of 4; CM, 0 out of 8). Thus, the result suggests that most of the tests examined with reference to constructional errors associated with problem items had the following issues: (a) clues to the correct answer, (b) instruction related issues (no or incomplete instruction), (c) time for completion of items not indicated on the test, (d) implausible distractors, and (e) ambiguous items/more than one correct answer as opposed to cluing and linking items.

Preliminary Analysis in Determining the Appropriate Statistical Tool to Test for Research Hypotheses 1 and 2

The scale of measurement for test length, number of good items, and number of problem items falls under parametric level of measurement; therefore, there was the need to examine the assumptions for bivariate parametric correlation test. However, the use of the statistical tool based on meeting the assumption that the variables were both measured on the ratio scale of measurement does not make it appropriate for good statistical results. Consequently, it was imperative to investigate other assumptions that underlie the use of parametric test of correlation between two variables: (a) the variables should be normally distributed, (b) linear relationship should exist between the variables, (c) no significant outliers for each of the variables, and (d) the variables should exhibit homoscedasticity.

Test of normality for the variables

Results on normality test for test length, number of good items and number of problem items are presented in Table 18.

Table 18 – *Tests of Normality for Test Length, Number of Good Items and Number of Problem Items*

	Kolmogorov-Smirnov			Shapiro-Wilk		
Variables	Statistic	df	Sig.	Statistic	df	Sig.
Test Length	.282	47	.000	.845	47	.000
Good Items	.128	47	.051	.944	47	.025
Problem Items	.144	47	.016	.923	47	.004

Source: Field data (2019)

From Table 18, Shapiro-Wilk test was significant for test length (W(47) = .845, p < .001), number of good items (W(47) = .944, p = .025), and number of problem items (W(47) = .923, p = .004). This means that the data set for each of the variables was not normally distributed. This was due to the degree of skewness and kurtosis observed with each of the variables. When a given data is skewed, it suggests the presence of outliers or extreme values at the higher or lower end of a given distribution. Statistics on the skewness and kurtosis associated with each of the variable are presented in Table 19.

Table 19 –*Skewness and Kurtosis for Test length, Number of Good item and Number of Problem Items*

	Skewness		Kurtosis	
Variables	Statistic	Std. Error	Statistic	Std. Error
Test Length	.936	.347	.918	.681
Number of Good Items	.818	.347	.576	.681
Number of Problem Items	1.066	.347	1.080	.681

Source: Field data (2019)

As it could be seen from Table 19 that the distribution for test length is positively skewed likewise the number of good items and the number of problem items. Moreover, the kurtosis values for each of variables means that there was the presence of peaked distribution, with very little variation in the data. One of the ways of achieving reasonable normality by dealing with skewness and the impact of outliers on statistical outcomes is transforming the data. Therefore, all the variables were transformed using square root transformation. After transforming the variables, test of normality was run and the results are presented in Table 20.

Table 20 –*Tests of Normality Based on Square Root Transformation of Test Length, Number of Good Items and Number of Problem Items*

	Kolmogorov-Smirnov			Shapiro-Wilk		
Transformed variables	Statistic	df	Sig.	Statistic	df	Sig.
Test Length	.254	47	.000	.877	47	.000
Number of Good Items	.101	47	.200	.975	47	.404
Number of Problem Items	.122	47	.076	.966	47	.192

Source: Field data (2019)

From Table 20, Shapiro-Wilk test of normality for the number of good items (W(47) = .975, p = .404) and the number of problem items (W(47) = .966, p = .192) appears to be not significant. This means that the independent data distributions for the number of good items and the number of problem items were assumed normal. However, the test of normality based on Shapiro-Wilk test was significant for test length (W(47) = .877, p < .001). This means that the distribution of test length was not normal based on the test. According to Coolican (as cited in Mayers, 2013), if the statistics for skewness and kurtosis for each of factor is divided by the standard error, it also helps to determine whether a given data set is normally distributed or not. The data is normally distributed when z-cores for skew and kurtosis are within ± 1.96. Table 21 presents descriptive information on the z-scores for skew and kurtosis for test length. From Table 21, examination of the z-scores for skew (1.29) and kurtosis (1.31) for test length depicts that test length was assumed to be normal.

Table 21 –*Z-Scores for Skew and Kurtosis for Square Root-Transformed Test Length Values*

	Skewness		Kurtosis	
Description	Statistic	Std. Error	Statistic	Std. Error
Test Length	.45	.35	.89	.68
z-scores(skew/kurtosis)	1.29		1.31	

Source: Field data (2019)

Examining linear relationship between test length (SQRT_VI) and good items (SQRT_GI) (using transformed data)

The graphical relationship between test length and the number of good items is presented by Figure 4.

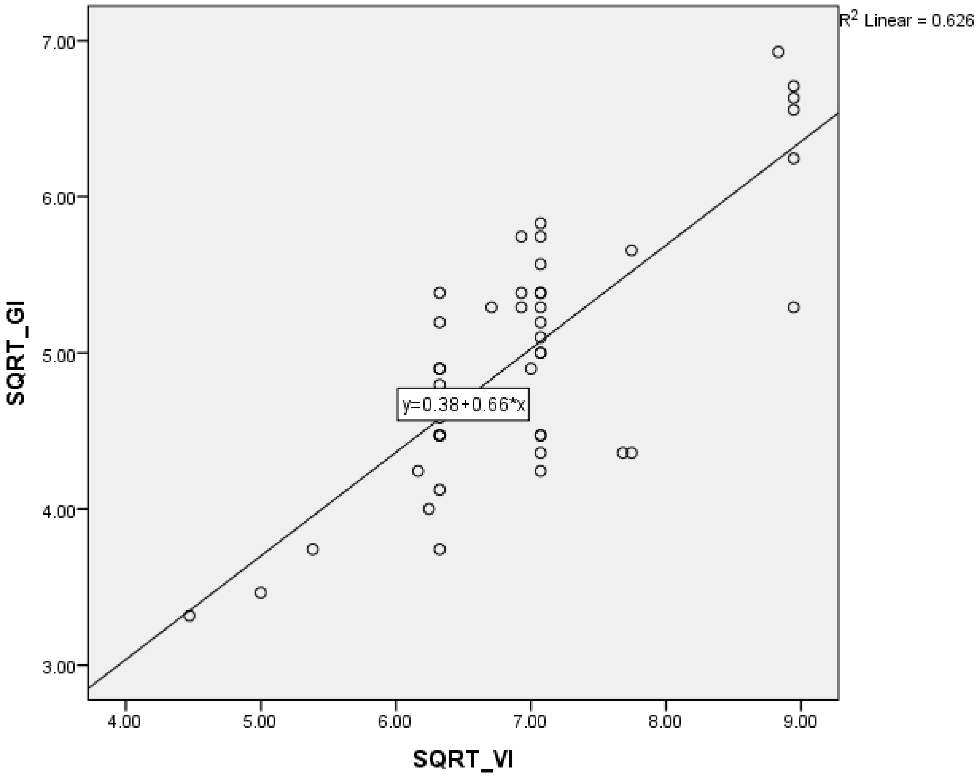


Figure 4 –Scatterplot Showing the Relationship between Test Length and the Number of Good Items

Source: Field data (2019)

Visual examination of Figure 4 depicts that there was positive linear relationship between test length and the number of good items (per the transformed data). This means that the assumption of linearity between the variables was not violated.

Test of linear relationship between test length (SQRT_VI) and problem items (SQRT_PI) (using transformed data)

Figure 5 presents pictorial information on the relationship between test length and the number of problem items.

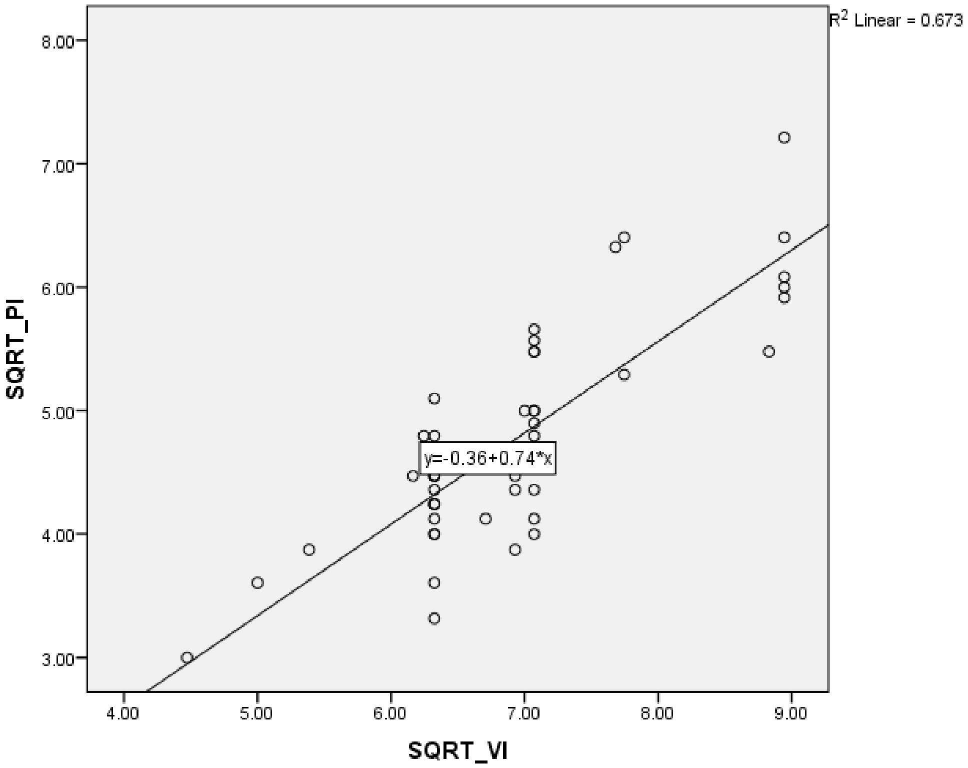


Figure 5 –Scatterplot Indicating the Relationship between Test Length and the Number of Problem Items

Source: Field data (2019)

From Figure 5, the relationship between test length and the number of problem items per the scatterplot was linear and positive in direction. Therefore, the assumption for linearity was met for test length and the number of problem items.

Examining the assumption of homoscedasticity for test length and the number of good items

Figure 6 presents result for examination of homoscedasticity for test length and the number of good items.

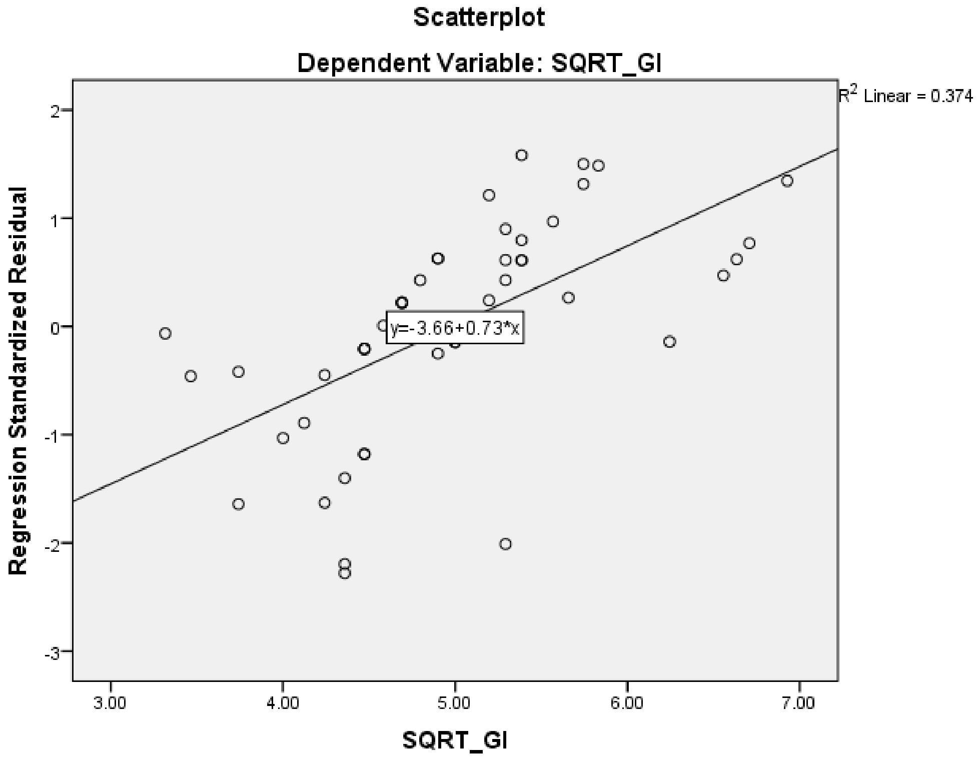


Figure 6 –Scatterplot for Examining the Assumption of Homoscedasticity for Test Length and the Number of Good Items

Source: Field data (2019)

Visual inspection of Figure 6 indicates that the assumption of homoscedasticity was not violated for the variables test length and the number of good items. That is, the variability of the residuals in the residuals of test length about the number of good items appeared similar at all levels of the number of good items.

Examining the assumption of homoscedasticity test length and the number of problem items

Figure 7 presents result on homoscedasticity for test length and the number of problem items.

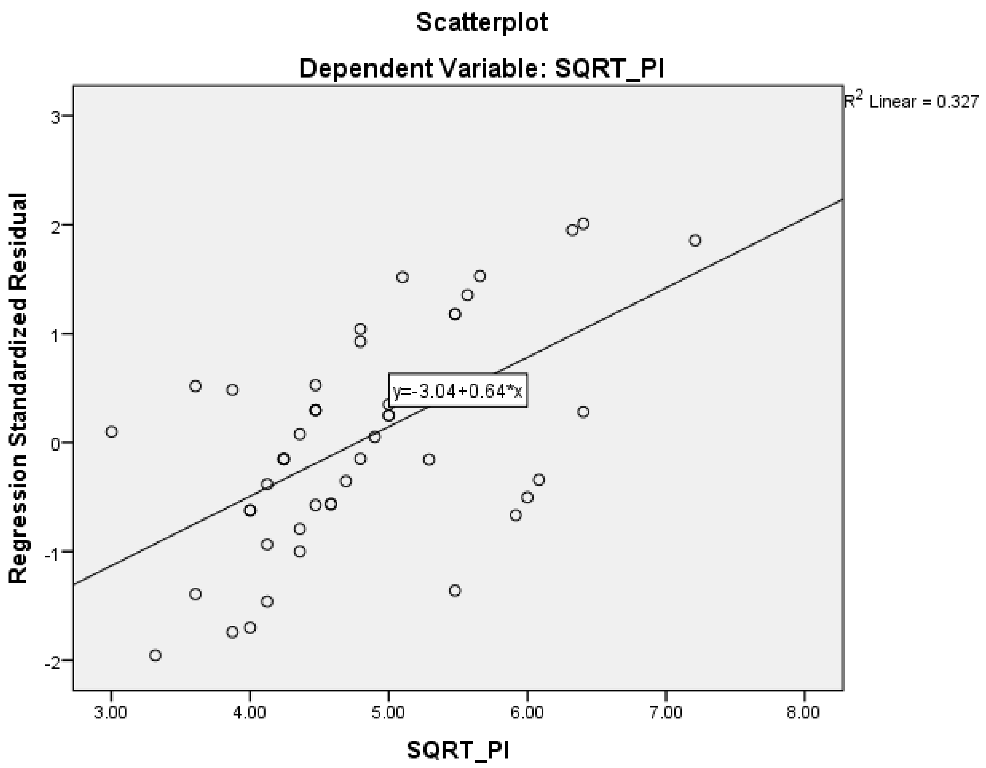


Figure 7 –Scatterplot for Examining the Assumption of Homoscedasticity for Test Length and the Number of Problem Items

Source: Field data (2019)

Visual inspection of Figure 7 shows that the assumption of homoscedasticity was not violated for the variables test length and the number of problem items. That is, the variability of the residuals in the residuals of test length about the number of problem items appeared similar at all levels of the number of problem items.

Based on the preliminary analysis on the use of bivariate correlational test, the distribution of the data on each of the variables was normal, the variables were linearly related, and homoscedasticity was assumed for the variables. These assumptions which had not been violated made it most appropriate to use Pearson Product-Moment Correlation Coefficient (r) to find answers to research hypotheses 1 and 2.

Research Hypothesis 1: There is no statistically significant relationship between test length and the number of good items produced by senior high school teachers in the Kwahu-South District.

This hypothesis sought to investigate if there was statistically significant relationship between test length and the number of good items produced by SHS teachers in the KSD. The result based on the use of PPMCC is presented in Table 22.

Table 22 –*Correlation between Test Length and the Number of Good Items*

Description				Number of good items
Test Length	Pearson Correlation			.791**
	Sig. (2-tailed)			.000
	N			47
	Bootstrap	Bias		-.008
		Std. Error		.073
		BCa 95% Confidence Interval	Lower	.632
			Upper	.891

Source: Field data (2019)

**. Correlation is significant at the .01 level (2-tailed).

Result from Table 22 shows that there was significant high positive relationship (r (45) = .79, p < .001, 2-tailed) between test length and the number of good items produced by the research participants. This is supported by bootstrap result based on 2000 bootstrap samples. That is, r = .79 with 95% Bias-corrected and accelerated (BCa) confidence interval [.632; .891] is significant. This implies that as the teachers increase their test length, the more likely they will produce more of good items. The coefficient of determination (r^2) for r of .79 is .62.

Research Hypothesis 2: There is no statistically significant relationship between test length and the number of problem items found in the multiple-choice test items constructed by senior high school teachers in the Kwahu-South District.

This hypothesis sought to investigate if there was statistically significant relationship between test length and the number of problem items found in the multiple-choice test items constructed by SHS teachers in the KSD. The result on the relationship between the variables is presented in Table 23.

Table 23 –*Correlation between Test Length and the Number of Problem Items*

	Description			Number of problem items
Test Length	Pearson Correlation			.820**
	Sig. (2-tailed)			.000
	N			47
	Bootstrap	Bias		-.002
		Std. Error		.043
		BCa 95% Confidence Interval	Lower	.716
			Upper	.887

Source: Field data (2019)
**. Correlation is significant at the .01 level (2-tailed).

Result from Table 23 shows that there was significant high positive linear relationship (r (45) = .82, p < .001, 2-tailed) between test length and the number of problem items produced by the research participants. This is supported by bootstrap result based on 2000 bootstrap samples. That is, r = .82 with 95% Bias-corrected and accelerated (BCa) confidence interval [.716; .887] is significant. This implies that as the teachers increase their test length, the more likely they will produce more of problem items. The r^2 for r of .82 is .67.

Preliminary Analysis in Determining the Appropriate Statistical Tool to Test for Research Hypotheses 3 and 4

The scale of measurement for teachers' test construction competencies, the quality of multiple-choice items (the proportion of good items) and years of teaching falls under parametric level of measurement. Consequently, assumptions for parametric test correlation were examined for the variables that constitute research hypothesis 3 and research hypothesis 4. The nature of research hypothesis 3 suggested the use of Pearson Product-Moment Correlation Coefficient while the nature of research hypothesis 4 called for the use of Pearson Partial Correlation technique.

Nevertheless, the use of the statistical tool based on meeting the assumption that the variables are measured on at least an interval scale does not make it a reliable measure for good statistical results. Subsequently, it was vital to investigate other assumptions that underlie the use of parametric correlational test which include: (a) the variables should be normally distributed, (b) linear relationship should exist between or among the variables, (c) no significant outliers for each of the variables, and (d) the variables should exhibit homoscedasticity.

Normality test for the variables

Table 24 contains results on normality test for years of teaching, teachers' test construction competencies, items' quality (in terms of proportion of good items).

Table 24 –*Normality Test for Years of Teaching, Teachers' Test Construction Competencies, Items' Quality (in Terms of Proportion of Good Items)*

	Kolmogorov-Smirnov			Shapiro-Wilk		
Variables	Statistic	df	Sig.	Statistic	df	Sig.
Years of Teaching	.132	47	.039	.931	47	.009
Teachers' Test Construction Competencies	.188	47	.000	.931	47	.008
Proportion of Good Items	.100	47	.200	.970	47	.271

Source: Field data (2019)

Table 24 shows that Shapiro-Wilk test was not significant for proportions of good items (W(47) = .970, p = .271), indicating that normal distribution was assumed for proportions of good items. However, the test of normality was significant for years of teaching (W(47) = .931, p = .009) and Teachers' test construction competencies (W(47) = .931, p = .008), meaning the distribution of data on years of teaching and teachers' test construction competencies were not normal. This was due to the degree of skewness and kurtosis observed with each of the variables. When a given data is skewed, it suggests the presence of outliers or extreme values at the higher or lower end of a given distribution. Statistics on the skewness and kurtosis associated with the variables (years of teaching and teachers' test construction competencies) are presented in Table 25.

Table 25 –*Skewness and Kurtosis Values for Years of Teaching and Teachers' Test Construction Competencies*

	Skewness		Kurtosis	
Variables	Statistic	Std. Error	Statistic	Std. Error
Years of Teaching	.75	.35	-.11	.68
Test Construction Competencies	.75	.35	-.17	.68

Source: Field data (2019)

As it could be seen from Table 25, the distribution for years of teaching and test construction competencies were positively skewed. One of the ways of achieving reasonable normality by dealing with skewness and the impact of outliers on statistical outcomes is transforming the data. Therefore, all the variables were transformed using square root transformation. After transforming the variables, test of normality was run and the results is presented in Table 26.

Table 26 –*Tests of Normality Based on Square Root Transformation of Years of Teaching and Test Construction Competencies*

	Kolmogorov-Smirnov			Shapiro-Wilk		
Transformed Variables	Statistic	df	Sig.	Statistic	df	Sig.
Years of Teaching	.083	47	.200	.964	47	.156
Test Construction Competencies	.185	47	.000	.937	47	.014

Source: Field data (2019)

From Table 26, Shapiro-Wilk test of normality was not significant for years of teaching (W(47) = .964, p = .156), but significant for test construction competencies (W(47) = .937, p = .014). This implies that normality is assumed for years of teaching. However, normality is not assumed for the variable, test construction competencies. Subsequently, there was the need to explore the z-scores of skew and kurtosis for the square root-transformed variable, test construction competencies. The result is presented in Table 27. From Table 27, examination of the z-scores for skew (1.96) and kurtosis (1.41) for test length depicts that normality was assumed for test construction competencies.

Table 27 –*Z-Scores of Skew and Kurtosis for Square Root-Transformed Variable (Test Construction Competencies-SQRTTC)*

	Skewness		Kurtosis	
Transformed Variable	Statistic	Std. Error	Statistic	Std. Error
SQRTTC	.68	.35	-.28	.68
z-scores(skew/kurtosis)	1.96		.41	

Source: Field data (2019)

Assumption of linearity between variables

Test construction competencies and proportions of good items

Figure 8 communicates the kind of relationship between test construction competencies (SQRTTC) and proportions of good items (SQRT_PGI). Visual examination of the scattergram (Figure 8) shows that there was linear relationship between test construction competencies and proportions of good items which in positive.

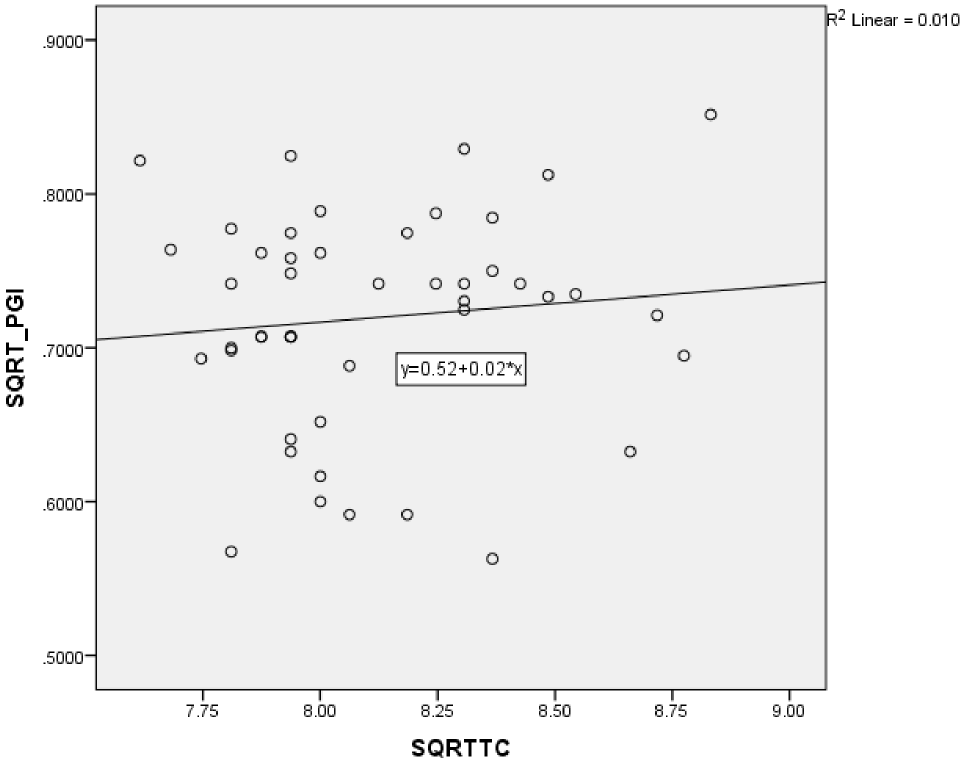


Figure 8 –Scattergram Communicating the Relationship between Test Construction Competencies and Proportions of Good Items

Source: Field data (2019)

Years of teaching and test construction competencies

Figure 9 indicates the kind of relationship between years of teaching (SQRT_EDYOT) and test construction competencies (SQRTTC). It is evident from Figure 9 that there was zero or no relationship between years of teaching and test construction competencies.

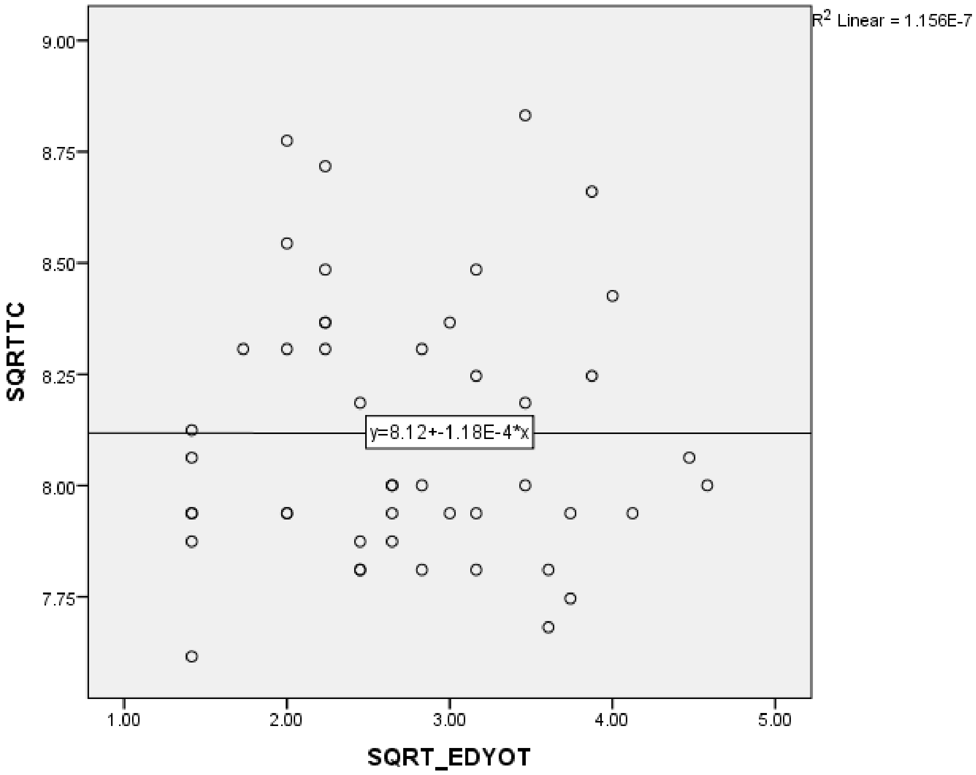


Figure 9 –Scattergram Showing the Kind of Relationship between Years of Teaching and Test Construction Competencies

Source: Field data (2019)

Years of teaching and proportions of good items

Figure 10 shows the relationship between years of teaching (SQRT_EDYOT) and proportions of good items (SQRT_PGI). The evidence from the scattergram (Figure 10) shows that there was negative linear relationship between years of teaching and proportions of good items.

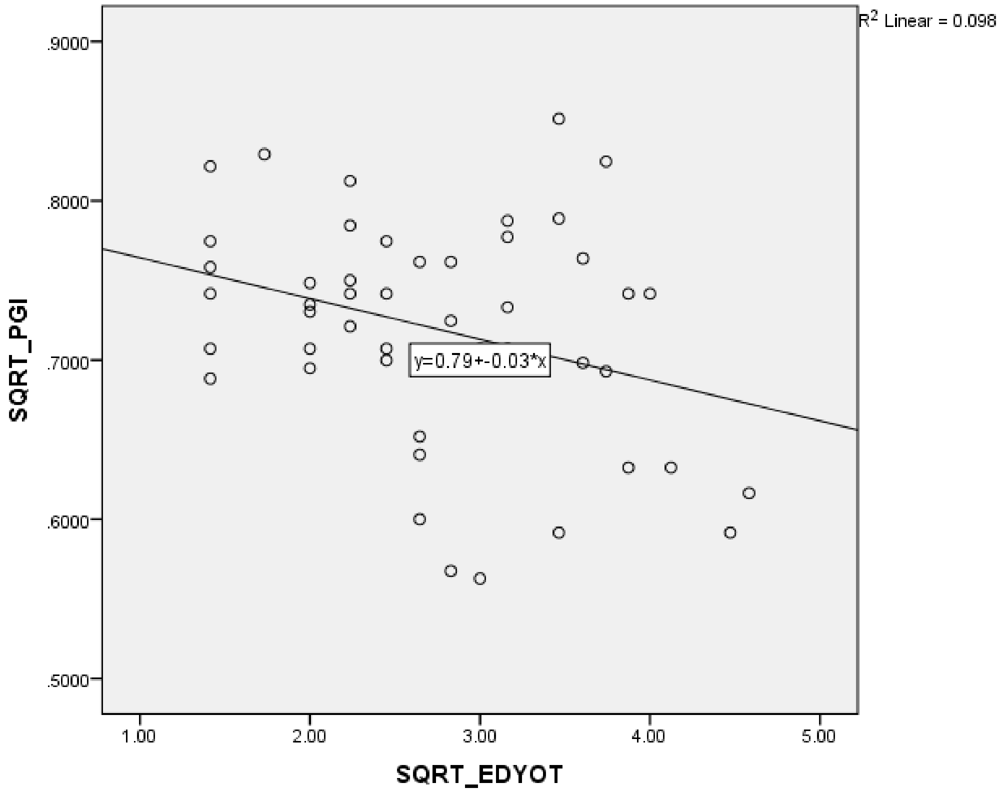


Figure 10 –Scattergram Showing the Pictorial Relationship between Years of Teaching and Proportions of Good Items

Source: Field data (2019)

Assumption of homoscedasticity between variables

Test construction competencies and proportions of good items

Scatterplot indicating homoscedasticity between test construction competencies (SQRTTC) and proportions of good items (SQRT_PGI) is presented in Figure 11.

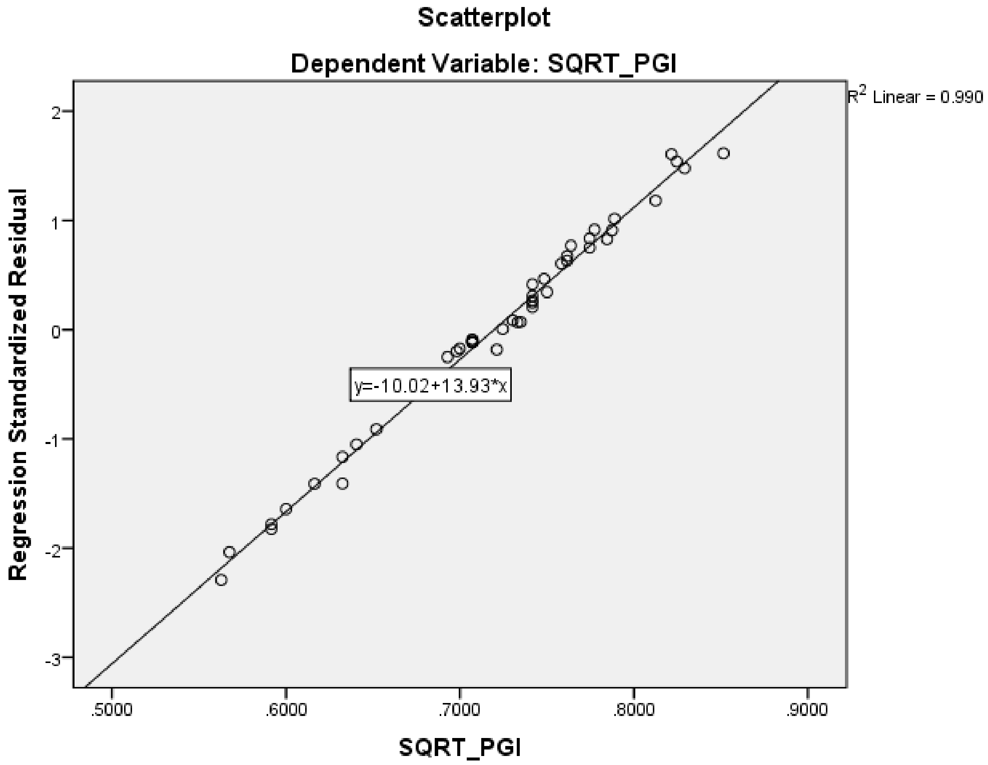


Figure 11 –Scatterplot Demonstrating Homoscedasticity between Test Construction Competencies and Proportions of Good Items

Source: Field data (2019)

From Figure 11, it can be observed that the assumption for homoscedasticity was not violated for the variables test construction competencies and proportions of good items. This means that the variability in scores for test construction competencies appeared constant at all levels of the proportions of good item.

Years of teaching and test construction competencies

Figure 12 is a graphical representation of homoscedasticity between years of teaching (SQRT_EDYOT) and test construction competencies (SQRTTC).

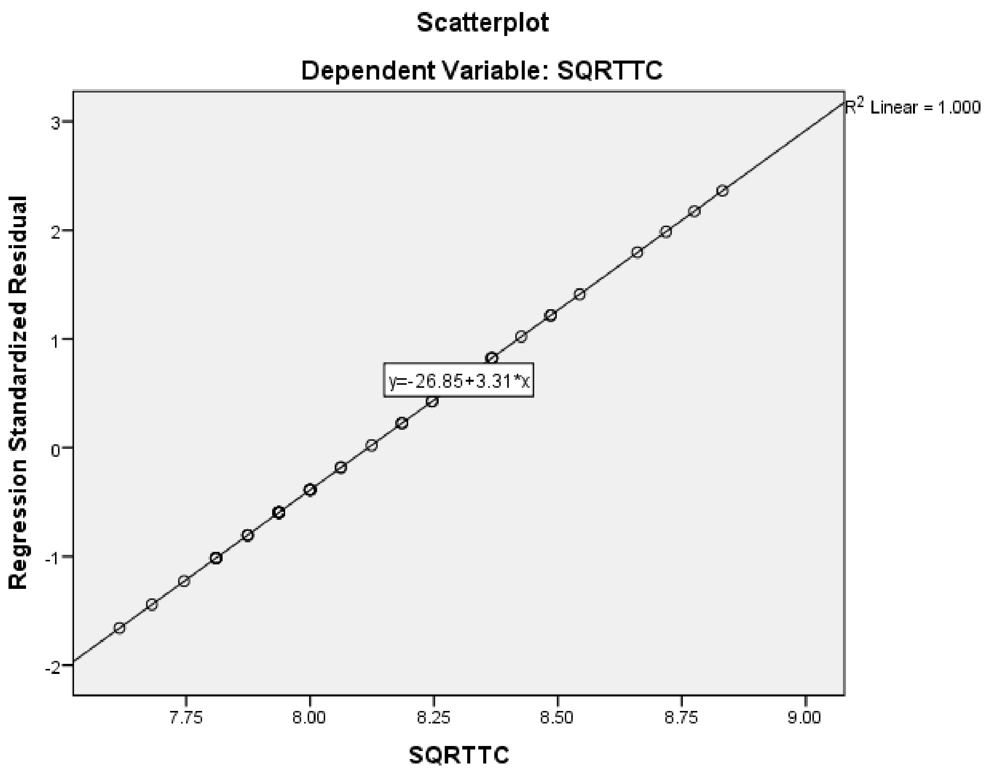


Figure 12 –Scatterplot Representing Homoscedasticity between Years of Teaching and Test Construction Competencies

Source: Field data (2019)

It can be observed from Figure 12 that the assumption for homoscedasticity was not violated for the variables years of teaching and test construction competencies. This means that the variability in years for years of teaching appeared constant at all levels of test construction competencies.

Years of teaching and proportions of good items

Figure 13 indicates homoscedasticity between Years of teaching (SQRT_EDYOT) and the proportions of good items (SQRT_PGI).

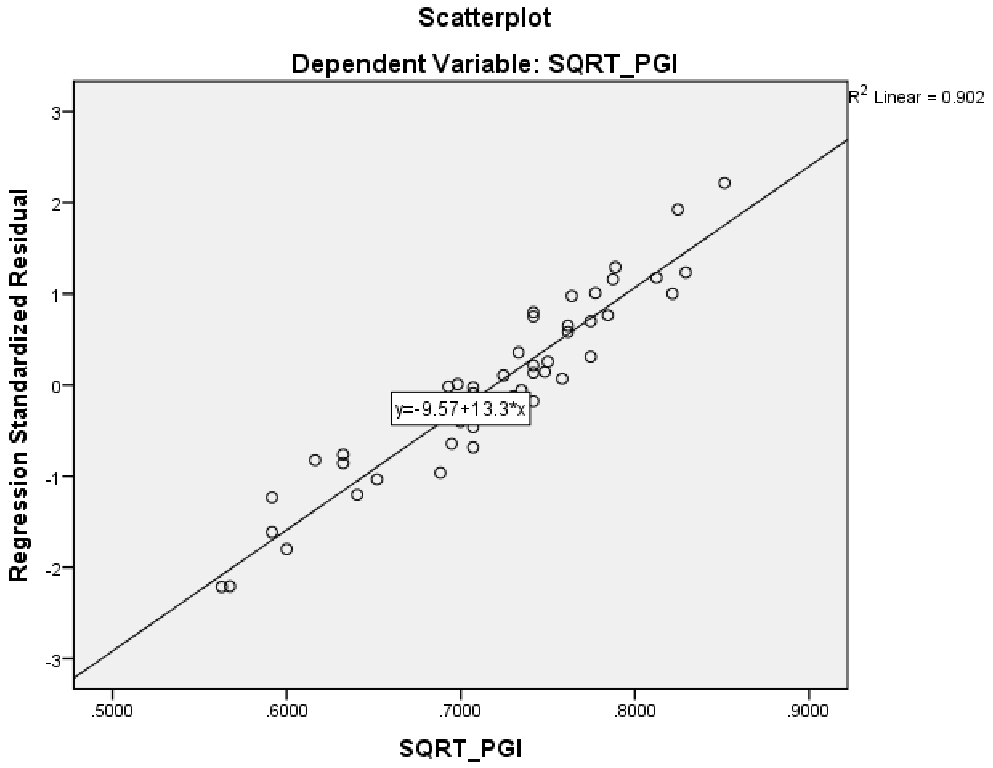


Figure 13 –Scatterplot Indicating Homoscedasticity between Years of Teaching and Proportions of Good Items

Source: Field data (2019)

From Figure 13, the assumption for homoscedasticity was not violated for the variables years of teaching and proportions of good items. This means that the variability in years for years of teaching appeared similar at all levels of the proportions of good items.

The nature of research hypothesis 3 required that the relationship between the two variables be established. From the preliminary analysis on the use of bivariate correlational test, the distribution of the data on each of the variables was normal, the variables were linearly related, and homoscedasticity was assumed for the variables. These assumptions which had been met made it most appropriate to use Pearson Product-Moment Correlation Coefficient (r) to find answer to research hypothesis 3.

Research Hypothesis 3: There is no statistically significant relationship between multiple-choice test construction competencies of teachers and the quality of multiple-choice test items in the Kwahu-South District.

Research hypothesis 3 sought to examine if there was any statistically significant relationship between multiple-choice test construction competencies of teachers and the quality of multiple-choice test items (proportions of good items) in the KSD. The result is presented in Table 28.

Table 28 –*Correlation between Multiple-Choice Test Construction Competencies and Proportions of Good Items*

	Description			Proportions of good items
Test Construction Competencies	Pearson Correlation			.102
	Sig. (2-tailed)			.496
	N			47
	Bootstrap[c]	Bias		-.001
		Std. Error		.151
		BCa 95% Confidence Interval	Lower	-.218
			Upper	.405

Source: Field data (2019)

c. Unless otherwise noted, bootstrap results are based on 2000 bootstrap samples

From Table 28, it can be seen that the relationship between test construction competencies and the proportions of good items (items' quality) is not significant. That is, r = .10 with 95% Bias-corrected and accelerated (BCa) confidence interval [-.218; .405] was not significant. Hence, high or low values in the items' quality was not related to high or low scores in the teachers' test construction competencies. By implication, there is no significant likelihood that a teacher with high level of test construction competencies will produce importantly more good items than problem items.

Research Hypothesis 4: Teachers' years of teaching has no statistically significant effect on the relationship between multiple-choice test construction competencies of teachers and the quality of multiple-choice test items in the Kwahu-South District.

The focus of research hypothesis 4 was to examine the effect of the teachers' years of teaching on relationship between their multiple-choice test construction competencies and the quality of multiple-choice test items they construct in the KSD. Variables were measured at least on an interval scale. To examine the effect of years of teaching on relationship between the variables of interest, there was the need to test for the assumptions of Partial Correlation. The assumption of linear relationship between teachers' construction competencies and years of teaching was violated. That is, zero relationship was observed, though the assumption of linearity was met for the relationship between years of teaching and the quality of multiple-choice test items. Nevertheless, there was the need to examine the relationship between years of teaching and the quality of multiple-choice test items (proportions of good items) using PPMCC. The result is presented in Table 29.

Table 29 –*Correlation between Years of Teaching and Proportions of Good Items*

	Description			Proportions of good items
Years of Teaching	Pearson Correlation			-.313*
	Sig. (2-tailed)			.032
	N			47
	Bootstrap	Bias		.002
		Std. Error		.131
		BCa 95% Confidence Interval	Lower	-.560
			Upper	-.037

Source: Field data (2019)
*. Correlation is significant at the .05 level (2-tailed).

Table 29 shows that the correlation coefficient obtained in examining the relationship between years of teaching and the quality of multiple-choice test items was significant (r (45) = -.31, p = .032). This is supported by bootstrap result based on 2000 bootstrap samples. That is, r = -.31 with 95% Bias-corrected and accelerated (BCa) confidence interval [-.560; -.037] was significant. This means that there was moderate negative linear relationship between years of teaching and the quality of multiple-choice test items. Therefore, as years of teaching increases, the items' quality of multiple-choice test decreases (or the rate of producing problem items increases against good items). The r^2 for r of -.31 is .10.

Discussion of Findings from the Results of the Study

This section discusses the finding(s) associated with each of the research questions and hypotheses in relation to published literature and empirical findings on test construction competencies and the quality of multiple-choice tests. Based on the objectives of the study, the thematic framework for discussion is outlined as follows.

1. Multiple-choice test construction competencies of teachers in assessing students learning outcomes
2. Characteristics of the multiple-choice tests (using quantitative and qualitative item analysis)
3. Multiple-choice test construction competencies and the quality of multiple-choice test items
4. Effect of teachers' years of teaching on multiple-choice test construction competencies of teachers and the quality of multiple-choice test items

Multiple-choice test construction competencies of teachers

Research question 1 helped to find answers to the competencies of teachers in constructing multiple-choice tests. The finding revealed that generally, most of the participants in the KSD possessed an appreciable level of competencies in constructing multiple-choice items. In other words, most of the research participants possessed competencies in achieving content validity, handling items' alternatives and assembling test items.

From the aforesaid, it implies that the teachers are able to (a) employ their ability, to an appreciable level, in measuring what students have achieved in relation to instructional objectives, (b) craft alternatives in order to discriminate students with respect to a particular content achieved at the end of instructional session(s), and (c) organise test items and format the test to permit students to demonstrate their ability on a given content. This observation can be associated with the fact that most of the participants were first degree holders with a background in education. According to Chau (as cited in Hamafyelto et al., 2015), teacher's test construction competence is directly related to ensuring the quality of a test. Consequently, these competencies possessed by the classroom teachers should help them to craft good multiple-choice tests for the assessment of students' learning outcomes.

Further exploration of multiple-choice test construction competencies indicated that the teachers found it very easy to exhibit competence in assembling test items, and easy to demonstrate competence in achieving content validity. However, they found it quite difficult to demonstrate competence in handling the items' alternatives. This means that the teachers were not all that good at demonstrating competencies in making options independent of each other, crafting options that are grammatically consistent with the stem, presenting options in some logical order (chronological, numerical or alphabetical), and making options approximately equal in length.

Burton, Sudweeks, Merrill and Wood (1991) have indicated that good multiple-choice test items are more demanding and take a lot of time to craft as compared to other types of test items. In addition, coming up with plausible distractors for the items requires a certain amount of skill. Nevertheless, this skill may be improved through experience, study, and practice. Rivera (2007) also believes that classroom teachers can master the writing of test items through practice. Therefore, classroom teachers should practically be exposed to item writing skills; especially, crafting options with good quality. According to Maba (2017) competence as ability is modifiable and new experiences can be integrated. Consequently, the new experiences gained by teachers as a result of exposure to constant practice in terms of ensuring the quality of items' options can lead to the integration and modification of their competencies in constructing multiple-choice tests.

Characteristics of the multiple-choice tests (using quantitative and qualitative item analysis)

The characteristics of teacher-made tests in the KSD was investigated as a means of finding answers to research question 2, research question 3, research hypothesis 1 and research hypothesis 2. For research question 2, finding indicated that most of the test items constructed by the research participants were described as good items per their respective difficulty and discrimination indices. However, further analysis indicated that the average value for number of good items constructed by the classroom teachers was not statistically greater than the average value for number of problem items.

Subsequently, it can be said that for test characteristics (in terms of difficulty and discrimination indices), generally, the multiple-choice test items used for assessing the students' achievement lacked a suitable level of psychometric properties because the total number of good items was not statistically different from the total number of problem items. This finding supports Agu et al. (2013) observation that, in Nigeria, the quality of classroom assessment tests lacks proper psychometric properties hence facing criticisms. According to Furr and Bacharach (2014), construction of test items to discriminate among those who have mastered a given content area from those who have not is the responsibility of the classroom teacher. Therefore, from Nitko's (2001) perspective, to improve reliability of assessment results, teachers are required to match assessment difficulty to students' ability levels.

Additional finding revealed by further examination of the problem items was that most of the items had unacceptable difficulty indices as compared to the discrimination indices. Thus, in assessing the number of good items, the impact of unacceptable difficulty levels was greater than that of unacceptable discrimination indices. With a careful inspection of the difficulty levels, most of the items were described as difficult (that is, had p-values < .30). This suggests that probably (a) the test items had issues of content validity, (b) the students had a poor understanding of the difficult topics that were treated, (c) there were some ambiguous items since students were choosing more of the distractors as compared to the correct answer (d) the classroom teachers did not have adequate knowledge about the characteristics of their students.

With item difficulty, from Nitko's (2001) point of view, teachers should ensure that the test they construct contains items that are not too difficult or too easy for their students. However, it could be observed that most of the items the teachers constructed were described as difficult for their students. Consequently, the quality of the assessment results used in grading the students is questionable. According to Amedahe and Asamoah-Gyimah (2016), reliability of an assessment is affected when test difficulty is not matched to the ability of the students involved. To minimise this effect, Nitko calls on classroom teachers to tailor test items to students' ability levels.

For the discrimination indices, in norm-referencing, items that do not differentiate among students, or produce negative discrimination indices should be discarded or avoided (Kubiszyn & Borich, 2013; Nitko, 2001). Furthermore, these items should not be considered in terms of the total number of items that make up students composite score in a given achievement test (Crocker & Algina, 2008). Yet, these items were used in assessing students learning outcomes and evaluating their performance.

According to Hambleton and Jones (1993) classical true-score theory items analysis procedures have the potential to provide invaluable information concerning constructional flaws such as implausible distractors, and double negatives. Therefore, informed by this assertion, research question 3 was established to identify other characteristics in terms of multiple-choice format and item constructional errors associated with teacher-made multiple-choice tests through qualitative evaluation of the tests (BM and CM).

Generally, findings on research question 3 revealed that most of the tests were identified with the problem of a detectable pattern of the correct answer as compared to the use of font size that students could find difficult to see and read. Moreover, most of the tests examined with reference to constructional errors associated with problem items had the following issues: (a) clues to the correct answer, (b) instruction related issues (no or incomplete instruction), (c) time for completion of items not indicated on the test, (d) implausible distractors, and (e) ambiguous items/more than one correct answer as opposed to cluing and linking items. The findings supports Amedahe's (1989) observations that test constructed by classroom teachers were not devoid of constructional flaws. Amedahe and Asamoah-Gyimah (2016), Joshua (2005), Kubiszyn and Borich (2013), Morrow et al. (2000) and Nitko (2001) have stated that the presence of format and constructional errors reduces the quality of assessment results.

Findings based on research questions 2 and 3 raise issues about the quality of the assessment results among SHSs in the KSD. Agu et al. (2013) have indicated that when teacher-made tests are low in quality, school administrators and teachers will not be able to make available support and educational opportunities that each student needs. In other words, lack of or low degree of validity of test results leads to undependable inferences about student learning based on which educational decisions such as promotion and selection of students for educational opportunities would be wrongfully made (Amedahe & Asamoah-Gyimah, 2016; Gareis & Grant, 2015).

Ali (1999), Ujah (2001) and Silker (2003) have emphasised that construction of good test items requires the use of skills through which classroom teachers can construct and design a test with accuracy, objective communication, correct use of language, items validation and the right choice of grading scales. Given the aforesaid, the classroom teachers should exhibit competencies in constructing multiple-choice test items that would help improve the quality of the assessment results. Simon (as cited in Ovat & Ofem, 2017) has stated that it is poor test construction that influences academic dishonesty, and examination malpractices among most secondary schools in Nigeria. Therefore, where the classroom teachers refuse to employ or apply high or appreciable levels of competencies in constructing test items, it will lead to poorly constructed multiple-choice tests which, in turn, would contribute to academic dishonesty and examination malpractices during the assessment of students' achievement.

Finding based on examination of research hypothesis 1 showed that there was a significant high positive relationship between test length and the number of good items produced by the research participants. Finding from research hypothesis 2 also gave contradicting evidence that there was a significant high positive linear relationship between test length and the number of problem items produced by the research participants. However, the respective correlation coefficients for research hypothesis 1(r = .79) and research hypothesis 2(r = .82) means that the probability of producing problem items was slightly greater than the probability of producing good items when increasing test length.

The findings in relation to research hypotheses 1 and 2 imply that in increasing test length, critical attention should be given to producing well-written multiple-choice test items, so that test length will have higher positive relationship with the number of good items and very low positive or higher negative relationship with problem items (that is where if there are problem items). This is because from the findings, in practical terms, just increasing test length does not automatically guarantee reliable and valid assessment results. According to Crocker and Algina (2008), increasing test length will improve assessment results only when the test constructor pays critical attention towards producing well-written items free from technical flaws. Besides, the items should have appropriate difficulty and discrimination indices (Allen & Yen, 2002; Crocker & Algina, 2008).

Multiple-choice test construction competencies and the quality of multiple-choice test items

The relationship between multiple-choice test construction competencies and the quality of the multiple-choice items was examined by research hypothesis 3. Finding on research hypothesis 3 gave the evidence that the relationship between test construction competencies and the quality of multiple-choice items constructed by the classroom teachers was not significant. That is, high or low scores of the teachers' test construction competencies were not significantly related to an increase or decrease in the items' quality of multiple-choice tests. In other words, high or low scores of the teacher's test construction competence was not significantly related to either producing importantly more good items than problem items, or crafting more problem items than good items.

This finding as presented contradicts Chau's (as cited in Hamafyelto et al., 2015) perspective that teacher's test construction competence is directly related to ensuring the quality of the test he or she constructs. If the teachers' test construction competencies were not significantly related to the items' quality of multiple-choice tests, then it appears that though they possessed appreciable or high levels of competencies, they paid inadequate attention to ensure the quality of the multiple-choice test items. Therefore, the problem items influenced the items' quality in a manner that it did not correlate to the teachers' self-report on their ability in constructing multiple-choice test items. As Kubiszyn and Borich (2013), Amedahe and Asamoah-Gyimah (2016) have indicated, problem items because of test-related factors such as the use of font size that students find difficult to read, unclear instructions, and ambiguous items, clues to correct answers are present, make assessment results less valid for relevant educational decisions concerning students and classroom teachers.

Effect of teachers' years of teaching on multiple-choice test construction competencies of teachers and quality of multiple-choice test items

Based on the literature reviewed, theoretically, years of teaching influences test construction competencies and quality of test items; and test construction competencies are also related to the quality of test items. In view of that the effect of years of teaching on the theoretical relationship between test construction competencies and quality of multiple-choice test items was explored.

Findings that emerged from examining research hypothesis 4 point out that years of teaching was probably not a covariate to the relationship (if any) between test construction competencies and the quality of multiple-choice test items per the sample evidence. Preliminary hypothesis (research hypothesis 3) towards investigating research hypothesis 4 showed that there was no significant relationship between the teachers' test construction competencies and the quality of multiple-choice test items they constructed. Therefore, if no evidence of significant relationship, then years of teaching had nothing to confound. Moreover, in testing for the assumptions of Pearson Partial Correlation, there was no relationship between test construction competencies and years of teaching. This implies that the teachers' years of experiences in crafting and assessing students with multiple-choice items neither add up nor reduce their multiple-choice test construction competencies significantly. Thus, the teachers' multiple-choice test construction competencies might be due to their educational experiences with courses related to educational assessment of students.

Though years of teaching had no relationship with teachers' multiple-choice test construction competencies, it had significant relationship with the quality of the multiple-choice items. By implication, for an increase in years of teaching, there is the likelihood that the items' quality of multiple-choice test will decrease (or the number of problem items will be increasing while the number of good items decreasing). Information from a study conducted by Tshabalala et al. (2015) showed that most of the classroom teachers did not consider validity and reliability when constructing and marking teacher-made tests therefore affecting the quality of assessment results. Chan (2009) and Osadebe (2015) have also made similar observations that classroom teachers give inadequate attention to the quality of assessment instruments. Inferably, in this present study, the nature of relation observed with the quality of the multiple-choice test items and years of teaching could mean that as years of teaching increases, classroom teachers pay little attention to ensuring the quality of the multiple-choice tests they construct (or producing test items of good quality).

Benjafield (2010) has expressed that one's experiences affect his or her behaviour. Therefore, inadequate attention to ensuring the quality of multiple-choice test items might be embedded in the nature of classroom teachers' experiences with constructing the test items. Marso and Pigge (1992) in their study found that teachers "believe testing, evaluation, and grading activities are among their more demanding and less pleasant classroom responsibilities" (p. 25). Downing (2003) has stated that teachers perceive test construction procedures as waste of time and non-motivating. According to Burton et al. (1991), in general, good multiple-choice test items are more demanding and time-consuming to craft than other types of test items.

Situations or activities perceived as demanding, time-consuming, non-motivating and stressful can decrease one's commitment towards pursuing them with good interest. Kinyua and Okunya (2014) have pointed out that lack of teachers' commitment to good practice reduces the quality of assessment results. This might help explain the lack of significant correlation between the respondents' self-reported multiple-choice test construction competencies and the quality of the multiple-choice test items. Besides, it seems as years of teaching go by, teachers' experiences in the classroom or within the educational setting at large reduce their motivation and commitment towards crafting multiple-choice test items; therefore, resulting in observed decrease in the quality of multiple-choice test items they construct.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Overview of the Study

The study sought to investigate multiple-choice test construction competences of classroom teachers in the KSD. The correlation research design was used to examine the teachers' multiple-choice test construction competencies and the quality of the items they developed, as well as the effect of years of teaching on such relationship. The study was guided by three research questions and four hypotheses. With the use of purposive sampling technique, the study covered only 47 participants (n = 47) out of the 157 teachers. Questionnaires were administered to 47 teachers who voluntarily participated in the study. The data collected was analysed using means and standard deviations, frequency count, percentages, and Pearson Product-Moment Correlation Coefficient.

Summary of Findings

Finding based on research question 1 showed that, generally, most of the teachers in the KSD possess appreciable level of competencies in constructing multiple-choice items. In other words, most of the research participants possessed high levels of competencies in achieving content validity, handling items' alternatives and assembling test items.

Further exploration of multiple-choice test construction competencies indicated that the teachers found it very easy to exhibit competence in assembling test items, and easy to demonstrate competence in achieving content validity. However, they found it quite difficult to demonstrate competence in handling the items' alternatives.

Findings based on research question 2 indicated that most of the test items constructed by the research participants are described as good items per their respective difficulty and discrimination indices. However, further analysis indicated that the average value for the number of good items produced by the classroom teachers was not statistically greater than the average value for the number of problem items produced. Hence, it can be said that with respect to test quality, generally, the multiple-choice test items used for assessing students' achievement lacked a suitable level of psychometric properties; for the total number of good items produced was not statistically different from the total number of problem items.

Additional finding revealed by further examination of the problem items was that most of the items had unacceptable difficulty indices as compared to the discrimination indices. This means that in assessing the number of good items, the impact of unacceptable difficulty levels was greater than that of unacceptable discrimination indices. With a careful inspection of the difficulty levels, most of the items were described as difficult (that is, had p-values < .30).

Generally, findings in relation to research question 3 revealed that most of the tests were identified with the problem of 'detectable pattern of correct answer' as compared to the 'use of font size that students could find difficult to see and read'. Moreover, most of the tests examined with reference to constructional errors associated with problem items had the following issues: (a) clues to the correct answer, (b) instruction related issues (no or incomplete instruction), (c) time for completion of items not indicated on the test, (d) implausible distractors, and (e) ambiguous items/more than one correct answer as opposed to ‘cluing and linking items’.

Finding based on examination of research hypothesis 1 showed that there was a significant high positive relationship between test length and the number of good items produced by the research participants. Therefore, by implication, as test length increases there is a high likelihood that teachers will produce more multiple-choice test items of good quality.

Finding from research hypothesis 2 also gave contradicting evidence that there was a significant high positive linear relationship between test length and the number of problem items produced by the research participants. However, the respective correlation coefficients for research hypothesis 1(r = .79) and research hypothesis 2(r = .82) suggested that the probability of producing problem items was slightly greater than the probability of producing good items when increasing test length.

Finding for research hypothesis 3 gave the evidence that the relationship between test construction competencies and the quality of the multiple-choice test items was not significant. This means high or low score of a teacher’s test construction competence was not significantly related to either producing noticeably more good items than problem items or more problem items than good items.

Findings based on research hypothesis 4 showed that years of teaching is probably not a covariate to the relationship (if any) between test construction competencies and the quality of multiple-choice test items per the sample evidence. Preliminary hypothesis (research hypothesis 3) towards investigating research hypothesis 4 showed that there was no significant relationship between teachers' test construction competencies and the quality of multiple-choice test items they constructed. Thus, if no evidence of significant relationship, then years of teaching had nothing to confound.

Moreover, there was no relationship between test construction competencies and years of teaching (r^2 = .00). This implies that the teachers' years of experiences in crafting and assessing students with multiple-choice items neither add up nor reduce their multiple-choice test construction competencies. Hence, the teachers' observed test construction competencies might be due to their educational experiences with courses related to educational assessment of students.

Though years of teaching had no relationship with teachers' multiple-choice test construction competencies, it had significant negative relationship with the quality of the multiple-choice test items. By implication, as years of teaching increases, there is the likelihood that the items' quality of multiple-choice test will decrease. In other words, as years of teaching increases, the likelihood that the teachers will produce problem items will increase resulting in decrease in the number of good items.

Conclusions

As revealed in the literature, teachers' test construction competencies is key in ensuring that assessment results are of good quality. Nevertheless, self-report of the classroom teachers on their multiple-choice test construction competencies did not match well with the findings revealed as a result of direct assessment of the quality of the test items using quantitative item analysis approach. That is, generally, they reported appreciable levels or high levels of multiple-choice test construction competencies; however, the multiple-choice test items for assessing the students' achievement lacked a suitable level of psychometric properties since the total number of good items produced was not statistically different from the total number of problem items.

Besides, though years of teaching had no significant relationship with teachers' multiple-choice test construction competencies, as years of teaching increased, the items' quality of multiple-choice test also decreased. It implies that as years of teaching increases, the classroom teachers pay little attention to the quality of the multiple-choice items, which results in producing more problem items against the number of good items. Therefore, the increase in number of problem items, as years of teaching increases, influenced the items' quality in a way that it did not show significant relationship with the teachers' self-reported multiple-choice test construction competencies. With the assumption that the respondents were honest with their responses, it suggests that the teachers paid inadequate or little attention to ensure that problem items that reduce the quality of multiple-choice test items was avoided or adequately minimised.

Considering test-related factors that affect students' responses to multiple-choice items, qualitative evaluation of the problem items revealed item format and constructional errors such as detectable pattern of correct answer, implausible distractors, clues to answers, and ambiguities that affect quality of assessment results. As Kubiszyn and Borich (2013), Amedahe and Asamoah-Gyimah (2016) have indicated, problem items because of test-related factors such as ambiguous items, and clues to correct answers, make assessment results less valid for relevant educational decisions concerning students and classroom teachers.

Recommendations

Bearing in mind the conclusions drawn from the study on the basis of the research findings, the following recommendations are made:

District Directorate of Education

Years of teaching did not show any significant relationship with the teachers' test construction competencies acquired through taking a course related to educational assessment of students. Accordingly, it is recommended that District Directorate of Ghana Education Service place more emphasis on exposing classroom teachers to more practical ways of constructing multiple-choice items especially on how to effectively handle items' alternatives. This can be achieved through workshops, or training in test construction.

School Authorities and Classroom Teachers

As test length increases, the likelihood of producing good items increases. Therefore, in terms of the use of multiple-choice items to assess students learning outcomes, school authorities should aim at ensuring and encouraging teachers to craft relatively more items. However, caution should be exercised to ensure that the items are well constructed through effective moderation and qualitative examination of the test items to deal with format and constructional flaws. This particular caution is essential because it was also revealed that as test length increases, the probability of producing problem items was little above the likelihood of producing good items.

In addition, test construction competencies showed no significant relationship with the quality of the multiple-choice test items. This implied that the presence of problem items influenced the quality of the multiple-choice tests in a manner that it showed no significant relationship with the teachers' self-reported appreciable or high levels of multiple-choice test construction competencies. Samples of problem items examined revealed test-related factors such as detectable pattern of correct answer, implausible distractors, clues to answers, and ambiguities that affect the quality of test items. In view of that, it is recommended that school authorities and classroom teachers pay critical attention to these factors and put up appropriate measures that will improve the reliability and validity of results obtained from assessments that involve the use of multiple-choice test items.

Government, School Authorities, and Circuit Supervisors

It was observed that as years of teaching increases, there is the likelihood that the items' quality of multiple-choice test will decrease. In other words, as years of teaching increases, the likelihood that the teachers will produce problem items will increase resulting in decrease in the number of good items. Therefore, it appears that as years of teaching increases, classroom teachers pay little attention to the quality of multiple-choice items when constructing them. Therefore, it is recommended that the government, school authorities, and circuit supervisors should look for possible ways of encouraging classroom teachers to give more attention to the quality of multiple-choice test items they construct.

Suggestions for Further Research

Taking into consideration the fact that the scope of the study was delimited to ensure its feasibility, it is recommended that forthcoming research should focus on the following areas:

1. Teachers' perceptions and attitudes towards the construction of multiple-choice items in the Kwahu-South District.
2. Qualitative evaluation of objective items and essay items constructed by classroom teachers in the Kwahu-South District.
3. Replicating the study with wider range of population in other educational settings to explore the relationship between test construction competencies of classroom teachers and the quality of test items they construct.

REFERENCES

Adane, L. O. (2013). *Factors affecting low academic achievement of pupils in Kemp Methodist Junior High School in Aburi, Eastern region*. A thesis submitted to the University of Ghana, Ghana.

Adodo, S. O. (2013). Effects of two-tier multiple choice diagnostic assessment items on students' learning outcome in basic science technology (BST). *Academic Journal of Interdisciplinary Studies, 2*(2), 201-210.

Agu, N. N., Onyekuba, C., & Anyichie, C. A. (2013). Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: Need for a test construction skill inventory. *Educational Research and Reviews*, *8*(8), 431-439.

Ali, A. A. (1999). *Basic research skills in education*. Enugu: Orient Printing and Publishing.

Alkharusi, H. (2011). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience and in-service assessment training. *Turkish Science Education*, *8*(2), 39-47. Retrieved from http://www.tused.org.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press Inc.

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: 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.

Amedahe, F. K. (2002). *Fundamentals of educational research methods*. Mimeograph. UCC, Cape Coast.

Amedahe, F. K. (2014). *The issue of falling educational standards in Ghana: A perception or reality?* [Inaugural Lecture]. Cape Coast: University of Cape Coast Press.

Amedahe, F. K., & Asamoah-Gyimah, K. (2014). *Introduction to educational researc*h (5th ed.). Cape Coast: University Printing Press.

Amedahe, F. K., & Asamoah-Gyimah, K. (2016). *Introduction to measurement and evaluation* (7th ed.). Cape Coast: Hampton Press.

Anhwere, Y. M. (2009). *Assessment practices of teacher training college tutors in Ghana*. Unpublished Master's thesis, University of Cape Coast, Cape Coast.

Ary, D., Jacobs, C. L., Sorensen, C., & Razavieh, A. (2010). *Introduction to research methods in education* (8th ed.). Belmont, CA: Wadsworth.

Asamoah-Gyimah, K. (2002). *An evaluation of the practice of continuous assessment in the senior secondary schools in the Ashanti Region of Ghana.* Unpublished thesis, University of Cape Coast, Cape Coast, Ghana.

Barker, C., Pistrang, N., & Elliot, R. (2002). *Research methods in clinical psychology: An introduction for students and practitioners* (2nd ed.). Canada: John Willey and Sons.

Battle, J., & Lewis, M. (2002).The increasing significance of class. The relative effects of race and socioeconomic status on academic achievement. *Journal of Poverty*, *6(2)*, 21-35.

Benjafield, J. G. (2010). *A history of psychology* (4th ed.). New York: Oxford University Press.

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices* (2nd ed.). Retrieved from http://scholarcommons.usf.edu/oa-tesxtbook/3.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. White Plains, NY: Pearson Education.

Bunch, M. B. (2012). *Aligning curriculum, instruction, and assessment*. Retrieved from http://measurementinc.com/sites/default/files/2017-08/Aligning%20Curriculum%2C%20Assessment%2C%20and%20Instruction%20%28M.%20Bunch%2C%202012%29.pdf.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty*. Retrieved from https://testing.byu.edu/handbooks/ betteritems.pdf.

Butler, S. M., McColskey, W., & O'Sullivan, R. (2005). *How to assess student performance in science: Going beyond multiple-choice tests* (3rd ed.). [Associated with SERVE Center for Continuous Improvement at UNCG]. Retrieved from https://files.eric.ed.gov/fulltext/ED513873.pdf

Callahan, C. M. (2006). *Assessment in the classroom: The key to good instruction (The practical strategies series in gifted education)*. Waco, TX: Prufrock Press, Inc.

Chan, K. K. (2009). Using test blueprint in classroom assessments: why and how. Paper presented at the 35th *International Association for Educational Assessment* (IAEA) Annual Conference, Brisbane, Australia.

Choi, K., Nam, J. H., & Lee, H. (2001). The effects of formative assessment with detailed feedback on students' science learning achievement and attitude regarding formative assessment. *Science Education International*, *12*(2), 28- 34.

Clark, P. V. K., & Creswell, J. W. (2015). *Understanding research: A consumer's guide* (2nd ed.). New York: Pearson.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). New York: Routledge.

Cohen, R. J., & Swerdlik, M. J. (2010). *Psychological testing and assessment: An introduction to tests and measurement* (7th ed.). New York: McGaw-Hill.

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.

Domino, G., & Domino, M. L. (2006). *Psychological testing: An introduction* (2nd ed.). Cambridge: Cambridge University Press.

Dosumu, C. T. (2002). *Issues in teacher-made tests*. Ibadan: Olatunji and Sons Publishers.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 83-100.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). New Jersey: Prentice-Hall.

Elharrar, Y. (2006). *Teacher assessment practices and perceptions: the use of alternative assessments within the Quebec educational reform* (Doctoral dissertation, Université du Québec à Montréal). Retrieved from http://www.archipel.uqam.ca/9490/1/D1368.pdf.

Feldman, M. S., Bell, J., & Berger, M. T. (2003). *Gaining access: A practical and theoretical guide for qualitative researchers.* Walnut Creek, CA: AltaMira.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). London: Sage.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). Boston: McGrew Hill.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). Boston: McGrew Hill.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). London: Sage.

Gareis, R. C., & Grant, W. L. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning* (2nd ed.). New York: Routledge.

Gattullo, F. (2000). Formative assessment in primary (elementary) ELT classes: An Italian case study. *Language Testing*, *17*(2), 278-288.

Gilley, J. W., & Steven, A. E. (1989). *Principles of human resource development*. New York: Addison Wesley Pub. Company. Inc.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: McMillan.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley.

Guskey, T. R. (2003). How classroom assessments improve learning. *Educational Leadership, 60*(5), 6–11.

Guskey, T. R., & Jung, L. A. (2013). *Answer to essential questions about standards, assessments, grading, & reporting*. Thousand Oaks, CA: Corwin.

Hamafyelto, R. S., Hamman-Tukur, A., & Hamafyelto, S. S. (2015). Assessing teacher competence in test construction and content validity of teacher made examination questions in commerce in Borno State, Nigeria. *Journal of Education, 5*(5), 123-128.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice*, *12*(3), 38-47.

Harpster, D. L. (1999). *A study of possible factors that influence the construction of teacher-made problems that assess higher-order thinking skills*. Doctoral Dissertation, Montana State University-Bozeman, College of Education, Health & human Development.

Harris, R. J. (2002*). What every parent needs to know about standardised tests: How to understand the tests and help kids score high!*. New York: McGraw-Hill.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*, 191-205. Retrieved from https://doi.org/10. 1177/1094428104263675

Healey, J. F. (2012). *Statistics: A tool for social research* (9th ed.). Belmont: Wadsworth, Cengage Learning.

Howitt, D., & Cramer, D. (2011). *Introduction to research methods in psychology* (3rd ed.). Harlow: Prentice Hall.

Huck, S. W. (2012). *Reading statistics and research* (6th ed.). Boston, MA: Pearson.

Izard, J. (2005). Overview of test construction. In K. N. Ross (Ed.), *Quantitative research methods in educational planning* (Module 6). Paris: International institute for educational planning/UNESCO. Retrieved from http://www2.tf.jcu.cz/~bauman/KPD_VTMP_ KVTMP/Educational_Research.pdf.

Johnson, R. B., & Christensen, L. B. (2004). *Educational research: quantitative, qualitative, and mixed approaches*. Boston, MA: Allyn and Bacon.

Joshua, M. T. (2005) *Fundamentals of test and Measurement in Education*. Calabar: University of Calabar Press.

Kinyua, K., & Okunya, L. O. (2014). Validity and reliability of teacher-made tests: Case study of year 11 physics in Nyahururu District of Kenya. *African Educational Research Journal, 2*(2), 61-71.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30*, 607-610.

Kubiszyn, T., & Borich, G. D. (2013*). Educational testing and measurement: Classroom application and practice* (10th ed.). Hoboken, NJ: Wiley.

Leedy, P. D., & Ormrod, J. E. (2013). *Practical research: Planning and design* (10th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.

Lucy, A. (2014). Tracing a beginning elementary teacher's development of identity for science teaching. *Journal of Teacher Education*, *65*(3) 223–240.

Maba, W. (2017). Teachers' sustainable professional development through classroom action research. *International Journal of Research in Social Sciences*, *7*(8), 718-732.

Magno, C. (2003). The profile of teacher-made test construction of the professors of University of Perpetual Help Laguna. *UPHL Institutional Journal, 1*(1), 48-55.

Magnusson, D. (1967). *Test theory*. Boston: Addison-Wesley.

Marso, R. N., & Pigge, F. L. (1989). *The status of classroom teachers' test construction proficiencies: Assessments by teachers, principals, and supervisors validated by analyses of actual teacher-made tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, California. (ERIC Document Reproduction Service No. ED306283).

Marso, R. N., & Pigge, F. L. (1992). *A summary of published research: Classroom teachers knowledge and skills related to the development and use of teacher-made tests*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 346 148).

Mayers, A. (2013). *Introduction to statistics and SPSS in psychology*. Boston: Pearson Education, Inc.

McMillan, J. H. (2000). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin publishing company.

McMillan, J. H. (2013). Why we need research on classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 3–16). Thousand Oaks: Sage.

Moore, D. S. (2001). *Statistics: Concepts and controversies* (5th ed.). New York: W. H. Freeman and Company.

Morrow, J. R., Jr., Jackson, A. W., Disch, J. G., & Mood, D. P. (2000). *Measurement and evaluation in human performance* (2nd ed.). Champaign, IL: Human Kinetics.

Morrow, J. R., Jr., Mood, D. P., Disch, J. G., & Kang, M. (2016). *Measurement and evaluation in human performance* (5th ed.). Champaign, IL: Human Kinetics.

Neuman, W. L. (2014). *Basics of social research: Qualitative and quantitative approaches* (3rd ed.). England: Pearson Education Limited.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Oduro-Okyireh, G. (2008). *Testing practices of senior secondary school teachers in the Ashanti region*. Unpublished master's thesis. University of Cape Coast, Cape Coast, Ghana.

Oliver, P. (2010). *The student's guide to research ethics* (2nd ed.). Berkshire: Open University Press.

Ololube, N. P. (2008). Evaluation competencies of professional and non professional teachers in Nigeria. *Studies in Educational Evaluation (SEE)*, *34*(1), 44-51.

Oosterhof, A. (2003). *Developing and using classroom assessments* (3rd ed.). Upper Saddle River, NJ: Merill/Prentice Hall.

Osadebe, P. U. (2015). Construction of valid and reliable test for assessment of students. *Journal of Education and Practice*, *6*(1), 51-56.

Osuola, E. C. (2001). *Introduction to research methodology* (3rd ed.). Onitsha, Nigeria: Africana F.E.P Publishers Ltd.

Ovat, S., & Ofem, J. U. (2017). Teachers variables and application of test blue prints in learners assessment in secondary schools in cross river state. *International Journal of Scientific Research in Education, 10(*1), 112-118.

Polit, D. F., & Beck, C. T. (2004). *Nursing research: Principles and methods*. Philadelphia, PA: Lippincott Williams & Wilkins.

Quaigrain, A. K. (1992). *Teacher–competence in the use of essay tests: A study of secondary schools in the Western Region of Ghana.* Unpublished thesis. University of Cape Coast, Cape Coast, Ghana.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4*, 1-11.

Rivera, J. E. (2007). *Test item construction and validation: Developing a statewide assessment for agricultural science education*. Unpublished dissertation, Cornell University. Retrieved from http://citeseerx.ist.psu. edu/viewdoc/download?doi=10.1.1.850.6608&rep=rep1&type=pdf.

Sanders, J. R., & Vogel, S. R. (1993). "3. *The Development of Standards for Teacher Competence in Educational Assessment of Students*". [Teacher Training in Measurement and Assessment Skills.5]. Retrieved from http://digitalcommons.unl.edu/burosteachertraining/5.

Sasu, E. O. (2017). *Testing practices of junior high school teachers in the Cape Coast metropolis*. Unpublished Mather's thesis, University of Cape Coast, Cape Coast, Ghana.

Shillingburg, W. (2016). *Understanding validity and reliability in classroom, school-wide, or district-wide assessments to be used in teacher/principal evaluations*. Retrieved from https://cms.azed. gov/home/GetDocumentFile?id=57f6d9b3aadebf0a04b2691a.

Silker, R. T. (2003). *Teachers and tests*. London: Basil Blackwell.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Erlbaum.

Tshabalala, T., Mapolisa, T., Gazimbe, P., & Ncube, A. C. (2015). Establishing the effectiveness of teacher-made tests in Nkayi district primary schools. *Nova Journal of Humanities and Social Sciences, 4*(1), 1- 6.

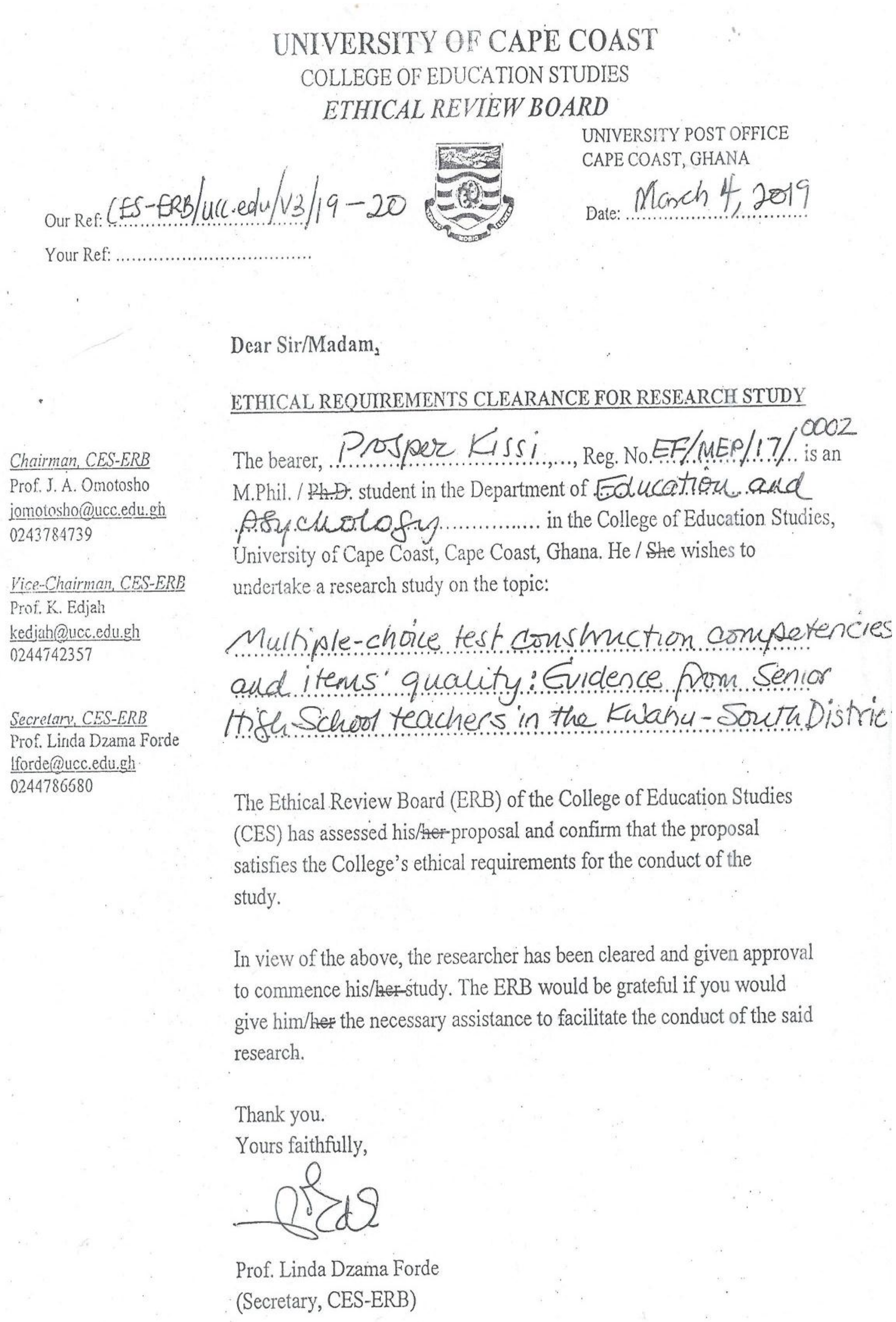
Ujah, E. U. (2001). *Development and validation of an introductory technology achievement test*. Unpublished M.Ed. Thesis, University of Nigeria, Nsukka.

Wiredu, S. G. (2013). *Assessment practices of tutors in the nurses' training colleges in the Western and Central regions of Ghana*. Unpublished Master's thesis, University of Cape Coast, Cape Coast.

APPENDICES

APPENDIX A

ETHICAL CLEARANCE



UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES
ETHICAL REVIEW BOARD

UNIVERSITY POST OFFICE
CAPE COAST, GHANA

Our Ref: CES-ERB/ucc.edu/V3/19-20

Date: March 4, 2019

Your Ref: ....

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

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

Secretary, CES-ERB
Prof. Linda Dzama Forde
lforde@ucc.edu.gh
0244786680

Dear Sir/Madam,

ETHICAL REQUIREMENTS CLEARANCE FOR RESEARCH STUDY

The bearer, Prosper Kissi, Reg. No. EF/MEP/17/0002 is an M.Phil. / ~~Ph.D.~~ student in the Department of Education and Psychology in the College of Education Studies, University of Cape Coast, Cape Coast, Ghana. He / ~~She~~ wishes to undertake a research study on the topic:

Multiple-choice test construction competencies and items' quality: Evidence from Senior High School teachers in the Kwahu-South Distric

The Ethical Review Board (ERB) of the College of Education Studies (CES) has assessed his/~~her~~ proposal and confirm that the proposal satisfies the College's ethical requirements for the conduct of the study.

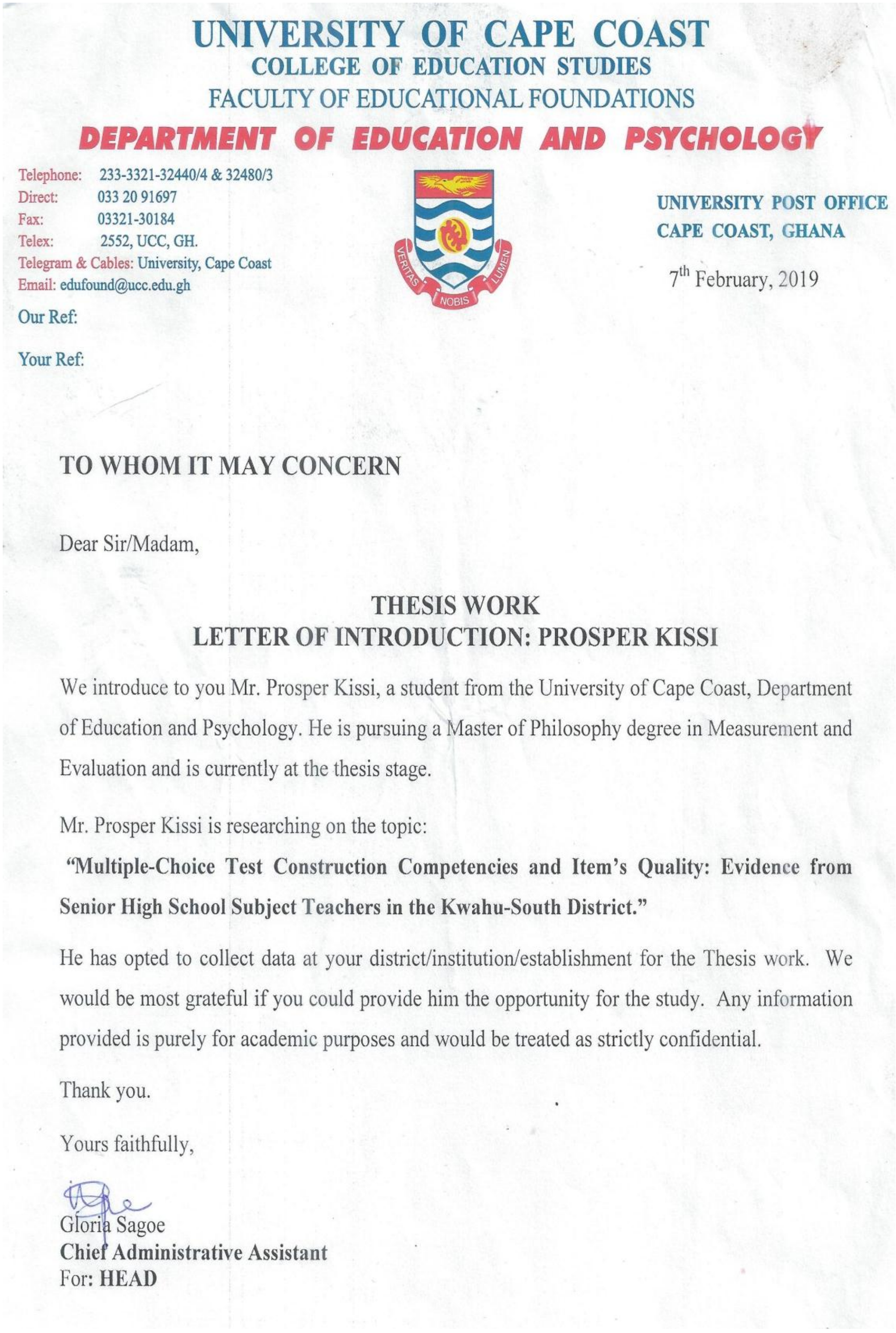
In view of the above, the researcher has been cleared and given approval to commence his/~~her~~ study. The ERB would be grateful if you would give him/~~her~~ the necessary assistance to facilitate the conduct of the said research.

Thank you.
Yours faithfully,

Prof. Linda Dzama Forde
(Secretary, CES-ERB)

APPENDIX B

INTRODUCTORY LETTER



UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES
FACULTY OF EDUCATIONAL FOUNDATIONS
DEPARTMENT OF EDUCATION AND PSYCHOLOGY

Telephone: 233-3321-32440/4 & 32480/3
Direct: 033 20 91697
Fax: 03321-30184
Telex: 2552, UCC, GH.
Telegram & Cables: University, Cape Coast
Email: edufound@ucc.edu.gh

UNIVERSITY POST OFFICE
CAPE COAST, GHANA

7th February, 2019

Our Ref:

Your Ref:

TO WHOM IT MAY CONCERN

Dear Sir/Madam,

THESIS WORK
LETTER OF INTRODUCTION: PROSPER KISSI

We introduce to you Mr. Prosper Kissi, a student from the University of Cape Coast, Department of Education and Psychology. He is pursuing a Master of Philosophy degree in Measurement and Evaluation and is currently at the thesis stage.

Mr. Prosper Kissi is researching on the topic:

"Multiple-Choice Test Construction Competencies and Item's Quality: Evidence from Senior High School Subject Teachers in the Kwahu-South District."

He has opted to collect data at your district/institution/establishment for the Thesis work. We would be most grateful if you could provide him the opportunity for the study. Any information provided is purely for academic purposes and would be treated as strictly confidential.

Thank you.

Yours faithfully,

Gloria Sagoe
Chief Administrative Assistant
For: **HEAD**

APPENDIX C

CONSENT FORM

Dear Sir/Madam,

I am to conduct a study with the purpose of investigating how Senior High School teachers in the Kwahu-South District write multiple-choice test items. I, therefore, write to seek your consent to voluntary participate in the study. There is no penalty for not participating. Moreover, there are no foreseeable risks associated with this study.

I must state that you have the right to withdraw from the study at any time without consequence. In addition, any information provided will be kept anonymous and treated as strictly confidential.

To obtain the necessary information to meet the purpose of the study, you will be required to:

1. respond to a questionnaire;
2. make available all of the following documents: (a) a copy of your latest end-of-semester self-constructed and administered multiple-choice test and (b) its marking scheme; and (c) students' responses on the administered end-of-semester multiple-choice test items.

Please complete the information below to document your agreement to participate

I ______________________________________ (your name) have been informed about the purpose of the study and the ethical issues involved. I understand that the information given will be used for the intended academic purpose. I also understand that I can withdraw from the exercise at any point in time that I wish to.

I do give my consent to participate in the study as I have received a copy of this consent form.

____________________________ ____________________________

Your signature Date

I have received this consent from the research participant and I will ensure anonymity and confidentiality of information provided.

____________________________ ____________________________

Candidate's signature Date

APPENDIX D

ITEMS FOR THE PILOT TESTING

SECTION B

This section assesses the extent each statement applies to you

INSTRUCTION

Carefully read each of the statements and decide how the statement applies to YOU. Please tick a box based on the following guide: Strongly Disagree **(SD)**, Disagree **(D)**, Agree **(A)**, and Strongly Agree **(SA)** to show the extent to which each of the statements applies to you.

Q/N	STATEMENTS	SD	D	A	SA
When constructing multiple-choice test, I :					
1.	match test items to instructional objectives (intended outcomes of the appropriate difficulty level)				
2.	make sure each item deals with an important aspect of content area				
3.	prepare marking scheme while constructing the items				
4.	pose clear and unambiguous items				
5.	give specific instructions on the test				
6.	present a definite, explicit and singular question or problem in the stem				
7.	include in the stem any word(s) that might otherwise be repeated in each option				
8.	emphasise negative word (e.g. by underlining and/or capitalising or bolding) in negatively stated stem				
9.	make the options grammatically consistent with the stem				
10.	make options independent of each other				
11.	use **"all of the above"** as part of the options to the stem of an item				
12.	avoid the use of **"none of the above"** as an option when an item is of the best answer type				
13.	make options approximately equal in length				
14.	present options in some logical order (e.g., chronological, most to least, alphabetical) when possible				
15.	include questions of varying difficulty				
16.	match items to vocabulary level of the students				
17.	give appropriate time for completion of test				
18.	use appropriate number of test items				
19.	number the test items one after the other				
20.	appropriately assign page numbers to the test				
21.	properly space the test items for easy reading				
22.	keep all parts of an item (stem and its options) on the same page				
23.	review test items for construction errors				

THANK YOU!!!

APPENDIX E

PARALLEL ANALYSIS BASED ON THE 23 ITEMS

```
Monte Carlo PCA for Parallel Analysis


3/28/2019   1:57:29 AM
Number of variables:     23
Number of subjects:     130
Number of replications: 100

++++++++++++++++++++++++++++++++++++++++++++++++++++
Eigenvalue #     Random Eigenvalue     Standard Dev
++++++++++++++++++++++++++++++++++++++++++++++++++++
      1              1.8500               .0852
      2              1.7123               .0632
      3              1.5948               .0568
      4              1.4952               .0419
      5              1.4079               .0376
      6              1.3294               .0357
      7              1.2624               .0347
      8              1.1906               .0346
      9              1.1251               .0327
     10              1.0604               .0312
     11              1.0033               .0305
     12              0.9474               .0298
     13              0.8931               .0313
     14              0.8376               .0295
     15              0.7883               .0311
     16              0.7358               .0316
     17              0.6848               .0306
     18              0.6396               .0296
     19              0.5911               .0274
     20              0.5447               .0275
     21              0.4895               .0260
     22              0.4398               .0293
     23              0.3768               .0309
++++++++++++++++++++++++++++++++++++++++++++++++++++
3/28/2019   1:57:34 AM

Monte Carlo PCA for Parallel Analysis
©2000 by Marley W. Watkins. All rights reserved.
****************************************************
```

APPENDIX F

ROTATED COMPONENT MATRIX FOR THE 23 ITEMS

	Description	Factors		
		1	**2**	**3**
	Percentage of variance explained (after rotation)	12.007	11.836	11.269
	Initial eigenvalue	3.706	2.483	1.887
Q/N	**When constructing multiple-choice test, I:**	**1**	**2**	**3**
2	make sure each item deals with an important aspect of content area	.760		
1	match test items to instructional objectives (intended outcomes of the appropriate difficulty level)	.715		
3	prepare marking scheme while constructing the items	.519		
17	give appropriate time for completion of test	.519		
4	pose clear and unambiguous items	.488		
15	include questions of varying difficulty	.413		
16	match items to vocabulary level of the students	.411		
21	properly space the test items for easy reading		.657	
23	review test items for construction errors		.646	
22	keep all parts of an item (stem and its options) on the same page		.589	
18	use appropriate number of test items		.544	
20	appropriately assign page numbers to the test		.524	
5	give specific instructions on the test		.493	
19	number the test items one after the other		.469	
8	emphasise negative word (e.g. by underlining and/or capitalising or bolding) in negatively stated stem			
13	make options approximately equal in length			.707
7	include in the stem any word(s) that might otherwise be repeated in each option			.648
14	present options in some logical order (e.g., chronological, most to least, alphabetical) when possible			.556
10	make options independent of each other			.547
12	avoid the use of "none of the above" as an option when an item is of the best answer type			.519
9	make the options grammatically consistent with the stem			.468
11	use "all of the above" as part of the options to the stem of an item			-.405
6	present a definite, explicit and singular question or problem in the stem			

Extraction Method: Principal Component Analysis

APPENDIX G

PARALLEL ANALYSIS AFTER REMOVAL OF ITEM 11

```
Monte Carlo PCA for Parallel Analysis


3/28/2019   2:40:46 AM
Number of variables:     22
Number of subjects:     130
Number of replications: 100

+++++++++++++++++++++++++++++++++++++++++++++++++++++
Eigenvalue #     Random Eigenvalue     Standard Dev
+++++++++++++++++++++++++++++++++++++++++++++++++++++
      1               1.8212               .0809
      2               1.6667               .0634
      3               1.5644               .0546
      4               1.4720               .0437
      5               1.3865               .0393
      6               1.3115               .0402
      7               1.2419               .0363
      8               1.1732               .0332
      9               1.0984               .0325
     10               1.0377               .0281
     11               0.9795               .0313
     12               0.9197               .0300
     13               0.8637               .0283
     14               0.8128               .0247
     15               0.7577               .0287
     16               0.7062               .0263
     17               0.6556               .0258
     18               0.6130               .0283
     19               0.5610               .0274
     20               0.5091               .0255
     21               0.4544               .0324
     22               0.3936               .0338
+++++++++++++++++++++++++++++++++++++++++++++++++++++
3/28/2019   2:40:51 AM

Monte Carlo PCA for Parallel Analysis
©2000 by Marley W. Watkins. All rights reserved.
*****************************************************
```

APPENDIX H

ROTATED COMPONENT MATRIX AFTER REMOVAL OF ITEM 11

Description		Factors		
		1	**2**	**3**
Percentage of variance explained (after rotation)		12.341	12.225	11.604
Initial eigenvalue		3.706	2.380	1.872
Q/N	**When constructing multiple-choice test, I:**	**1**	**2**	**3**
21	properly space the test items for easy reading	.660		
23	review test items for construction errors	.643		
22	keep all parts of an item (stem and its options) on the same page	.588		
18	use appropriate number of test items	.544		
20	appropriately assign page numbers to the test	.521		
5	give specific instructions on the test	.493		
19	number the test items one after the other	.469		
2	make sure each item deals with an important aspect of content area		.755	
1	match test items to instructional objectives (intended outcomes of the appropriate difficulty level)		.683	
3	prepare marking scheme while constructing the items		.535	
4	pose clear and unambiguous items		.521	
17	give appropriate time for completion of test		.510	
15	include questions of varying difficulty		.406	
16	match items to vocabulary level of the students		.401	
13	make options approximately equal in length			.717
7	include in the stem any word(s) that might otherwise be repeated in each option			.653
10	make options independent of each other			.569
14	present options in some logical order (e.g., chronological, most to least, alphabetical) when possible			.550
12	avoid the use of “none of the above” as an option when an item is of the best answer type			.516
9	make the options grammatically consistent with the stem			.491
6	present a definite, explicit and singular question or problem in the stem			
8	emphasise negative word (e.g. by underlining and/or capitalising or bolding) in negatively stated stem			

Extraction Method: Principal Component Analysis

APPENDIX I

TEACHERS' MULTIPLE-CHOICE TEST CONSTRUCTION COMPETENCE QUESTIONNAIRE (TTCCQ-MC)

The main aim of this questionnaire is to elicit the necessary responses relevant to the purpose of this study. Any information provided by participants is for academic purpose only. In view of this, confidentiality of information provided is fully assured to respondents.

SECTION A

This section collects data on the demographic variables of participants

INSTRUCTION

Please tick **[√]** the appropriate response where required

CODE:

A. Gender:

1. Male [ ]
2. Female [ ]

B. Subject teaching: *(Please indicate/tick the major subject teaching)*

1. Business Management [ ]
2. Core Mmathematics [ ]
3. Cost Accounting [ ]
4. Economics [ ]
5. English Language [ ]
6. Financial Accounting [ ]
7. Integrated Science [ ]

C. Which class do you teach? *(Please tick one class)*

1. Form **ONE** [ ]
2. Form **TWO** [ ]
3. Form **THREE** [ ]

D. Teacher Qualification **(Highest qualification)**:

1. Master's Degree: M.Phil. [ ]
   M.Ed. [ ]
   M.A.: With education [ ]*
   Without education [ ]
2. First Degree: With education [ ]
   Without education [ ]
3. Higher National Diploma (HND) [ ]
4. Diploma in Basic Education (DBE) [ ]

Any other: (Please specify in the space provided)……………………………………….

E. For how many years have you been teaching at the **senior high school level**? (Please write in the space provided)………………

SECTION B

This section assesses the extent each statement applies to you

INSTRUCTION

Carefully read each of the statements and decide how the statement applies to YOU. Please tick a box based on the following guide: Strongly Disagree **(SD)**, Disagree **(D)**, Agree **(A)**, and Strongly Agree **(SA)** to show the extent to which each of the statements applies to you.

Q/N	**STATEMENTS**	**SD**	**D**	**A**	**SA**
	When constructing multiple-choice test, I:				
1.	properly space the test items for easy reading				
2.	review test items for construction errors				
3.	keep all parts of an item (stem and its options) on the same page				
4.	use appropriate number of test items				
5.	give specific instructions on the test				
6.	appropriately assign page numbers to the test				
7.	number the test items one after the other				
8.	make sure each item deals with an important aspect of content area				
9.	match test items to instructional objectives (intended outcomes of the appropriate difficulty level)				
10.	prepare marking scheme while constructing the items				
11.	give appropriate time for completion of test				
12.	pose clear and unambiguous items				
13.	include questions of varying difficulty				
14.	match items to vocabulary level of the students				
15.	make alternatives approximately equal in length				
16.	include in the stem any word(s) that might otherwise be repeated in each alternative				
17.	present alternatives in some logical order (e.g., chronological, most to least, alphabetical) when possible				
18.	make alternatives independent of each other				
19.	avoid the use of **"none of the above"** as an option when an item is of the best answer type				
20.	make the alternatives grammatically consistent with the stem				

THANK YOU!!!

APPENDIX K

Multiple-Choice Test Error Analysis Checklist

Instruction: Record once if each error have occurred several times or once for each test.

Q/N	ERRORS IN CONSTRUCTING MULTIPLE-CHOICE TEST	Number of Occurrence Across Tests	Total
	Test format errors		
1.	Alternatives not presented in some logical order		
2.	Detectable pattern of correct answer		
3.	Horizontal arrangement of options		
4.	Options of items appearing in different columns/pages		
5.	Page numbers not assigned		
6.	Poor arrangement of items/spacing of test items		
7.	Use of font size difficult to see and read		
	Item construction errors		
8.	Ambiguous items/More than one correct answer		
9.	Central theme, task or problem not presented in the stem		
10.	Clues to the correct answer		
11.	Cluing and linking items		
12.	Grammatical, punctuation, and spelling		
13.	Heterogeneous options		
14.	Implausible distractors		
15.	Instructional related issues (No/ Incomplete instruction)		
16.	No answer		
17.	Not emphasising (e.g. bolding, underlying or capitalising) negative word in the stem		
18.	Time for completion of items not indicated on the test		
19.	Use of “all of the above”		
20.	Wrong answer		
21.	Wrong key to item		
22.	Wrong usage of “none of the above”		