UNIVERSITY OF CAPE COAST

TESTING PRACTICES OF SENIOR SECONDARY SCHOOL TEACHERS IN THE ASHANTI REGION OF GHANA

GEORGE ODURO-OKYIREH

2008

UNIVERSITY OF CAPE COAST

TESTING PRACTICES OF SENIOR SECONDARY SCHOOL TEACHERS IN THE ASHANTI REGION OF GHANA

BY

GEORGE ODURO-OKYIREH

A THESIS SUBMITTED TO THE DEPARTMENT OF EDUCATIONAL FOUNDATIONS OF THE FACULTY OF EDUCATION, UNIVERSITY OF CAPE COAST, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF PHILOSOPHY DEGREE IN MEASUREMENT AND EVALUATION

MARCH, 2008

DECLARATION

Candidate's Declaration

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

Candidate's Name:…… Date:……

Signature……

Supervisors' Declaration

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

Principal Supervisor's Name:…… Date:……

Signature:……

Co-Supervisor'sName:…….. Date:……

Signature:……

ABSTRACT

The study aimed at finding out whether Senior Secondary School teachers in the Ashanti Region of Ghana followed the basic principles in their testing practices. It also sought to find out whether pre-service training in testing contributes to competence in actual testing practice.

Cluster and simple random sampling techniques were adopted to select 265 teachers of Mathematics, Integrated Science and English Language in 26 Senior Secondary Schools for the study.

Eight research questions guided the study. A 52-item questionnaire and a 17-item observation guide were used for data collection. The reliability coefficient of the questionnaire was 0.70.

The study showed that to a great extent, teachers followed the basic principles in test construction, administration and scoring. Teachers applied seven out of 10 principles in test construction, 12 out of 18 principles in test administration and six out of nine principles in test scoring. Also, teachers reported they used both norm-referenced and criterion-referenced approaches in their test-score interpretation. Again, the findings indicated that pre-service instruction in educational measurement had a positive impact on actual testing practice, although the impact was quite subtle.

It was recommended that since competence in assessment is key to teacher effectiveness, every teacher must be given formal training in educational measurement and evaluation during pre-service training. Again, the teacher training universities in Ghana should accentuate the practical classroom aspects of the testing principles in their educational measurement courses to help teachers practicalise the theoretical knowledge they acquire.

ACKNOWLEDGEMENTS

I wish to express my appreciation to all who in diverse ways contributed to the success of this study.

My deepest gratitude goes particularly to my principal supervisor Dr. Y. K. A. Etsey, for his fatherly love, encouragement, guidance and support throughout the research and the entire M.Phil programme.

I am greatly indebted to Professor F. K. Amedahe, my co-supervisor, for his expert advice and suggestions during this research work. Through his intellectual interaction, a lot have been learnt.

I also express my sincere appreciation to all the lecturers of the Educational Foundations Department especially Professor Essuman, Dr. Frederick Ocansey, Messrs Koawo Edjah, and Godwin Awabil for the diverse ways they contributed to the completion of the M.Phil programme.

My appreciation also extends to my course mates, Mr. Eric Anane and Miss Justine V. A. Adzraku for their help and encouragement in diverse ways.

My last appreciation goes to all the Core Mathematics, English Language and Integrated Science teachers in the secondary schools selected for the study for accepting to participate in the study.

DEDICATION

This work is dedicated to my parents, Mr. and Mrs. Mark Oduro-Okyireh, my wife, Regina, and children, Emmanuel and Michelle.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

Background to the Study

It is absolutely impossible for anybody to study in an entire educational system without being exposed to a wide range of educational and psychological assessment procedures. This is because constantly in an educational system, decisions have to be made about students, curricula and programmes, and educational policies. According to Nitko (1996), decisions about students include managing classroom instruction, placing students into different types of programmes, assigning them to appropriate categories, guiding and counselling them, selecting them for educational opportunities and credentialling and certifying their competence. Decisions about curricula and programmes include decisions about their effectiveness (summative assessment) and about ways to improve them (formative assessment). In Ghana, decisions about educational policies are made at the national level. It is worth knowing, however, that educational assessments, of which in the Ghanaian educational system, tests predominate, provide some of the needed information for these types of decisions.

Tests are indispensable tools in every educational system. Tests and teaching are interwoven. Quaigrain (1992) has stated that "tests provide

needed information for evaluation. Without evaluation there cannot be feedback and knowledge of results. Without knowledge of results there cannot be any systematic improvement in learning" (p. 1).

In the Ghanaian educational system, standardised achievement, aptitude, and intelligence tests that are found in the developed countries such as the United States of America (USA), Canada and Great Britain are to a large extent non-existent. The tests that are conducted by the West African Examinations Council (WAEC) at the terminal points of the educational system cannot be said to be standardised since they do not meet all the standard characteristics of standardised achievement tests. Examples of the WAEC conducted tests are the Basic Education Certificate Examination (BECE) and the Senior Secondary School Certificate Examination (SSSCE). According to Linn and Gronlund (1995), the characteristics of a carefully constructed standardised achievement test include the following:

1. The test items are of a high technical quality. They have been developed by educational and test specialists, tried out experimentally (pretested) and selected on the basis of difficulty, discriminating power and relationship to a clearly defined and rigid set of specifications.
2. Directions for administering and scoring are so precisely stated that the procedures are standard for different users of the test.
3. Norms based on national samples of students in the grades where the test is intended for use are provided as aids in interpreting the scores.
4. Equivalent and comparable forms of the tests are usually provided as well as information concerning the degree to which the tests are comparable.

5. A test manual and other accessory materials are included as guides for administering and scoring the test, evaluating its technical qualities, and interpreting and using the results.

These characteristics of carefully constructed standardised achievement test make it clear the total absence of such tests in the Ghanaian classroom. In view of this situation, testing of achievement in the Ghanaian educational system is mainly through the use of informal classroom tests or teacher-made tests. But for a very effective and efficient instructional programme in any educational system, there is the need for both standardised achievement tests and teacher-made tests, in the sense that by way of their respective functions, one complements the other. Standardised achievement tests mainly measure outcomes and content common to the national curriculum, test basic skills and complex learning outcomes, but seldom reflect emphasis or timeliness of the classroom situation. Teacher-made tests on the other hand are generally well adapted to outcomes and content of the classroom. "They have the flexibility that affords continuous adaptation of measurement to new material and changes in procedure, and well adaptable to various-sized work units, but tend to neglect complex learning outcomes" (Linn & Gronlund, 1995, p. 367).

In an educational system such as the Ghanaian situation where standardised achievement tests are non-existent and all the information needed for important instructional decisions are provided by informal classroom tests, there is the need for teachers to always ensure that they follow the standard approved principles in the construction, administration and scoring of their tests and the interpretation of the test results. This way, they would be striving

to make their test scores more reliable so that the uses to which the test scores will be put will be as sound and appropriate as possible. This will minimise the negative consequences that students are likely to experience when their test results are used as intended.

The situation with respect to achievement testing in the Ghanaian educational system as discussed in the paragraphs above is a matter of concern. This is because this very indispensable educational exercise to a large extent has become the sole responsibility of the classroom teacher. Whether teachers are adequately prepared and professionally well equipped to execute this responsibility as expected is also a matter of concern. As pointed out by Amedahe (1989), irrespective of pre-service training, teachers in Ghana construct, administer, score and interpret the results of classroom achievement tests. He continued by noting that, while some teachers have received in their college courses, pre-service instruction concerning the construction, administration, scoring of tests and the interpretation of test results, others have not. The possible effects of this state of affairs seem to be seen in the testing practices of many secondary school teachers in Ghana today. By general observation, the effects range from lifting test items from textbooks, test items testing only recall of facts, improper wording of test items resulting in ambiguity, unreasonable difficulty levels of items, unclear directions, unreasonable time limit allotment, subjective and inconsistent scoring, to test results that are interpreted wrongly or not interpreted at all.

In buttressing the point above, McDaniel (1994, p. 4), recorded that:

> Teachers as well as publishers sometimes succumb to writing quick memory level items. A study of 342 teacher-made tests revealed that most teachers use short answer tests measuring knowledge of facts, terms and principles . The tests require students to remember but not to apply knowledge. Such tests are easy to construct but they send the wrong signals to students about the things we value in education.

Touching on the consequences of using the results of improperly constructed teacher-made tests in making decisions about students, Sinclair (cited in Amedahe, 1989, p.2), stated that , "decisions of major consequences to the individual are increasingly being made on the basis of his performance in tests."

For a very sound instructional programme in the Ghanaian educational system, the onus rests with teachers to take time to create fair, focused and well-thought-out achievement tests. In this case teachers can have confidence in the evidence they gather and make meaningful judgements about students' performance and future instructional plans and decisions. McDaniel (1994, p.4), says, "there is nothing mysterious in constructing better classroom tests. Three steps will lift your tests out of the ordinary and provide a sounder basis for evaluating students' achievements. These are test planning, item analysis, and revision."

Statement of the Problem

Amedahe (1989) conducted a study to determine whether teachers in the secondary schools in the Central Region of Ghana in constructing, administering and scoring their classroom achievement tests followed the principles prescribed by testing experts. The main summary of the findings of the study was that, to a great extent teachers in the study did not follow the basic prescribed principles of test construction, administration and scoring of their classroom achievement tests. Among other things, Amedahe recommended that in order to refute or confirm the tentative findings of his study, an extensive research that will cover most of the subjects taught or the entire secondary school system is needed. Also, in order to generalise the study for teachers in Ghana, there was the need to study the testing practices of secondary school teachers in other regions in the country. It is based on these recommendations that this study was conducted.

The Ghana Education Service (GES), acting under directives from the Ministry of Education (MOE), employs both professional and non-professional personnel to teach. Every professionally trained Senior Secondary School (SSS) teacher is expected to have had at least a semester's course in educational measurement and evaluation during pre-service training and as such is expected to be guided by the basic testing principles laid down by measurement experts in his/her testing practices. However, a general observation easily reveals that the testing practices of Ghanaian SSS teachers are with lots of flaws that one begins to wonder whether training contributes to competence at all. It was based on the above problem that the researcher sought to examine the testing practices of teachers in the Senior Secondary

Schools (SSSs) in the Ashanti Region of Ghana in the light of the construction, administration, scoring of tests and the interpretation of the test results.

In a summary, the study was that of the goodness-of -fit between prescription and practice of educational achievement testing. Stated in question form, the main research problem is, to what extent do teachers in the SSSs in the Ashanti Region of Ghana adhere to the standard approved principles in classroom achievement testing?

Purpose of the Study

The study sought to determine the state of achievement testing in the SSSs in the Ashanti Region of Ghana. The study's main purpose was to find out the principles that SSS teachers in the Ashanti Region of Ghana used in the construction, administration and scoring of their classroom achievement tests and also how the results of these tests are interpreted. Knowledge of this would help establish whether the basic principles laid down by testing experts are being followed by these teachers in their testing practices. The study also sought to find out whether any differences existed between teachers who received instruction in educational measurement and those who did not, in terms of the construction, administration and scoring of tests and the interpretation of the test results.

Research Questions

In order to achieve the purpose of the study, the following research questions were answered under the four aspects of the problem under study. That is, test construction, administration, scoring and interpretation of test results.

1. Which principles do SSS teachers use in the construction of their classroom achievement tests?
2. What differences exist between SSS teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement in terms of the principles used in test construction?
3. Which principles do SSS teachers use in the administration of their classroom achievement tests?
4. What differences exist between SSS teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement in terms of the principles used in test administration?
5. Which principles do SSS teachers use in the scoring of their classroom achievement essay-type tests?
6. What differences exist between SSS teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement in terms of the principles used in test scoring?
7. How do SSS teachers interpret the results of their classroom achievement tests?

8. What differences exist between SSS teachers who received training in educational measurement and those who did not receive training in educational measurement in terms of how they interpret their test results?

Significance of the Study

Given the extent of prevalence of classroom achievement tests in Ghanaian schools and the variety of uses to which the results from these tests are put, there was the need for research into the testing practices of teachers.

The facts gathered by the study would help determine the state of affairs with respect to achievement testing in the Ghanaian educational system. This, it is believed, would help teachers who received instruction in educational measurement to be up and doing and put their acquired knowledge into practice since testing principles were related to practice throughout the study.

Also various weak spots have been identified in the testing practices of teachers in the Ashanti Region. Positive suggestions have been given as a means of addressing these flaws. It is hoped that these suggestions would help all teachers to improve on their testing practices.

Finally, it is hoped that the study would serve as an important reference source for students and teachers in achievement test development and also be supplementary to studies already undertaken in this area of achievement testing practices in Ghana.

Delimitations of the Study

The study was confined to teachers in the SSSs in the Ashanti Region of Ghana only. The study covered teachers teaching Core Mathematics, English Language and Integrated Science in Forms One, Two and Three in 26 of the 82 government assisted SSSs in the Ashanti Region.

Form One was chosen because it is the entry point of senior secondary education and a step above the Junior Secondary School (JSS) and one would be interested in the way teachers and students cope with testing. Form Two is of particular interest to the researcher. This is the midpoint of the senior secondary educational system and after the initial preparatory and adjustment stage in Form One, the researcher wanted to find out what happens there in terms of testing. Form Three was included in the study because here, students are being prepared for the SSSCE and the researcher was interested in finding out how teachers go about testing their students to achieve WAEC standards. Forms One, Two and Three constitute the entire SSS educational system.

The study was confined to teachers of English Language, Core Mathematics and Integrated Science. These subjects were chosen for the study because they are core subjects that are offered by all SSS students and also lend themselves to both objective-type and essay-type testing.

Furthermore, the study was confined to the construction, administration and scoring of classroom achievement tests and the interpretation of the results of these tests. These are the four main aspects of classroom achievement test development.

SSSs all over the country use classroom achievement tests. But the study covered only 26 schools selected randomly from the Ashanti Region.

This restriction was dictated by time and financial constraints on the part of the researcher.

Organisation of the Rest of the Report

Chapter two of this report centres on the literature related to the study. The literature entails both theoretical and empirical reviews. Chapter three describes the methodology adopted for the study. It examines the population, the sample, the sampling technique for the study, the research design, the research instruments, the pre-testing procedure, the validity and reliability of the instruments, the data collection procedure and the analysis of data.

Chapter four presents the research results and discussion of the findings in relation to the reviewed literature. Chapter five gives relevant conclusions and recommendations based on the research findings.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

Introduction

In this chapter, relevant literature has been reviewed. The review was organised under two broad sections, which are theoretical and empirical. Under the theoretical review, literature was reviewed on the following subtopics:

1. Philosophical foundations of testing
2. Historical development of testing
3. Opposition to testing
4. The need for tests
5. Classroom achievement tests
6. Construction of classroom achievement tests
7. Administration of classroom achievement tests
8. Scoring of classroom achievement tests
9. Interpretation of the results of classroom achievement tests
10. Summary

The empirical review on the other hand dealt with the review of available related research findings on the construction, administration and scoring of classroom achievement tests and the interpretation of the test results.

Theoretical Review

Philosophical Foundations of Testing

Philosophically and scientifically, the act of intellectual assessment is a quest for truth and reality (Flanagan, Genshaft & Harrison, 1997). According to Messick (cited in Flanagan et al., 1997), the act of intellectual assessment is a means by which the examiner's hypotheses are identified and then tested within the context of the scientific method. In substantiating the assertion above, Flanagan et al. (1997) stated that, "in the spirit of a true Cartesian philosophy, if the method of inquiry can be made correct, truth will reveal itself; in this case the true pattern of an individual's underlying skills and abilities" (p. 8). They continued that intellectual assessment represents a key factor in both the applied and theoretical sides of psychology's quest for understanding human intellectual functioning.

On the issue of whether test results represent the reality of an individual's underlying abilities, the work and influence of such prominent classical philosophers as Socrates, Plato and Aristotle are most profound. For instance, for Plato, authentic knowledge is only made possible through a "systematic, coherent account of reality in which each conclusion is rationally justified and that what is particular, observable and concrete must be understood in terms of higher-level principles that are comprehensive, theoretical and abstract" (Ittenbach & Lawhead, cited in Flanagan et al., 1997, p.19).

It is worthy to note that at the most basic level, it is this quest for discovering the fundamentals of truth and reality that marked the starting point of all scientific thought.

Historical Development of Testing

Those who do not take cues from history are often compelled to repeat the mistakes of the past. Consequently, in reviewing literature on the state of the art of the assessment of human cognitive abilities, it is appropriate to look back on some of the forces that have shaped the development of these measures of intellectual ability with the view to understanding why they have the form and substance they do have.

The attempts to measure human cognitive abilities can be traced to a time early in the history of imperial China (DuBois, cited in Anastasi, 1982; DuBois, cited in Cunningham, 1986; Ebel, cited in Amedahe, 1989; Flanagan et al., 1997). According to Flanagan et al. (1997) and DuBois (cited in Anastasi, 1982 & Cunningham, 1986), because the Chinese had no hereditary aristocracy, they developed a standardised civil service testing programme as far back as 2200 B.C and this programme lasted for about 4000 years. It was, however, discontinued when Alfred Binet displayed his scale for measuring intelligence in 1905. Flanagan et al. and Dubois pointed out that the tests covered the examinee's knowledge of civil law, military affairs, agriculture, revenue and geography and that civil servants were tested every three years for the purposes of initial appointments and continuance in employment.

In the West, in England, civil service ability testing was adopted during the middle portion of the 19th century (Cunningham, 1986; Flanagan et

al., 1997). Cunningham (1986) continued by noting that the Chinese method of selecting government employees was used as a basis for the establishment of the Indian civil service. He concludes that "the first British civil service commission was set up in 1850" (p. 3).

In the USA, testing began in the later part of the 19th century (DuBois, cited in Cunningham, 1986; Flanagan et al., 1997). Dubois pointed out that following the successful use in England of the Chinese method of selecting government employees, the method was adopted in the USA. He pointed out that the first civil service was established in 1883.

Formal testing in schools (paper and pencil tests) began with the introduction of paper in the 12th century (Dubois, cited in Cunningham, 1986). According to Cunningham (1986), assessment by means of written tests was first used by the Jesuits at St Ignatio. He noted that the development of academic tests was pioneered in Britain, particularly in the University of London. Under its initial charter, testing and awarding of degrees were recognised as a legitimate basis for decision making. It is worth noting however, that, prior to this period, academic testing (oral testing) in schools had already begun. As stated by DuBois (cited in Anastasi, 1982), among the ancient Greeks, testing was an established adjunct to the educational process. Tests were used to assess the physical as well as intellectual skills. Anastasi (1982) pointed out that the Socratic method of teaching with its interweaving of testing and teaching has much in common with today's programmed learning. On the account of Ebel (cited in Amedahe, 1989) and Anastasi (1982), from their beginnings in the Middle Ages, European universities relied

on formal examinations in awarding degrees and honours. These examinations, however, were largely oral.

Test development like many other aspects within psychology and education is a product of many contributors and disciplines throughout history. Notable among the early thinkers were the following personalities.

Charles Darwin (1809 − 1882), a trained physician and later a clergyman, published the book "The Origin of the Species" in 1859. He was an important factor in the increased acceptance of individual differences (Cunningham, 1986; Flanagan et al., 1997).

British mathematician and physicist, Sir Francis Galton (1822−1911), is generally recognised as the founder of formal testing. Galton's most important contributions were his emphasis on individual differences which is the corner stone of the field of psychological measurement, his initial attempts to establish norms and standard scores, and his laying of the foundation for the development of the correlation coefficient. Credit for coining the term "mental test" in 1890 however, is given to the American psychologist, James Mckeen Cattel (1860−1944). Galton and Cattel worked together to propel the field of mental testing forward in large and definable units (Anastasi, 1982; Cunningham, 1986; Flanagan et al., 1997).

A mathematician and a statistician of the first order, Karl Pearson (1857−1936), who was a student of Galton needs to be acknowledged. He derived the mathematical underpinnings of regression (then referred to as reversion), correlation, and covariation of observable phenomena in a manner that allowed Galton to make inferences about unobservable phenomena (Anastasi, 1982; Cunningham, 1986; Flanagan et al., 1997). It is worthy of

noting that the correlation coefficient of Galton and Pearson continues to be used as the basis for reliability and validity coefficients in educational and psychological testing today.

Alfred Binet (1857 − 1911), a French psychologist, developed the first intelligence test that measured high level mental functioning called the Binet-Simon test in 1905 together with Theodore Simon (1872 − 1961). Later, he developed two additional scales, the 1908 and 1911 scales (Binet & Simon, cited in Anastasi, 1982; Cunningham, 1986; Amedahe, 1989; Schultz & Schultz, cited in Flanagan et al., 1997).

Louis Terman is credited with the modification of the Binet-Simon tests in 1916 and coming out with the Stanford-Binet test which was the first well-standardised and carefully developed intelligence test. With the ongoing development in the field of measurement at the time, the use of the Stanford-Binet test as an individual intelligence test declined after the introduction of the Wechsler tests developed by David Wechsler (1896 − 1981) in 1939 (Cunningham, 1986; Wechsler, cited in Flanagan et al., 1997).

During World War I, Arthur Otis (1886 − 1964), under the tutorship of Louis Terman in 1917, developed the first group tests of intelligence which were used to screen recruits for intellectual fitness. Arthur Otis is further credited with the design and introduction of multiple-choice and other objective-test items (Anastasi, 1982; Cunningham, 1986; Flanagan et al., 1997). It is worth noting that achievement testing in Ghanaian schools today involves the use of multiple-choice and other objective type tests.

Discoveries, innovations and development continued in the field of educational measurement over the years and by 1945 many of the theories and

principles used in educational testing today had been developed (Amedahe, 1989).

Opposition to Testing

According to Cunningham (1986), opposition to testing has its roots in the philosophy of egalitarianism. This philosophy emphasises on a belief in equal treatment for all, with the view that there is something wrong with any practice in or out of education that emphasises individual differences. There is the belief that jobs and educational opportunities are rights that should be available and accessed by every person. Cunningham (1986) continued by noting that psychological testing is undoubtedly an indispensable aspect of the educational process that seems to uncover a raw strain of competitiveness in any society. This is because depending on how one performs, it promises rewards to some, and an unhappy life for others, thereby, magnifying the social inequalities that already exist. For the fact that differences in ability exist, testing makes them more pronounced.

Flanagan et al. (1997) have also pointed out that one of the most frequently voiced criticisms of psychological tests is that they are simplistic and do not reflect the way the human mind really works. They have argued that "intelligent behaviour is much more than making marks on a piece of paper, answering vocabulary questions, solving matrix analogies, solving number series problems or performing any of the other tasks that are used as indicators of intellectual functioning" (p. 8). To this end, the plausible deduction is that decisions based on test results are not likely to be wholly valid.

The possibility of bias in tests has been a major concern since the days of early test developers (Ebel & Frisbie, 1991; Flanagan et al., 1997; Joint Committee on Standards for Educational and Psychological Testing [JCSEPT], 1999; Nitko, 1996). The above mentioned testing authorities contended that standardised tests in particular have been attacked for their alleged bias against racial or ethnic minorities in the USA in particular, and also against either males or females or students with poor reading skills. Flanagan et al. (1997) have pointed out that contemporary concerns with bias have become much more sophisticated, but the problem was recognised earlier. Flanagan et al. (1997) have, however, argued that during the period between 1925 and 1975, the psychological community responded to the criticisms with increased efforts at self-regulation and a careful and extended debate over the issue of test bias. Self regulation in this respect took the form of the development of professional guidelines for test development and use, which were revised in 1985 and most recently revised in 1999.

The Need for Tests

Despite the massive opposition to testing, stemming largely from the philosophy of egalitarianism and the fact that both the construction and taking of tests are rather unpleasant activities, tests are indispensable in both the educational and non-educational settings. According to Nunnally (1964), "tests supply some very important information that would be difficult, if not impossible, to obtain from any other means" (p. 92). The need for tests can be categorised into educational and non-educational needs.

Educational uses of tests

Nitko (1996) and Amedahe and Gyimah (2003) have classified the educational uses of tests into instructional management decisions, selection decisions, classification decisions, placement decisions, counselling and guidance decisions, and credentialling and certification decisions.

The instructional management decisions refer to all the classroom decisions taken by the teacher on the basis of the assessment results of students. Firstly, tests provide useful information for instructional diagnosis and remediation. The classroom teacher constantly needs to diagnose his instruction and remediate the aspects which have been defective (Amedahe & Gyimah, 2003). This is made possible through feedback from students to the teacher. In instructional diagnosis and remediation, the teacher engages in diagnostic testing to identify which students need remedial help or special attention. According to Nitko (1996), diagnosis involves identifying both the appropriate content and the features of the learning activities in which a student should be engaged to attain the learning target.

Secondly, tests are used in the modelling of learning targets. According to Nitko (1996), "assessments define for students what the teacher wants them to learn." (p. 9). He continued by noting that students can always compare their current performance on the learning targets with the desired performance. The teacher can then teach his students to detect the ways in which their performance is matching the desired performance and the ways in which it is deficient. In this way, the teacher can direct his teaching on the remediation of any identified deficiency and students are also able to know what is important

to learn once they are able to evaluate their own performance vis-à-vis the desired learning targets.

Thirdly, tests are needed for the provision of motivation for students, rewarding those who have prepared well in advance and providing negative consequences for those who have not prepared well. The frequency of an individual's behaviour is increased by reinforcement. Hence, it can be reasonably concluded that tests cause students to study more in the sense that the motivation derived from tests as a result of performing well can activate and direct their learning by sustaining their interest (Cunningham, 1986; Ebel & Frisbie, 1991; Gronlund, 1988; Nitko, 1996).

Fourthly, tests are used for the assignment of grades to students. The grades or symbols (A, B, C,…) that the classroom teacher reports, represent his /her formal evaluation or judgement of the quality or worth of his/her students' achievement of the important learning objectives (Amedahe & Gyimah, 2003; Ebel & Frisbie, 1991; Nitko, 1996). It is worth noting that assessment results of which test results constitute the most important part as it is in the Ghanaian educational system provide the basis for the assignment of grades. Ebel and Frisbie (1991) have cautioned here that to serve effectively the purpose of stimulating, directing and rewarding students' effort to learn, grades must be valid. To achieve this, the highest grades must go to those students who have demonstrated the highest level of achievement with respect to the course objectives.

On the issue of selection decisions, sometimes, an institution decides whether some persons are acceptable for specific programmes while others are not. Those not acceptable are rejected and are no longer the concern of the

institution (Amedahe & Gyimah, 2003; Cronbach, 1960; Nitko, 1996). An educational institution often uses test results to provide part of the information on which selection decisions are based. Typical examples are the selection of candidates for admission into SSSs in Ghana which is based on the test scores of students at the end of the Junior Secondary School and university admissions in Ghana which are based on the test scores of students at the end of the SSS.

Tests provide the basis for the grouping of children with reference to their ability to profit from different types of school instruction and the identification of the intellectually retarded and the gifted (Cunningham, 1986). Nitko (1996) has pointed out that sometimes, based on test results, a decision is made that result in a person being assigned to one of several different but unordered categories of programmes. According to Cronbach and Glaser (cited in Nitko, 1996), these types of decisions are called classification decisions. These decisions result in either assigning students in the same classroom to different groups for effective instruction or assigning students to special education classes. Cunningham (1986) however cautioned test users about the over reliance on test results in assigning students to special education classes by pointing out that "intelligence tests are only one component of the assessment of students referred for possible placement in special classes" (p. 11).

On the issue of placement decisions, Cronbach (1960); Kubiszyn and Borich (1984) and Nitko (1996) have pointed out that placement decisions are made after an individual has been accepted into an educational programme. Cronbach, Kubiszyn and Borich, and Nitko, continued by noting that

placement decisions basically involve using assessment results or test data to determine where in a programme an individual is best suited to begin work. Such decisions are characterised by assigning individuals to different levels of the same general type of instruction or education based on their ability, with no one rejected by the institution (Cronbach & Glaser, cited by Nitko, 1996). Promotion in Ghanaian schools from one class or form to another which in most cases is based on the performance in tests of the previous class is an example of a placement decision.

Counselling and guidance decisions involve using assessment results, with test data inclusive, to help students in exploring and choosing careers and in directing them to prepare for the careers they select (Anastasi, 1982; Amedahe & Gyimah, 2003; Kubiszyn & Borich, 1984; Nitko, 1996;). Amedahe and Gyimah (2003) have explained that guidance is one of the students' personnel services provided in a non-instructional setting to cater for the needs of students including educational, emotional, and moral and adjustment needs. Nitko (1996) and Amedahe and Gyimah (2003) have agreed with the fact and argued that due to the complexities involved in guidance and counselling decisions, test data must always be combined with other assessments such as interviews, interest inventories, various aptitude tests and personality questionnaire together with additional background information on students and discussed with students in a series of counselling sessions in order to help students make good decisions.

On credentialling and certification decisions, Nitko (1996) and Amedahe and Gyimah (2003) explained that they are concerned with assuring that a student has attained a certain standard of learning. Credentialling and

certification may be mandated by state legislation as in the USA and executed by an external examining body at the state level. In Ghana, certification and credentialling of students is done by the WAEC. With the introduction of the practice of continuous assessment as a result of the educational reforms in 1987, Ghanaian classroom teachers contribute 30% of the total marks for certification of students at the JSS and SSS levels (Amedahe, 2000; Pecku, 2000).

Non-educational uses of tests

According to Anastasi (1982), one of the first problems that stimulated the development of psychological tests was the identification of the mentally retarded. Over the centuries the uses of tests have been quite diverse with various non-educational applications.

Anastasi (1982) has pointed out that non-educational uses of tests include clinical applications in the area of the examination of the emotionally disturbed, the delinquent and other types of behaviour deviants. According to Cronbach (1960), clinical uses of tests are mainly found in the diagnosis and classification of mental patients to determine the type of treatment suitable for them.

The selection and classification of industrial personnel represent another major non-educational application of tests (Anastasi, 1982; Cronbach, 1960). Anastasi (1982) claimed that from the assembly-line operator to top management, tests have proved helpful in such matters as hiring, job assignment, transfer, promotion or termination. According to Cronbach (1960) and Anastasi (1982), testing constitutes an important part of the total personnel

programme. A typical example is the application of psychological testing in the selection and classification of military personnel worldwide. Anastasi (1982) argued that from simple beginnings in World War I, the scope and variety of psychological tests employed in military circumstances underwent a phenomenal increase during World War II.

Cronbach (1960), however, asserted that where people are assigned to different levels of work, rather than to distinctly different types of work, the decision becomes a placement decision. This is exemplified in a case of choosing officer candidates from among enlisted men where men, not chosen as officers, remain in the army and are assigned other duties. This is a placement decision.

Tests are also used in counselling in non-educational settings. Anastasi (1982) claimed that the use of tests as an integral part of the information gathering process in counselling has broadened in scope. She continued by noting that from a narrowly defined guidance, concerning educational and vocational plans, tests are now used in all aspects of the person's life such as the emotional well-being, effective interpersonal relations, self understanding and personal development.

Psychological tests are currently being used in the solution of a wide range of practical problems including basic research (Anastasi, 1982; Cronbach, 1960). "Nearly all problems in differential psychology, for example, require testing procedures as a means of gathering data" (Anastasi, 1982, p. 4). Cronbach (1960) and Anastasi (1982) pointed out again that psychological tests provide standardised tools for the evaluation of treatments such as the outcomes of psychotherapy, the impact of community programmes

and the influence of environmental variables on human performance. Further, in industry, questions about treatment or management can be decided by suitable tests. The effectiveness of training can be judged by performance tests while supervision and personnel policies can be judged by tests of attitudes and morals.

Classroom Achievement Tests

Classroom achievement tests are generally teacher-made tests (McDaniel, 1994). These tests are constructed by teachers to test the amount of learning done by students or their attainment at the end of a course unit, term or at the end of an academic year (Amedahe, 1989). According to Mehrens and Lehmann (1991), teacher-made tests usually measure attainment in a single subject in a specific class or form or grade.

The predominance of teacher-made tests in every educational set up is given credence by the conclusions of studies by Herman and Dorr-Bremme and Stiggins and Bridgeford (cited in Mehrens & Lehmann, 1991) that, in the face of the ever-increasing use of portfolios and performance tests to assess student progress, teacher-made tests are mostly the major basis for evaluating student progress in school.

The main purpose of teacher-made tests has been delineated by measurement experts (Ebel & Frisbie, 1991; Etsey, 2004; Gronlund, 1988; Kubiszyn & Borich, 1984; Mehrens & Lehmann, 1991). All these authorities have agreed with the fact that the main purpose of a teacher-made test is to obtain valid, reliable, and useful information concerning students'

achievement and thus contribute to the evaluation of educational progress and attainments for the total improvement of classroom teaching and learning.

Teacher-made tests can be classified in a variety of ways. According to Mehrens and Lehmann (1991), one type of classification is based on the type of item format used — essay-type versus objective-type. Another classification is based on the stimulus material used to present the tests to students—verbal versus non-verbal, while other classifications may be based on the purposes of the tests and the use of the test results—criterion-referenced versus norm-referenced, achievement versus performance, and formative versus summative.

The teacher-made test classification that is most popular with testing experts is the classification based on the type of item format used, which classifies tests into objective-type tests and the essay-type tests (Cunningham, 1986; Etsey, 2004; Gronlund, 1985; Nunnally, 1964; Tamakloe et al, 1986). The aforementioned testing experts have contended that essay-type tests can either be the extended or the restricted response types while objective-type tests can take the form of the short-answer, true-false, matching or multiple-choice.

Construction of Classroom Achievement Tests

The basic principles for the construction of teacher-made tests have been developed over the years by a number of educational measurement experts (Amedahe, 1989). While some of the test construction principles are general and apply to any type of test, others are specific and apply solely to the particular type of test under construction.

From available literature, the test construction principles that the researcher judged as most comprehensive and practicable in the classroom testing situation were those postulated by Tamakloe, Atta and Amedahe (1996) and Etsey (2004). These are in eight steps. The steps are:

a) define the purpose of the test,

b) determine the item format to use,

c) determine what is to be tested,

d) write the individual items,

e) review the items,

f) prepare the scoring key,

g) write directions, and

h) evaluate the test.

According to Gronlund (1988), "the key to effective achievement testing is careful planning" (p. 15). It is during the planning stage that the purpose of the test must be determined. As already pointed out in the literature, tests can be used for a number of purposes. It is worthy of note, however, that each type of test use typically requires some modification of the test design and thereby determines the type of item format to be used.

The second step of the planning stage is the determination of the item format to use. As stated earlier in the literature, the most common item formats in classroom achievement testing are the essay- and the objective-types. According to Etsey (2004), it is sometimes necessary to use more than one item format in a single test. This is because depending on the purpose of the test, one item format cannot be used exclusively to measure all learning outcomes. According to Mehrens and Lehmann (1991), the choice of an

appropriate item format depends on factors such as the purpose of the test, the time available to prepare and score the test, the number of students to be tested, the skills to be tested, the difficulty level desired, the physical facilities available for reproducing the test, the age of the students and the teacher's skill in writing the different types of items.

The final step of the planning stage is the determination of what is to be tested or measured. According to Etsey (2004), the teacher at this point should determine the chapters or units of the course content that the test should cover as well as the knowledge, skills or attitudes to be measured. Instructional objectives need to be defined in terms of student behaviours and linked to what has been stressed in class. A test plan made up of a table of specifications should be made. The table of specifications matches the course content with the instructional objectives (Etsey, 2004). With the total number of items on the test in mind, the specification table helps to avoid overlapping in the construction of the test items, helps to determine the weighting of learning outcomes with respect to content areas, and makes sure that justice is done to all aspects of the course, thereby helping to ensure the content validity of the test.

After the planning stage, actual writing of the individual test items follows. Tamakloe et al. (1996) and Etsey (2004) have pointed out that whichever test item types that are being constructed must follow the basic principles laid down for them. There are, however, general guidelines that according to Mehrens and Lehmann (1991) and Etsey (2004), apply to all types of tests. These include:

1. The table of specifications must be kept before the teacher and continually referred to as the items are written.
2. The test items must be related to and match the instructional objectives.
3. Well-defined items that are not vague and ambiguous must be formulated. Grammar and spelling errors must be checked. Textbook or stereotyped language must be avoided.
4. Excessive verbiage and complex sentences must be avoided.
5. The test items must be based on information that students should know.
6. More items than are actually needed in the test must be prepared in the initial draft. Mehrens and Lehmann (1991) suggested that the initial number of items should be 25% more while Hanna (cited in Amedahe, 1989) has suggested 10% more items than are actually needed in the test.
7. Items of varying levels of difficulty must be used. This, however, depends on the purpose of the test.
8. The items and the scoring keys must be written as early as possible after the material has been taught.
10. The test items must be written in advance (at least two weeks) of the testing date to permit reviews and editing.

After the items have been written, Tamakloe et al. (1996) call the next stage the item preparation stage. At this stage the test items must be reviewed and edited. Etsey (2004) has suggested that the items must be critically examined at least a week after writing them. He has emphasised that where

possible, fellow teachers or colleagues in the same subject area should review the test items. Reviewing and editing the items are for the purpose of removing or rewording poorly constructed items, checking difficulty level of items, checking the length of the test, and the discrimination level of the items (items must discriminate between low- and high-achievers). All test items should be checked for technical errors and irrelevant clues.

After reviews and editing, the test items can now be assembled. In assembling test items, the following points must be considered (Etsey, 2004; Kubiszyn & Borich, 1984; Mehrens & Lehmann, 1991; Tamakloe et al., 1996).

1. The items should be arranged in sections by item formats. The sections must progress from easier formats (true-false) to more difficult formats (interpretive exercises and essay).
2. Within each section or format, the items must be arranged in order of increasing difficulty. One way of achieving this is to group items in each format according to the instructional objectives being measured and make sure that they progress from simple to complex. According to Mehrens and Lehmann (1991), such a grouping has the advantage of helping the teacher to ascertain which learning activities appear to be most readily understood by students, those that are least understood and those that are in-between. According to Hambleton and Traub (cited in Mehrens & Lehmann, 1991), ordering items in ascending order of difficulty leads to better performance than either a random or hard-to-easy ordering. Lafitte (cited in Mehrens & Lehmann, 1991) on the other hand, has reported inconclusive data. Although, empirical

evidence is also inconclusive about the effectiveness of using statistical item difficulty as a means of ordering items, Sax and Cromack (cited in Mehrens & Lehmann, 1991), Mehrens and Lehmann (1991) and other testing experts have recommended that for lengthy or timed tests, items should progress from the easy to the difficult—if for no other reason than to instill confidence in the examinee, especially at the beginning. It should be noted however, that, the use of statistical item difficulty or item difficulty indexes by the classroom teacher seems impracticable to a large extent (Kubiszyn & Borich, 1984; Tamakloe et al., 1996). This is because statistical item difficulty data are always gathered after test administration or test try-outs and teacher-made test items are usually not pre-tested. Mehrens and Lehmann (1991) however, recommended that subjective judgement must be relied on to determine difficulty level of items. They have stated that "teachers could only categorise their items as difficult, average or easy" (p. 71).

3. The items must be spaced and numbered consecutively so that they are not crowded and can easily be read.
4. All stems and options must be together on the same page and if possible, diagrams and questions must be kept together.
5. If a diagram is used for a multiple-choice test, the diagram must be placed above the stem.
6. A definite response pattern to the correct answer must be avoided.

In addition to the above, Gronlund (1985) and Etsey (2004) have recommended that for objective-type tests, the options must be written vertically below the stem rather than across the page. Further, Etsey (2004)

has suggested that test items can also be arranged according to the order in which they were taught in class or the order in which the content appeared in the textbook.

After the test items have been assembled, the next task is the preparation of the scoring key, the marking scheme or the scoring rubric (Etsey, 2004). The marking scheme according to Etsey (2004) and Amedahe and Gyimah (2003), must be prepared when the items are still fresh in the teacher's mind and always before the administration of the test. This way, defective items that do not match their expected responses would be recognised and reviewed. For objective-type tests, correct responses to items should be listed. For essay-type tests, points or marks should be assigned to various expected qualities of responses. Mehrens and Lehmann (1991) have pointed out that if the teacher considers it prudent to have differential weighting for different essay questions, then factors such as the time needed to respond, the complexity of the question, and emphasis placed on that content area during the instructional phase must be considered.

Immediately following the preparation of the marking scheme is the writing of clear and concise directions for the entire test and sections of the test. Here, the time limit for the test must be clearly stated. As argued by Nunnally (1964), and Ebel and Frisbie (1991), a good working rule is to try to set a time limit such that about 90 percent of the students will feel that they have enough time to complete the test. Directions according to Etsey (2004), must include penalties for undesirable writings, number of items to respond to, where and how the answer should be written, credits for orderly presentation of material (where necessary), and mode of identification of examinees.

The last stage of the test construction process is the evaluation of the test on the criteria of clarity, validity, practicality, efficiency and fairness.

Clarity refers to how simply and clearly the items are written vis-à-vis the ability level of the testees and the material the test is measuring. It also refers to the kinds of knowledge the test is measuring and how adequately the test items relate to the content and course objectives (Amedahe & Gyimah, 2003; Etsey, 2004; Tamakloe et al., 1996).

Validity bothers on how closely the test represents the material presented in the course unit or chapter and how faithfully the test reflects the difficulty level of the material taught in class. The issue of validity here establishes the content validity evidence of the test (Amedahe & Gyimah, 2003; Etsey, 2004; Tamakloe et al., 1996).

On practicality, consideration is given to whether students will have enough time to complete the test. It also bothers on whether there are enough materials (chairs, tables, answer booklets) to present the test and complete it effectively (Amedahe & Gyimah, 2003; Etsey, 2004; Tamakloe et al., 1996).

Efficiency bothers on finding out whether the test is the best way to measure the desired knowledge, skill or attitude. Consideration must also be given to the problems that might arise due to material difficulty or shortage and these expected problems well catered for (Amedahe & Gyimah, 2003; Etsey, 2004; Tamakloe et al., 1996).

On the fairness criterion, consideration is given to whether students have been given advance notice of the test, whether students have been adequately prepared for the test, and whether students understand the testing procedures. Consideration is also given to how the lives of students are

affected as a result of the possible uses to which the test scores are put (Amedahe & Gyimah, 2003; Etsey, 2004; Tamakloe et al., 1996). After this comprehensive evaluation of the test, the test can be submitted to be processed for subsequent administration.

Administration of Classroom Achievement Tests

The guiding principle in test administration is to provide all examinees with a fair chance to demonstrate their achievement on what is being measured (Gronlund, 1985; Tamakloe et al., 1996). The need to maintain uniform conditions in test administration cannot be over-emphasised. This is especially essential for the test to yield consistent, reliable and valid scores without much influence of chance errors. This is emphasised by the JCSEPT (1999) by stating that, "reasonable effort should be made to assure the integrity of the test scores by eliminating opportunities for test takers to attain scores by fraudulent means" (p. 64). "This calls for ensuring a congenial psycho-physical atmosphere for test taking" (Tamakloe et al., 1996, p. 214). This was also emphasised by Airasian (cited in Amedahe & Gyimah, 2003) that test administration is concerned with the physical and psychological setting in which students take their tests.

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

weighting of content areas, the kinds of items (objective-types or essay-types) on the test, how the test will be scored and graded, and the importance of the results of the test.

The physical conditions that need to be in place to ensure maximum performance on the part of students include adequate work space, quietness in the vicinity, good lighting and ventilation and comfortable temperature (Etsey, 2004; Gronlund, 1985; Lindquist, cited in Tamakloe et al., 1996). Adequate work space is very essential for test administration because when tables and chairs are closely arranged together, students will not have the independence to work on their own. This will in no doubt lead to students copying from each other. In addition, tables provided for the examination must be conducive to the testing materials being used. For example, in Practical Geography examinations where topographical sheets are used, each student could use two tables or desks in order to get adequate work space (Tamakloe et al., 1996).

Noise and distraction in the testing environment should be kept at the barest minimum if not eliminated completely. Interruptions within and outside the testing room has the tendency of affecting student's performance (Mehrens & Lehmann, 1991; Tamakloe et al., 1996). Etsey (2004) has pointed out that it is helpful to hang a "Do Not Disturb. Testing in Progress" sign at the door of the testing room to warn people to keep off. Good lighting is important in effective test administration. This facilitates students' reading of instructions and test items without straining their eyes, thereby working faster (Gronlund, 1985). "Good ventilation and comfortable temperature should be assured since their absence could create unrest or uneasiness in testees making concentration difficult" (Tamakloe et al., 1996, p. 215). Other basic physical conditions are

that, all testing equipment must be in the room and readily available, and also, all possible emergencies during test administration must be expected and well catered for.

The psychological conditions in test administration, on the other hand, include the position of the invigilator, timing of the test, threatening behaviours of invigilators, and interruption to give instructions and announcements (Etsey, 2004; Bernstein, cited in Amedahe, 1989; Gronlund, 1985; Tamakloe et al., 1996). A study on the examiner as an inhibiting factor, carried out by Bernstein (1953) and reported by Amedahe (1989) found out that, the presence of the examiner tended to inhibit the performance of those students who were nervous. The crux of the matter is that if the mere presence of the examiner or invigilator could affect the performance of students who are nervous, then there is no doubt that the position of the invigilator is very significant to the performance of students on examinations. Etsey (2004) has recommended that the invigilator should stand where all students could be viewed and move among the students once a while to check malpractices. Such movements should not disturb the students. He must be vigilant. Reading novels or newspapers, making of and listening to telephone calls, dozing off and chatting are not allowed.

The timing of tests is very important. Tests must not be given immediately before or just after a long vacation, holidays or other important events where students are involved either physically or psychologically. Tests must also not be given when students would normally be doing something pleasant such as having lunch, athletics or other sporting activities as this will hamper students' concentration (Amedahe & Gyimah, 2003; Etsey, 2004).

Interruptions during testing, such as giving instruction, must be kept to the barest minimum and should always relate to the test. The time spent and time left to complete the test must be announced at regular intervals to enable students apportion their time to the test items. Where practicable, the time should be written on the chalkboard at 15-minutes intervals until near the end of the test when it could be changed every five minutes. Further, students should start the test promptly and stop on time (Amedahe & Gyimah, 2003; Etsey, 2004; Tamakloe et al., 1996).

Teachers should always work at minimising test anxiety in students during testing. They should therefore, avoid, warning students to do their best because the test is important, telling students that they must work faster in order to finish on time, threatening dire consequences of failure in the test, and threatening students with tests if they do not behave (Amedahe & Gyimah, 2003; Etsey, 2004 ; Tamakloe et al., 1996).

Scoring of Classroom Achievement Tests

If the scores of a test are to supply the needed information to students, parents, the school and other stake holders, then they must be reported in a manner that accurately communicates how well students performed (Nunnally, 1964). The statement above clearly spells out the third principal stage of the classroom testing process which is the task of scoring. Since the objective- and essay-type tests form the back bone of most teacher-made tests, it is pertinent that their scoring is examined.

According to Nunnally (1964), “scoring of objective-type tests is purely a mechanical problem which requires no special skill” (p. 140).

Nunnally continued by stating that, a scoring key stating the correct answer for each item should be available. Scoring in this case involves checking the student's response to each item with the scoring key to see whether it is correct. The simplest way to obtain a total score for individual students is to count up the number of correct answers. According to Nunnally, this would be, the number of true-false or multiple-choice items marked correctly, the number of correct matchings in matching items and the number of correct terms supplied in fill-in-items.

The scoring of essay-type tests on the other hand according to Tamakloe et al. (1996), is a highly important issue due to the fact that no matter how careful one is in writing the items, without equally taking careful steps to ensure consistency of scoring, the scores will not be reliable. The main reason for utmost care in the scoring of essay-type tests is the subjectivity involved. This is a major difference between the essay- and objective-type tests (Amedahe & Gyimah, 2003; Ebel & Frisbie, 1986; Etsey, 2004; Gronlund, 1988). The assertion above is substantiated by Kubiszyn and Borich (1984) in the words that:

> Essays tend to be difficult to score reliably. That is, the same essay answer may be given an A by one scorer and a B or C by another scorer. Or, the same answer may be graded A on one occasion, but B or C on another occasion by the same scorer! As disturbing and surprising as they may seem, these conclusions have long been supported by research findings (p. 99).

According to Ebel and Frisbie (1991) and Mehrens and Lehmann (1991), the decision on a method of scoring for essay-type tests depends to some extent on the type of score interpretation desired—norm-referenced or criterion-referenced and the amount of diagnostic information needed about individuals' responses. It also depends on the time and facilities available for reading the papers and whether the essay is of the restricted- or extended-response type.

The analytic and holistic methods are the two main methods of scoring essays. The analytic, point-score or the trait method (Nitko, 1996; Mehrens & Lehmann, 1991) basically involves the use in scoring of an already prepared list of points or ideas considered essential to a good answer to the question, together with the number of points (marks) allotted to each idea raised or discussed in the answer. This is known as a marking scheme, a scoring rubric or a scoring key. (Amedahe & Gyimah, 2003; Ebel & Frisbie, 1991; Etsey, 2004; Tamakloe et al., 1996).

Ebel and Frisbie (1991), Tamakloe et al. (1996), Amedahe and Gyimah (2003), and Etsey (2004), have contended that the analytic method is more applicable to the scoring of restricted-response essay-type tests. The main advantages of the analytic method, according to Tamakloe et al. (1996), include the fact that it can yield very reliable scores when used by a critical scorer. Also, the process of preparing the detailed marking scheme may bring to the teacher's attention errors such as faulty wording, extreme difficulty of items, and unrealistic time limits. This enables the teacher to review defective items. Again, the subdivision of the model answer makes possible the provision of valuable feedback to students concerning their strengths and

weaknesses, thereby making the method very useful, especially when criterion-referenced tests (CRTs) are employed. Finally, with the analytic method, it is easier to discuss or justify the marks or grades given to students.

Tamakloe et al. (1996) have asserted that the main disadvantages of the analytic method are that, it is very laborious and time consuming and so may be slower than the holistic method. Also, for essays that are not very well constructed, it is difficult to come out with well-defined elements in the scoring guide.

The holistic, global, rating or sorting method (Nitko, 1996; Mehrens & Lehmann, 1991) on the other hand requires the scorer to make a judgement of the overall quality of each student's response. With this method, the model answer is not subdivided into specific points or component parts. Rather, the model answer serves as a standard and each response is read for a general impression of its adequacy as compared to the standard (Amedahe & Gyimah, 2003; Etsey, 2004; Ebel & Frisbie, 1991; Mehrens & Lehmann, 1991; Tamakloe et al., 1996).

According to Mehrens and Lehmann (1991), in the establishment of the standards or anchor points upon which to judge the adequacy of students' responses, the teacher could prepare different answers corresponding to the various scale points or categories, or select students' already written papers and let the actual responses establish the various anchor points or standards. Mehrens and Lehmann (1991) have emphasised that the anchor points could be a 2-point scale—"acceptable" and "unacceptable" or a 5-point scale such as "superior quality", "above average quality", "average quality", below average quality" and "inferior quality". After the anchor points or standards have been

established, the rater does a first reading to sort the various responses into one of the various piles or categories depending on the quality of the answer in relation to the different samples. The next task, according to Mehrens and Lehmann (1991) and Etsey (2004), is a simple clerical task of assigning score values to the papers in the various categories upon a second reading of the papers in each category.

According to Nitko (1996) and Amedahe and Gyimah (2003), the advantages of holistic scoring are that it is faster than analytic scoring and also helps the rater to view the papers as a working whole. The method is also effective when large numbers of essays are to be read. The disadvantages on the other hand, are that because raters give a single overall mark and do not point out details to students concerning their strengths and weaknesses, there is no effective feedback to students to help them improve. Also, the scorer's own biases and errors can easily be masked and go unnoticed.

In order to improve objectivity in the scoring and reliability of the scores of essay-type tests, Mehrens and Lehmann (1991); Tamakloe et al. (1996); Amedahe and Gyimah (2003); and Etsey (2004) have suggested the following techniques or principles to be adopted by scorers.

1. Prepare a form of scoring guide. This could either be an analytic scoring guide or a holistic scoring guide.
2. Constantly follow the marking scheme when scoring. It is one thing deciding to score all papers uniformly using a scoring guide and actually following the scoring guide constantly to achieve uniformity. Scorers should follow the marking scheme constantly as they score, as this reduces rater drift, which is the likelihood of either not paying

attention to the scoring guide or interpreting it differently as time passes. Scorers must also avoid being influenced by the first few papers they score since this can let them become too lenient or harsh in scoring other papers.

3. Score all responses item by item rather than script by script. Here, scorers must take one item at a time and score all the responses to it throughout before going to the next item. This principle is to minimise the carryover effect on the scores and thereby ensure consistency.

4. Randomly reshuffle the scripts when beginning to score each set of items. This will minimise the bias introduced as a result of the position of one's script .Research by Hales and Tokar (cited in Mehrens and Lehmann, 1991) has shown that a student's essay grade will be influenced by the position of his paper, especially if the preceding answers were either very good or very poor. Mehrens and Lehmann (1991) have pointed out that randomly reshuffling of scripts is especially significant when teachers are working with high- and low-level classes and read the best scripts first or last.

5. Score the scripts anonymously. Scripts should be identified by code numbers or any other means instead of the names of students. This principle is to reduce the "halo-effect". This happens when a scorer's general impression of a person influences how he scores his paper.

6. Keep previously scored items out of sight when scoring the rest of the items. This principle is to minimise the carryover effects and ensure consistency of the scores.

7. Try to score all responses to a particular item without interruption. This is to avoid unreliability of the scores as a result of the grader's standards varying markedly due to excessive interruptions in the course of scoring.
8. Score essay-type tests only when you are physically sound and mentally alert. This is to say that essays must be scored at a congenial time. This is because "it is known that consistency in scoring essay tests is a function of the time the paper is scored" (Tamakloe et al., 1996, p. 251). Over excitement, depression, and any type of psychological or mental disequilibrium will affect the consistency of the scores of essay-type tests.
9. Comments should be provided and errors corrected on the answer scripts for students to facilitate learning. This is especially important in formative assessments where the comments should be on students' weaknesses and strengths in answering various items.
10. The mechanics of expressions such as correct grammar usage, flow of expression, quality of handwriting, orderly presentation of material and spelling should be judged separately from subject matter correctness.

Interpretation of the Results of Classroom Achievement Tests

The last stage of the classroom achievement testing process is the interpretation of the results of the test in order to make them meaningful and achieve their intended use. "After a measure of achievement has been obtained, the results need to be put in a form that is easily interpretable" (Gronlund, 1988, p. 155).

There are two main types of test-score interpretations which are the norm-referenced interpretation (NRI) and criterion-referenced interpretation (CRI) (Amedahe & Gyimah, 2003; Etsey, 2004; Ebel & Frisbie, 1991; Gronlund, 1988; and Tamakloe et al., 1996).

Norm-referenced interpretation

In norm-referenced interpretation (NRI), a student's assessed performance is described in terms of his/her position in a reference group that has been administered the assessment (Nitko, 1996). In other words, a student's level of performance is described in relation to that of other members of the class. An example is reporting that a student's performance on a test is better than 75 percent of the class or a student places 12th out of 50 students in a class. In these examples, the reference group is the class and it is called the norm group.

According to Gronlund (1988), for comparison purposes, it is common for the classroom teacher to use the total raw score of each student to do a simple ranking of the raw scores. Scores that have been derived from the raw scores such as percentile ranks and stanines can also be used. In simple ranking of raw scores, the individual scores are arranged in rank order from high to low together with a frequency count to show the number of students earning each score. This method has a major limitation when it is used to communicate the test results to others. This is because how good a student's performance is, depends on the group size which if not quoted together with the rank, the rank alone is meaningless.

A percentile rank indicates a student's relative position in a group in terms of the percentage of group members scoring at or below the student's score (Gronlund, 1988; Nitko, 1996). For example, if a raw score of 35 equals a percentile rank of 70, it means that 70 percent of the group members had raw scores equal to or lower than 35. According to Gronlund (1988), the use of percentile ranks enables raw scores to be put on a scale that has the same meaning with different group sizes and that is readily understood by test users. Gronlund (1988) and Nitko (1996) have given a simple formula for converting raw scores to percentile ranks (PR). It is given by:

$$PR=\frac{\text{Number of students below score} + \frac{1}{2}(\text{Number of students at score})}{\text{Number in Group (N)}} \times 100\%$$

The stanine scale is a system of standard scores that divides the distribution of raw scores into nine parts. The lowest stanine score is 1, the highest is 9, and stanine 5 is the median score, located at the centre of the distribution. Stanines are normally distributed standard scores with a mean of 5 and standard to deviation of 2 (Ebel & Frisbie, 1991; Gronlund, 1988; Nitko, 1996). According Gronlund (1988) and Nitko (1996), the percentage of a group that falls within each stanine in a normal distribution is given by Table 1 below.

Table 1

Percentage of a Group in Each Stanine in a Normal Distribution

Stanine	1	2	3	4	5	6	7	8	9
Percentage	4	7	12	17	20	17	12	7	4

Gronlund (1988) continued by noting that the classroom teacher can simply assign the top 4% of the students a stanine 9, the next 7%, stanine 8, the next 12%, stanine 7, and so on.

The stanine system of standard scores provides a standard scale or a common yardstick by which scores on different tests by one group or different groups may be compared reasonably (Ebel & Frisbie, 1991; Gronlund, 1988). For example, students' score on classroom tests may be compared readily with their standing on standardised tests of achievement.

Criterion-referenced interpretation

Criterion-referenced interpretation (CRI) describes assessed performance in terms of the kinds of tasks a student with a given score can do (Glaser & Nitko, cited in Nitko, 1996). An example is reporting that a student can spell correctly 75% of the technical terms on 'respiration' in biology or a student can solve 15 out of 20 problems on factorisation of quadratic expressions in SSS One.

According to Gronlund (1988), for formative evaluation purposes in the classroom, tests that are termed criterion-referenced tests (CRTs) are commonly used to measure mastery of instructional objectives or learning targets. Gronlund has emphasised the need of some standard or cut-off score by stating that "when CRTs are used to distinguish between those who have mastered a given set of tasks and those who have not, some standard or cut-off score must be set" (p. 119). He continued by pointing out that the percentage-correct score is the most widely used method of judging whether learning targets have been mastered, in reporting the results of classroom CRTs.

According to Gronlund (1988), setting standards of acceptable performance on a CRT is difficult and frustrating due to the fact that many issues are involved and there are few clear guidelines to follow. Gronlund continued by suggesting that the teacher must rely on judgement based on his own teaching experience, and if a team is involved, on the experience of colleague teachers. "A relatively simple and practical procedure is to arbitrarily set a standard and then adjust it up or down as various conditions and experiences are considered" (Gronlund, 1988, p. 119). An example is shown below:

1. Set the performance standard or cut-off score of a multiple-choice test at 85 percent correct.
2. Increase the level if test is short.
3. Increase the level if essential for next stage of instruction.
4. Decrease the level if tasks have low relevance.
5. Decrease the level if tasks are extremely difficult.
6. Adjust the level up or down as teaching experience dictates.

According to Gronlund (1988) and Nitko (1996), both methods of test score interpretation are important to understand how well students are learning. NRI tells how an individual's test performance compares with that of others while CRI tells in specific performance terms what an individual can do without reference to the performance of others, so that if necessary, remedial work can be planned. Gronlund (1988) and Nitko (1996) have argued that since the terms norm-referenced and criterion-referenced only refer to the method of interpreting test scores, both types of interpretation could be applied to the same test. It could be reported that John did better than 90% of the

students (norm-referenced interpretation) by solving 15 of the 20 problems in Algebra (criterion-referenced interpretation).

Gronlund (1988), and Ebel and Frisbie (1991) have, however, argued that the two types of interpretations are likely to be most meaningful when the test is specifically designed for the type of interpretation desired. They continued that norm-referenced interpretation is facilitated by tests that provide a wide spread of scores so that reliable discrimination will be possible among students of different levels of achievement. This is done by using test items of average difficulty. Such tests give rise to the term norm-referenced tests (NRTs). Criterion-referenced interpretation on the other hand, is aided by tests that comprise items that are directly relevant to the learning outcomes being measured, irrespective of the difficulty level of the items. Such tests give rise to the term criterion-referenced tests (CRTs).

Gronlund (1988) has cautioned further that if the two types of interpretations are to be combined, then it is most likely to be effective where NRI is added to the performance description of a CRT. For example, it could be reported that, a student can find the product of two binomials and that only 20 percent of SSS One students can do this. Gronlund continued that when CRIs are added to tests designed for NRI, the descriptions of student performance are likely to be inadequate. This is because NRTs typically cover a wide range of learning outcomes with few items per outcome and so the performance descriptions will tend to be sketchy and unreliable.

Empirical Review

Not much has been done in the area of research into testing practices in Ghana. Readily available studies are those of Amedahe (1989) on the testing practices of secondary school teachers in the Central Region of Ghana and Quaigrain (1992) on teacher competence in the use of essay-type tests in the Western region of Ghana. According to Amedahe (1989), the problem of insufficient study in the field of classroom achievement testing appears to exist even in the advanced countries like the USA because the emphasis is rather laid on standardised testing. This claim was confirmed by the research of Gullickson and Ellwein (cited in Amedahe, 1989).

Testing Practices of Teachers in the United States of America

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

The first is a study to assess the testing skills and practices of 326 elementary and secondary school teachers in Ohio which was undertaken by Marso and Pigge (1989). The assessments included direct analysis of teacher-made tests as well as perceptual assessments of teachers' testing needs and proficiencies.

It was generally agreed that the testing proficiencies of teachers in the study were not adequate to meet classroom needs. Analyses revealed that typical teachers gave 50 or more formal teacher-made tests each year, for which they wrote most of their own questions. Matching exercises on teacher-made tests were particularly prone to error. Most teacher-made tests, except in mathematics and science, functioned at the knowledge level. Administrators'

and teachers' perceptual assessments of teachers' testing skills were negatively correlated with the results of direct analysis of teacher testing skills as displayed in their teacher-made tests.

Approximately 225 studies addressing the knowledge and skills of classroom teachers from kindergarten through grade 12 related to the development and use of teacher-made tests were reviewed by Marso and Pigge (1992). The following are the summary of the findings from the reviews concerning inadequacies in teachers' testing knowledge and training.

a) Limited expertise, support, and pre-service and in-service training are available to assist teachers in meeting their testing responsibilities.

b) Teachers view teacher-devised testing as positively influencing instruction and learning.

c) Most teacher-constructed tests contain many faults, and function almost exclusively at the recall level.

d) Teachers typically do not use test improvement strategies such as test blueprints and item analysis.

The third is a study to analyse seventh- and eighth-grade teachers' classroom tests in science and mathematics undertaken by McMorris and Boothroyd (1992). The major findings of the study were that, teachers used all major item formats. Science teachers favoured multiple-choice and mathematics teachers favoured computation items. Faults were found in 35% of completion items and 20% of multiple-choice items. Average test quality on 30 evaluative items was 5.4 on 7-point semantic differential scales. The quality of a teacher's test was best predicted by performance on a multiple-choice measurement competency test. The best predictor of the quality of test

variable was the score on the measurement competency test (r = 0.37). Test quality was also associated with the ability to detect item faults and self-report adequacy of measurement training.

The policy information report by Barton and Coley (1994) which provides a profile of state testing programmes between 1992 and 1993 as well as a view of classroom testing practices by states, school districts, schools or teachers gives another synopsis of the testing practices of teachers in the USA. Firstly, the report stated that the multiple-choice test remains dominant at the state level. In the classroom in contrast, non-multiple-choice items appear to be the predominant mode.

Secondly, on item type and the frequency of testing, with multiple-choice tests, the report stated that, at the fourth grade, six percent of students have teachers who give multiple-choice tests once or twice a week, 43% once or twice a month, and 51% yearly or never. The comparable percentages for eighth graders are 4, 30 and 66.

With problem sets, about half of the fourth graders have teachers who use problem sets once or twice a week, 39% once or twice a month, and nine percent yearly or never. The comparable percentages for eighth graders are 58, 32 and 10.

With written responses, 44% of fourth graders are given tests requiring written responses at least monthly, 16% once or twice a year and 40% never or hardly ever. For eighth graders, the comparable percentages are 44, 22 and 33.

With projects, portfolios and presentations, 20% of fourth graders are given these forms of assessment at least monthly, 25% once or twice a year

and 54% are never or hardly ever given these assessment types. For grade eight, the comparable percentages are 21, 32 and 47.

On testing equity, the report concluded that the patterns of traditional and alternative testing in the classroom are similar for students of different races, ethnicity, ability groups and resource adequacy.

Another report by the American Association for the Advancement of Science (AAAS, 1998) in describing the current assessment practices in the United States of America pointed out that, teacher-made tests are often as limited in measuring student thinking as their standardised counterparts (Stiggins & Conklin, cited by AAAS, 1998). The reasons are as follows.

First, teacher-made tests are mostly short-answer or matching items that place far more emphasis on students' recall than on students' thinking ability. Second, evidence suggests that because teachers do not receive proper training in effective assessment methods, they tend not to change the assessment methods they use as assessment needs change. Different assessments are needed to measure performance, effort and achievement, for instance, but teachers tend to use the same type of assessment, mainly tests, to measure all three. Third, because of limited time, teachers usually use the assessments that are found at the end of textbook chapters. These assessments include mostly short answer questions that call for only low level thinking skills and simple recall of factual knowledge (Centre for the Study of Testing, Evaluation & Educational Policy [CSTEEP], cited by AAAS, 1998).

The report further stated that even if teachers receive the training, time and resources that would allow them to widen their assessment practices, students themselves may be a barrier. Students, especially high school students

who have become test-wise sometimes oppose the more labour intensive format of assessments that entail performance tasks, answering essay-type items or providing possible solutions to open-ended items. Parents have also become used to report cards that contain letters and percentages and may question new approaches that are not clearly explained and justified.

The sixth is a study to assess teacher-made tests in secondary science and mathematics classrooms which was undertaken by Oescher and Kirby (1998) and published by the Educational Resources Information Centre (ERIC). The study covered the nature of classroom assessment, characteristics of teacher-made tests, item format, cognitive levels, quality of test items and teachers' confidence in testing skills. The results of the study indicated that the main areas where teachers lacked competence were the use of tables of specifications, development of higher order items, item formatting, and empirical analysis of test results.

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

Testing Practices of Teachers in England

In England, a review of research findings of a number of research studies by Crooks (1998) and Black (1993b) generally revealed an overall picture that was one of weak practice.

First, classroom evaluation practices generally encouraged superficial and rote learning, concentrating on recall of isolated details, usually items of knowledge which pupils soon forget. Second, teachers generally do not review the assessment questions that they use and do not discuss them critically with peers, so there is a little reflection on what is being assessed. Third, teachers over-emphasise the grading function while the learning function is under-emphasised. Fourth, there is a tendency on the part of teachers to use normative rather than a criterion approach which emphasises competition between pupils rather than personal improvement of each student. The evidence is that with such practices, the effect of feedback is to teach the weaker pupils that they lack ability, so that they are de-motivated and lose confidence in their own capacity to learn.

According to Black and William (1998), more recent researches have confirmed this general picture. Both in questioning and written work, teachers' assessments focus on low-level aims, mostly recall. There is little focus on such outcomes as speculation and critical reflection (Bol & Strage, 1996; Pijl, 1992; Senk, Beckman & Thompson, 1997, cited by Black and William, 1998). Students focus on getting through the tasks and oppose attempts to engage in risky cognitive activities (Duschl & Gitomer, cited by Black and William, 1998). Black and William (1998) added that, although teachers can foretell the performance of their students on external tests, their own tests do not tell them

what they must know about their students' learning (Lorsbach, Tobin, Briscoe & Lamaster, cited by Black and William, 1998).

Testing Practices of Teachers in Ghana

On test construction, the study of Amedahe (1989) showed that to a great extent, secondary school teachers in the Central Region did not follow the basic prescribed principles of classroom test construction. He also found out that there was no significant difference between the procedure used in constructing classroom achievement tests by teachers who received instruction in testing and those who did not, in terms of the accuracy of following prescribed test construction principles. He discovered further that there was a moderate relationship between the years of teaching experience and the accuracy with which teachers constructed their classroom achievement tests. The more experienced the teacher was in teaching, the more accurate he/she became in constructing his/her achievement tests.

Quaigrain (1992), on the other hand, found out that majority of the teachers in the study did advance planning of essay-type tests. This finding does not support wholly the first finding of Amedahe that to a large extent, teachers did not follow the basic prescribed principles in classroom test construction. Another finding of Quaigrain (1992) was that while some teachers reviewed their essay-type tests items, others did not review them. He found also that majority of the teachers did not indicate the score points which each item attracted on the question paper to guide students. Lastly on test construction, Quaigrain (1992) found out that majority of the teachers prepared their marking scheme after the examination while few prepared their

marking scheme before the test was taken. These findings of Quaigrain generally support the first finding of Amedahe that to a great extent, the teachers in the study did not follow the basic prescribed principles of classroom test construction.

On whether pre-service training in testing held anything good for actual testing practice, Quaigrain (1992) found that there was a significant positive relationship between pre-service training in educational measurement and competence in the use of essay-type tests. The point-biserial correlation coefficient (r_{pbis}) was 0.43. This finding, however, does not support the second finding of Amedahe that there was no significant difference between the procedure used in constructing classroom achievement tests by teachers who received instruction in testing and those who did not, in terms of the accuracy of following prescribed test construction principles.

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

On test administration, Amedahe (1989) found that, teachers in the study generally observed good physical and psychological conditions when administering their classroom achievement tests. This was a very good indication for classroom achievement test development.

On the scoring of classroom achievement tests, Amedahe (1989) reported that teachers in the schools used mainly the analytic method in scoring their essay-type tests. Again, teachers in the schools scored their essay-type tests either item by item or script by script. On the part of Quaigrain (1992), he found that majority of teachers in the schools used the analytic method in scoring their essay-type tests. Also, almost half of the teachers scored their essay-type tests item by item while the other half scored them script by script. These two findings of Quaigrain are in total agreement with Amedahe's findings. The analytic method of scoring seems to be very popular with classroom teachers and this may be attributed to the numerous advantages it holds over the holistic method of scoring, especially in formative testing.

A comparison of the two studies reviewed above reveals quite similar findings. It could, therefore, be concluded that for a period of three years from 1989 to 1992, Quaigrain's study came to confirm the findings of Amedahe's study to a large extent.

Summary

From the literature review, the state of the art of achievement testing in the USA, England and Ghana are given in the following paragraphs.

In the USA, most teachers-made tests contain many faults and most, except in mathematics and science, function exclusively at the recall level. On item type, teachers use all major item formats. Teachers typically do not use test improvement strategies such as test blueprints and item analysis. Limited expertise, support and pre-service and in-service training are available to assist

teachers meet their testing responsibilities. On the whole, teachers in the USA are generally not well prepared in assessment (Stiggins, 1999). Only a few states require competence in assessment as a condition for licensure.

In England, classroom evaluation practices generally encourage superficial and rote learning, concentrating on recall of isolated details, usually items of knowledge, which pupils soon forget. Teachers over-emphasis the grading function while the learning function is under-emphasised. On test-score interpretation, teachers use normative rather than a criterion approach which empasises competition between pupils, rather than personal improvement of each student. On the whole, there is a case of general weakness and lack of competence in assessment on the part of teachers in England.

In Ghana, the studies reviewed have shown that to a great extent, teachers generally, do not follow the basic prescribed principles in test construction. On the relationship between pre-service training and actual testing practice, the first study found no relationship while the second study three years later, found a weak positive relationship. On test administration, teachers generally observe good physical and psychological conditions. Finally, on test scoring, teachers use the analytic method and score their essays either script by script or item by item.

CHAPTER THREE

METHODOLOGY

This chapter discusses the methodology adopted in carrying out the study. The methods and approaches as described in the chapter are under nine sub-sections. These are the Research Design, Population, Sample and Sampling Procedure, Research Instruments, Pre-testing Procedure, Validity and Reliability of the Instruments, Training of Assistants, Data Collection Procedure and Data Analyses.

The main rationale for the study was to find out whether teachers in the SSSs in the Ashanti Region of Ghana follow the basic principles of constructing, administering, scoring classroom achievement tests and interpreting the results of these tests. It also sought to find out the differences (if any) that exist between teachers who had training in testing and those who did not, in terms of the principles used in the construction, administration and scoring of classroom achievement tests and the interpretation of the test results. This customarily involved a critical inquiry into how teachers construct, administer, score and interpret the results of their achievement tests.

The Research Design

The research design chosen for the study is the descriptive sample survey. According to Amedahe (2004), "descriptive research is research which specifies the nature of a given phenomenon" (p. 50). Gay (cited in Amedahe, 2004), explains that descriptive research involves the collection of data in order to test hypotheses or answer research questions concerning the current status of the subjects of the study.

The descriptive research design was deemed best for the study because, according to Best and Khan (cited in Amedahe, 2004), descriptive research is concerned with the conditions or relationships that exist, such as determining the nature of prevailing conditions, practices and attitudes, opinions that are held, processes that are going on, or trends that are developed. This, however, was the main purpose of the study. It was, to collect data in order to answer research questions concerning the current status of achievement testing in the SSSs in the Ashanti Region. Another reason for the adoption of the descriptive research design is that it is suitable for either a quantitative or qualitative research where there is the formulation of hypotheses or research questions to be tested or answered in order to describe situations (Amedahe, 2004). Also, the descriptive survey affords the opportunity to select a sample from the population being studied and then make generalisations from the study of the sample (Ary, Jacobs and Razavieh, 1990; Gay, 1992). A major advantage of the descriptive research design is that it often employs the method of randomization so that error may be estimated when population characteristics are inferred from observation of samples (Amedahe, 2004). Lastly, the descriptive design is highly regarded by policy

makers in the social sciences where large populations are dealt with, and is widely used in educational research since data gathered by way of descriptive survey represent field conditions (Osuala, 1991).

Irrespective of the strengths of the descriptive survey mentioned above, Fraenkel and Wallen (2000) identified the weaknesses of the descriptive survey as (1) difficulty in ensuring that the questions to be answered are clear and not misleading, (2) getting respondents to answer questions thoughtfully and honestly is a setback, and (3) getting a sufficient number of questionnaires completed and returned so that meaningful analysis can be made is also a set back. Osuola (2001) in buttressing the points on the weaknesses of the descriptive research, pointed out that, "designing a quality investigation requires particular attention to two central factors: appropriate sampling procedures, and precision in defining terms in eliciting information" (p. 201). He continued by adding that, while descriptive research is a prerequisite for finding answers to questions, it is not in itself sufficiently comprehensive to provide answers and that it cannot also provide cause-and-effect relationships.

Despite the difficulties and setbacks of the descriptive research outlined above, it was still deemed appropriate for the study because of the potential that it held for achieving the main purpose of the study.

The Population

According to Amedahe (2004), the target group about which a researcher is interested in gaining information and drawing conclusions is what is known as the population. It is a group of individuals who have one or more characteristics in common that are of interest to the researcher. In this

study, the target population was the set of teachers in all the public SSSs in the Ashanti Region. They were 3182 in number (Ashanti Regional Education Office, 2005). For the purpose of the study, the accessible population consisted of teachers of Core Mathematics, English Language and Integrated Science in the 82 public SSSs in the Ashanti Region. It was made up of 565 teachers of Core Mathematics, 576 teachers of English Language and 605 teachers of Integrated Science giving rise to a total of 1,746 (Researcher's Field Data, March, 2006).

Sample and Sampling Procedure

A sample is a proportion of the population selected for observation and analysis. According to Sarantakos (1997), a sample enables the researcher to study a relatively smaller number of units in place of the target population and to obtain data that are representative of the target population.

In this study, a sample size of 265 teachers was used. This was made up of 91, 80, and 94 teachers of Core Mathematics, English Language and Integrated Science, respectively.

The sampling procedures adopted for the study were the cluster and simple random sampling. Ten districts were selected from the 21 districts in the Ashanti Region by simple random sampling. Random sampling with replacement was used. These districts formed the clusters from which cluster of schools were randomly selected. These districts had a total of 56 SSSs.

To come out with a representative sample of schools in the 10 districts, the researcher obtained the names of all the SSSs in each district from the Ashanti Regional Education Office. These schools formed the clusters from which

appropriate proportional representations of schools were determined. The names of the various schools were coded to ensure anonymity and from this list, another sampling frame was built. The proportionate number of schools in each district required for the study was then determined by dividing the total number of schools to be sampled for the study which was 26, by the total number of schools in the districts sampled which was 56. This gave 0.46 or 46%. The product of the number of schools in each district and 0.46 gave the number of schools to be selected from the particular district. For example, Ejisu/Juaben District had five schools and multiplying by 0.46 gave 2.3 which was approximated to two schools. Afigya Sekyere District had six schools and multiplying by 0.46 gave 2.75 which was approximated to three schools. The procedure used in selecting the schools from each district was the simple random sampling, which was drawing from a pool with replacement. In all, 26 schools were selected from the 10 districts. The list of the 26 schools and the number of teachers sampled are in Appendix A.

Table 2 shows the distribution of the selected districts and the number of schools sampled from each district.

Table 2

Distribution of Sampled Schools in Each District

Serial No.	Name of District	No. of SSSs in District	No. of SSSs Sampled in District
01	Afigya Sekyere	6	3
02	Amansie East	4	2
03	Obuasi Municipal	3	1
04	Asante Akim North	4	2
05	Bosomtwe–Atwima/Kwanwoma	3	1
06	Ejisu / Juaben	5	2
07	Kumasi Metropolis	17	8
08	Kwabre	6	3
09	Sekyere West	4	2
10	Sekyere East	4	2
	Total	56	26

In selecting the required number of teachers from the sampled schools, the researcher visited the schools personally and contacted the assistant headmasters and heads of departments for the number of the teachers teaching Integrated Science, Core Mathematics and English Language in Forms One, Two and Three. This gave rise to a sample size of 509 for the study which forms about 29.15% of the accessible population. With the case of teachers teaching Forms One, Two and Three, if one was selected for Form One first, he/she was excluded from the selection process in Forms Two and Three and

vice versa. This was to facilitate independent sampling and to ensure that a teacher appeared only once in the study to avoid duplication of responses during data collection.

In selecting schools for the observation of the conditions under which teachers tested their students, the simple random sampling procedure, where there was selection from a pool with replacement was adopted. A sample size of 10 schools which forms about 12.20% of the population of schools under study was selected. According to Amedahe (2004), in most quantitative studies, a sample size of 5% to 20% of the population size is sufficient for generalization purposes, depending upon the size of the population.

Research Instruments

The data collection instruments used for the study were a questionnaire, and an observation guide. The choice of a questionnaire is based on the assertion of Osuola (2001) that, "they are particularly advantageous whenever the sample size is large enough to make it uneconomical for reasons of time or funds to observe or interview every subject" (p. 268). The questionnaire consisted of four sections, A, B, C and D. (See Appendix B). The items on the questionnaire were mainly close-ended with only one being open-ended. The first section (Section A, items 1 to 5), centred on the background information of the respondents. The second section (Section B, items 6 to 20) concentrated on the basic principles and some other considerations in test construction.

The third section (Section C, items 21 to 38) dealt with the principles of test administration. Under this, both the physical and psychological

conditions under which teachers administered their tests were examined. The fourth section (Section D, items 39 to 52) centred on the principles teachers used in scoring their essay-type tests as well as how teachers interpreted the results of their tests.

Most of the items on the questionnaire were multiple-scored on a four-point Likert type scale with a few being dichotomously scored. The items on the Likert type scale were scored ranging from three (3) for "Always" to zero (0) for "Never". The Likert type scale was chosen because according to Gyimah (2002), in measuring the views and impressions of teachers on an on-going practice, it is the most simple, but equally efficient approach when considered alongside with social-distance scales, Thurstone scales and the scalogram analysis. It was adopted also to ensure effective analysis of the data even though it restricts free expression and perception of respondents in a study.

Pre-testing Procedure

The questionnaire was pre-tested in three SSSs in the Central Region of Ghana. These were Breman Asikuma Secondary School, Agona Nsaba Presbyterian Secondary School and Obiri Yeboah Secondary School in Assin Fosu. Blank sheets of papers were added to the questionnaire for respondents to express their views in writing on the clarity, ambiguity, biases, inconsistencies and problems in all aspects of the questionnaire. Forty-six teachers teaching Core Mathematics, English Language and Integrated Science were involved in the pre-testing exercise. It should be noted that the location of these three schools has similar socio-cultural characteristics with

that of the schools in the Ashanti Region where the study was done. Teachers in the two regions therefore have similar characteristics.

Feedback from the pre-testing helped to revise items that were either ambiguous or appeared not to measure what it was intended for. For example, item 2 was reworded and item 12 clarified.

Validity and Reliability of the Instruments

The content validity of the questionnaire was established by submitting the questionnaire to lecturers of the Educational Foundations Department whose area of specialisation is educational measurement and evaluation and research methods, and have expert knowledge in validation of research instruments, for review. This helped to establish the content validity of the questionnaire.

The reliability (internal consistency) of the questionnaire for the main study was estimated using Cronbach's co-efficient alpha. According to Cronbach (cited in Ebel & Frisbie, 1991), co-efficient alpha can provide a reliability estimate for a measure composed of items of varying point values such as essays or attitude scales that provide responses such as "strongly agree" and "strongly disagree" with intermediate response options. The Cronbach's co-efficient alpha for the main study was 0.70.

Observation was used to obtain information on both the physical and psychological conditions under which students were tested in the schools. This observation was important to bolster the information that was gathered from the questionnaire on test administration. The researcher adapted and used the observation guide developed by Amedahe (1989) for the same purpose. This instrument was based on the principles of test administration and the

conditions testing experts deem appropriate to testing students in order to ensure students' maximum performance. The instrument was pretested at Edinaman Secondary School in the Central Region and revised. This catered for its content validity. Appendix C shows the observation guide.

Training of Observation Assistants

Five students from the University of Cape Coast and University of Ghana who were on vacation were recruited and trained by the researcher to assist in the observation. The training session took the form of an hour briefing on what to observe and how to use the observation guide to record what is observed.

Data Collection Procedure

The questionnaire was administered personally by the researcher to all the 509 teachers involved in the study in the 26 selected SSSs in the Ashanti Region. The researcher used a period of two weeks to travel to all the sampled schools to administer the questionnaire. Respondents were given a period of two weeks to respond to the questionnaire after which the researcher travelled round again to the schools for collection.

The observations were done in the 10 schools selected (for observation) by the researcher together with the five trained assistants. To ensure the reliability of the results from the observations, the observations were done by two people at a time in each classroom. Results were compared after each observation. Each teacher was observed on two occasions.

The data collection process started on 28th June, 2006 and ended on 31st July, 2006, thus, spanning a period of one month. Appendices D and E

show the dates on which the questionnaires were administered and collected from the various schools while appendix F shows the dates on which the teachers were observed in the schools. Out of the 509 questionnaires administered, 265 representing 52.06% were retrieved.

It should be noted that the researcher made about three follow-ups to many of the schools to collect the completed questionnaires. The low return rate of the questionnaire is simply due to the poor attitude of Ghanaian teachers toward completion of questionnaire and research.

Data Analysis

The responses to the questionnaires were first edited, coded and scored. The editing procedure was to check whether respondents had followed directions correctly, and whether all items had been responded to. Section A was on some background information of the respondents. These responses were analysed using frequency and percentage tables.

For Sections B, C and D, items 6 to 15 and 21 to 38 were assigned the weights of 3, 2, 1 and 0 for “always”, “very often”, “sometimes” and “never”, for positive items respectively. The items with negative implications were 37 and 38 (Section C) and so the weights were reversed directly for them. Items 16 to 18 and 41 to 48 were dichotomously scored. Items 19, 20, 39, 40 and 49 to 52 were scored for a point each.

Research Question One

Which principles do SSS teachers use in the construction of their classroom achievement tests?

The respondents' responses to items 6 to 15 which represented 10 test construction principles were used in answering this research question. The data on this research question were analysed using frequency and percentage tables. A frequency and percentage table showing the proportion of responses on the response options, "always", "very often", "sometimes" and "never" for each of the principles was used. For each principle, the proportions for "always and "very often" were added and taken as one, while the proportions for "sometimes" and "never" were also added and taken as one. The binomial test (z-test) was further used to test whether or not the observed proportion for "always" and "very often" for each of the principles was significantly different from a hypothesised proportion of 50%. For each principle that the observed proportion for "always" and "very often" was found to be significantly different from and higher than the hypothesised proportion of 50%, a conclusion was drawn that, that principle is used by respondents and vice versa.

Research Question Two

What differences exist between SSS teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement in terms of the principles used in test construction?

The scores for the respondents' responses to items 6 to 15 which represented 10 test construction principles were used to answer this research question. The t-test for independent samples was computed between teachers

who received instruction in educational measurement and those who did not receive instruction in educational measurement for each of the principles. The result of the t-test was then used to determine the principles in which differences existed between teachers who received instruction in educational measurement and those who did not, in terms of the application of the principles. The level of significance was 0.05.

Research Question Three

Which principles do SSS teachers use in the administration of their classroom achievement tests?

The respondents' responses to items 21 to 38 which represented 18 test administration principles were used in answering this research question. The data on this research question were analysed using frequency and percentage tables. A frequency and percentage table showing the proportion of responses on the response options, "always", "very often", "sometimes" and "never" for each of the principles was used. For each principle, the proportions for "always and "very often" were added and taken as one, while the proportions for "sometimes" and "never" were also added and taken as one. The binomial test (z-test) was further used to test whether or not the observed proportion for "always" and "very often" for each of the principles was significantly different from a hypothesised proportion of 50%. For each principle that the observed proportion for "always" and "very often" was found to be significantly different from and higher than the hypothesised proportion of 50%, a

conclusion was drawn that, that principle is used by the respondents and vice versa.

Research Question Four

What differences exist between SSS teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement in terms of the principles used in test administration?

The scores for the respondents' responses to items 21 to 38 which represented 18 test construction principles were used to answer this research question. The t-test for independent samples was computed between teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement for each of the principles. The result of the t-test was then used to determine the principles in which differences existed between teachers who received instruction in educational measurement and those who did not, in terms of the application of the principles. The level of significance was 0.05.

Research Question Five

Which principles do SSS teachers use in the scoring of their classroom achievement essay-type tests?

Items 40 to 48 which represented nine essay-type test scoring principles were used to answer this research question. A frequency and

percentage table was used to find the proportion of respondents' responses on each option under item 40. Items 41 to 48 were dichotomously scored. A frequency and percentage table showing the proportions of responses on Yes and No was used to determine the test scoring principles that respondents used and those they did not use in scoring their essay-type tests.

Item 39 on the questionnaire asked respondents to indicate the method they used in scoring their essay-type tests. This was either the analytic or the holistic method. A frequency and percentage table was used to find the proportion of respondents' responses on each of the methods. It is worthy of noting that it is after a method of scoring has been chosen that the principles of essay-type tests scoring are applied to ensure consistency in scoring and reliability of the test scores.

Research Question Six

What differences exist between SSS teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement in terms of the principles used in the scoring of essay-type tests?

The scores for the respondents' responses to items 40 to 48 which represented nine test scoring principles were used to answer this research question. The t-test for independent samples was computed between teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement for each of the nine principles. The result of the t-test was then used to determine the principles in which

differences exist between teachers who received instruction in educational measurement and those who did not, in terms of the application of the principles. The level of significance was 0.05.

Research Question Seven

How do SSS teachers interpret the results of their classroom achievement tests?

Items 49 to 52 addressed this research question. Frequency and percentage tables based on participants' responses on the methods they used in their test-score interpretation and the methods they employed under norm-referenced and criterion-referenced interpretations were used to answer this research question. Content analysis and direct quotations were also used especially for item 44 which was an open-ended question.

Research Question Eight

What differences exist between SSS teachers who received instruction in testing (i.e., educational measurement) and those who did not receive instruction in testing in terms of how they interpret the results of their classroom achievement tests?

The scores for the respondents' responses to items 49 to 52 which represented four issues on test-score interpretation were used to answer this research question. The t-test for independent samples was computed between teachers who received instruction in educational measurement and those who

did not receive instruction in educational measurement for each of the issues. The result of the t-test was then used to determine the issues on test-score interpretation in which differences existed between teachers who received instruction in educational measurement and those who did not, in terms of how the teachers handled those issues. The level of significance was 0.05.

CHAPTER FOUR

RESULTS AND DISCUSSION

The study aimed at finding out whether SSS teachers follow the basic prescribed principles in the construction, administration and scoring of classroom achievement tests, and how they interpret the results of these tests.

This chapter presents the results of the analyses and discussion of the findings of the study. The data were analysed through frequency and percentage tables, the binomial test and the t-test for independent samples as presented in the previous chapter.

Results

Background Information

The study was carried out in 26 SSSs in the Ashanti Region of Ghana, with a sample size of 265. The number of respondents from each school ranged from five to 20. The 26 SSSs were located in 15 towns. (See Appendix G for the distribution of towns in which SSSs are located).

Form(s) and Subject(s) Taught by Respondents

Item 4 of the questionnaire requested respondents to indicate the Form(s) and the subject(s) they taught. Table 3 shows the Form(s) taught by respondents.

Table 3

Form(s) Taught by Respondents

Form	Frequency	Percent (%)
One	49	18.5
Two	31	11.7
Three	20	7.5
One and Two	47	17.7
One and Three	11	4.2
Two and Three	14	5.3
One, Two and Three	93	35.1
Total	265	100.0

Table 3 indicates that majority of the respondents representing about 62.3% taught more than one Form. Among those teaching more than one Form, 93 (35.1%) taught all three Forms. A total of 49 (18.5%) taught Form One only, 31 (11.7%) taught Form Two only, and only 20 (7.5%) taught Form Three only. In sum, teachers teaching all the three Forms were the dominant group in the study.

The subject(s) taught by respondents is / are shown in Table 4.

Table 4

Subject(s) Taught by Respondents

Subject	Frequency	Percent (%)
English Language	80	30.2
Core Mathematics	91	34.3
Integrated Science	94	35.5
Total	265	100.0

From Table 4, 80 (30.2%) of the respondents were Mathematics teachers, 91 (34.3%) were English Language teachers while 94 (35.5%) were Integrated Science teachers. The study, therefore, gave a near even distribution of teachers in the three subject areas.

Pre-service Training in Educational Measurement

Item 5 of the questionnaire asked respondents to state whether they had ever taken a course in educational measurement. The result is shown in Table 5.

Table 5

Respondents who took a Course or did not take a Course in Educational Measurement

Status	Frequency`	Percent (%)
Took a Course in Educational measurement	182	68.7
Did not take any Course in Educational Measurement	83	31.3
Total	265	100.0

From Table 5, more than two-thirds of the respondents, 182 (68.7%) indicated that they have taken a course in educational measurement whereas the remaining 83 (31.3%) indicated that they had never taken a course in educational measurement. It is worth noting that all the respondents with no educational measurement training were either holders of a bachelor of arts (B.A) or bachelor of science (B.Sc.) degrees with no professional educational component.

Research Question 1

Which principles do SSS teachers use in the construction of their classroom achievement tests?

Research question 1 sought to find out the kind of principles that SSS teachers use in the construction of their achievement tests. Ten test construction

principles were outlined in the questionnaire (items 6-15, Section A, Appendix B) and ranged from defining the purpose of the test to the final evaluation of the test before it is submitted for typing and subsequent administration.

Appendix H gives the percentage distribution of the responses to these items. This was based on a 4−point Likert type scale ranging from “Always”, “Very Often”, and “Sometimes” to “Never”.

Table 6 gives the binomial test (z-test) of the observed proportions for “Always” and “Very Often” for all the test construction principles against a hypothesised (test) proportion of 50% (0.50) at a level of significance (alpha level) of 0.05, two-tailed.

Table 6

Binomial Test of Proportions for Test Construction Principles

Principle	Category	N	Observed Prop.	Test Prop.	P-value
1.Define the purpose of the test.	Sometimes/Never	67	.25	.50	.000
	Always/Very Often	198	.75		
2. Relate the instructional objectives of the subject matter to the test.	Sometimes/Never	31	.12	.50	.000
	Always/Very Often	243	.88		

Table 6 (continued)

Principle	Category	N	Observed Prop.	Test Prop.	P-value
3. Select the test format suitable for testing the stated objectives.	Sometimes/Never	44	.17	.50	.000
	Always/Very Often	221	.83		
4. Use a table of specifications to decide the items on the test.	Sometimes/Never	178	.67	.50	.000
	Always/Very Often	87	.33		
5. Prepare more items than needed in the test.	Sometimes/Never	178	.67	.50	.000
	Always/Very Often	87	.33		
6. Write test items in advance (at least two weeks) of the test date to permit review.	Sometimes/Never	109	.41	.50	.005
	Always/Very Often	156	.59		
7. Prepare marking scheme as soon as the test items are written.	Sometimes/Never	48	.18	.50	.000
	Always/Very Often	217	.82		
8. Review test items after they have been set aside for a few days by reading over the items.	Sometimes/Never	110	.42	.50	.007
	Always/Very Often	155	.58		

Table 6 (continued)

Principle	Category	N	Observed Prop.	Test Prop.	P-value
9. Write clear and concise directions for the entire test and sections of the test.	Sometimes/Never	20	.08	.50	.000
	Always/Very Often	245	.92		
10. Evaluate the test as a whole on the criteria of clarity, validity, practicality and fairness.	Sometimes/Never	168	.63	.50	.000
	Always/Very Often	97	.37		

There is an indication that majority of the respondents in the study applied most of the test construction principles either always or very often (Please refer to Appendix H).

On the principle of defining the purpose of the test, 74.7% of the respondents indicated they applied this principle always or very often. Only 25.3% indicated they applied the principle sometimes or never applied the principle at all. From Table 6, the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). This means that teachers in the study practised this principle to an appreciable level (always or very often). With the principle of relating the instructional objectives of the subject matter to the test, 88.4% of the respondents indicated they always or very often applied the principle

while only 11.6% applied it sometimes or never applied the principle at all. It could be seen from Table 6 that the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). Hence, the respondents in study practised this principle. On the next principle, "selecting the test format suitable for testing the stated objectives," the response pattern was similar. Majority, 83.0% indicated they applied this principle always or very often. Only 17.0% indicated they either applied the principle sometimes or never. Again, Table 6 shows that the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). Teachers in the study, therefore, practised this principle.

Concerning the principle of using a table of specifications to determine the items on the test, there was a reverse of the response pattern. Only 32.7% always or very often applied this principle. As many as 47.5% indicated they applied this principle only sometimes while 19.6% never used the table of specifications at all. Table 6 indicates that the observed proportion for always and very often is lower than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). This means that teachers in the study did not apply this principle to any appreciable level. The next principle, "prepare more items than needed in the test or examination," attracted responses from the participants that were similar to the case of the use the table of specifications. Here, 33.2% of the respondents indicated they always or very often applied this principle while 66.8% applied the principle only sometimes or never at all. The indication from Table 6 is that teachers in the study did not apply this principle to any appreciable level. This is because the observed proportion for

always and very often is lower than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). On the principle of writing the test items in advance (at least two weeks) of the test date to permit reviews and editing, the respondents maintained their initial response pattern but with a somewhat even distribution of responses from always to sometimes. A little more than half of the respondents (59.2%) indicated they either applied this principle always or very often, while a little less than half (40.8%) either applied the principle sometimes or never. Table 6 indicates that the respondents in the study practised this principle. The observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). The next principle was that of preparing the marking scheme as soon as the test items are written. Here, 57.0%, 25.3%, 14.0% and 3.8% of the respondents indicated their practice of this principle always, very often, sometimes and never, respectively. It could be seen from Table 6 that teachers in the study practised this principle since the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$).

Concerning the principle of reviewing the test items after being set aside for a few days by reading over the items, the distribution of the responses was quite even from always (34.0%), very often (26.4%), to sometimes (36.2%). A very small proportion (3.4%) of the respondents indicated they never used this principle. Table 6 shows that respondents in the study practised this principle. The observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). Regarding the principle of writing clear and

concise directions for the entire test and sections of the test, as many as 92.2% of the respondents indicated they did it always or very often while only 7.8% indicated they did it either sometimes or never. It is clear from Table 6 that respondents in the study practised this principle. The observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). Lastly, concerning the principle of evaluating the test as a whole on the criteria of clarity, validity, practicality, efficiency and fairness, as many as 63.4% of the respondents indicated they did it only sometimes or never at all while only 36.6% indicated they did it either always or very often. This principle is not practised to any appreciable level by teachers in the study since Table 6 shows that the observed proportion for always and very often is lower than and significantly different from the hypothesised proportion of 50% ($p < 0.05$).

The analysis above indicates that teachers generally practised to an appreciable level ("Always" and "Very Often"), seven of the test construction principles. This gives the number of principles not practised at all or not to any appreciable level ("Never" or "Sometimes") by teachers as three.

The principles used by teachers are:

i. defining the purpose of the test,

ii relating the instructional objectives of the subject matter to the test,

iii. selecting the test format suitable for testing the stated objectives,

iv. writing the test items in advance (at least two weeks) of the test date to permit review and editing,

v. preparing the marking scheme as soon as the test items are written,

vi. reviewing the test items after they have been set aside for a few days,

and

vii. writing clear and concise directions for the entire test and sections of it.

The principles not used frequently by teachers are:

i. using a table of specifications to determine the items on the test,

ii. preparing more items than needed in the test, and

iii. evaluating the test as a whole on the criteria of clarity, practicality, validity, efficiency and fairness.

Research Question 2

What differences exist between SSS teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement in terms of the principles used in test construction?

It must be noted that teachers who received instruction in educational measurement during pre-service training are referred to as trained in testing in this report. To answer this question, the responses to the items on test construction principles were used. The mean scores, standard deviations and t-values of respondents trained in testing and those not trained in testing for each principle are shown in Table 7. Level of significance was 0.05.

Table 7

Results of t-test of Independence for Application of Test Construction Principles by Respondents

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
1. Define the purpose of the test.	Not trained in testing.	83	2.020	.869	-1.676	.095
	Trained in testing.	182	2.210	.814		
2. Relate the instructional objectives of the subject matter to the test.	Not trained in testing.	83	2.250	.794	-1.999	.047
	Trained in testing.	182	2.440	.660		
3. Select the format suitable for testing the stated objectives.	Not trained in testing.	83	2.020	.897	-3.268	.001
	Trained in testing.	182	2.360	.704		
4. Use a table of specifications to determine the items on the test.	Not trained in testing.	83	1.120	.847	-1.352	.177
	Trained in testing	182	1.270	.868		

Table 7 (continued)

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
5. Prepare more items than needed in the test.	Not trained in testing.	83	1.140	.857	-.392	.696
	Trained in testing.	182	1.190	.947		
6. Write test items in advance of the test date to permit review and editing.	Not trained in testing.	83	1.590	.988	-2.576	.011
	Trained in testing.	182	1.920	.945		
7. Prepare marking scheme as soon as the test items are written.	Not trained in testing.	83	2.290	.863	-.692	.490
	Trained in testing.	182	2.370	.862		
8. Review test items after they have been set aside for a few days by reading over the items.	Not trained in testing.	83	1.690	.810	-2.479	.014
	Trained in testing.	182	1.980	.943		
9. Write clear and concise directions for the entire test and sections of the test.	Not trained in testing.	83	2.480	.786	-2.412	.017
	Trained in testing.	182	2.720	.642		

Table 7 (continued)

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
10.Evaluate the test as a whole on the criteria of clarity, validity, practicality and fairness.	Not trained in testing.	83	2.290	.863	-.672	.450
	Trained in testing.	182	2.370	.862		

From Table 7, five out of the 10 test construction principles indicated statistically significant differences between respondents who received instruction in testing and those who did not receive instruction in testing. These are:

i. 'Relate the instructional objectives of the subject matter to the test': not trained in testing ($\underline{M}$ = 2.250, $\underline{SD}$ = 0.794) and trained in testing ($\underline{M}$ = 2.440, $\underline{SD}$ = 0.660), $\underline{t}_{(263)}$ = -1.999, $\underline{p}$ < 0.05.

ii. 'Select the test format suitable for testing the stated objectives': not trained in testing ($\underline{M}$ = 2.020, $\underline{SD}$ = 0.897) and trained in testing ($\underline{M}$ = 2.360, $\underline{SD}$ = 0.704), $\underline{t}_{(263)}$ = -3.268, $\underline{p}$ < 0.05.

iii. 'Write the test items in advance of the test date to permit reviews and editing': not trained in testing ($\underline{M}$= 1.590, $\underline{SD}$ = 0.988) and trained in testing ($\underline{M}$ = 1.920, $\underline{SD}$ = 0.945), $\underline{t}_{(263)}$= -2.576, $\underline{p}$ < 0.05.

iv. 'Review the test items after they have been set aside for a few days by reading over them': not trained in testing ($\underline{M}$ =1.690, $\underline{SD}$ =0.810) and trained in testing ($\underline{M}$ =1.980, $\underline{SD}$ = 0.943), $\underline{t}_{(263)}$= -2.479, $\underline{p}$ < 0.05.

v. 'Write clear and concise directions for the entire test and sections of the

test': not trained in testing ($\underline{M}$ = 2.480, $\underline{SD}$ = 0.786) and trained in testing ($\underline{M}$ = 2.720, $\underline{SD}$ = 0.642), $\underline{t}_{(263)}$ = -2.412, $\underline{p}$ < 0.05.

From the analysis above, it could be concluded that in all the five principles listed above, respondents who received instruction in testing indicated that they applied the principles more frequently than their counterparts who did not receive instruction in testing.

Research Question 3

Which principles do SSS teachers use in the administration of their classroom achievement tests?

This question sought to find out the principles that SSS teachers used in administering their classroom achievement tests. Eighteen statements covering preparation of students in advance for the test and actual test administration (invigilation) principles were outlined in the questionnaire (items 21-38, Section B, Appendix B).

Appendix I gives the levels (percentages) of practice of the respondents in the principles outlined. The analysis was based on a 4-point Likert type scale ranging from "Always", "Very Often", and "Sometimes" to "Never".

Table 8 shows the binomial test (z-test) of the observed proportion for "Always" and "Very Often" for each of the test administration principles against a test proportion of 50% (0.50) at a level of significance of 0.05, two-tailed.

Table 8

Binomial Test of Proportions for Test Administration Principles

Principle	Category	N	Observed Prop.	Test Prop.	P-value
1. Make students aware of when (date & time) the test will be given.	Sometimes/Never	24	.09	.50	.000
	Always/Very Often	241	.91		
2.Make students aware of the conditions under which test will be given.	Sometimes/Never	82	.31	.50	.000
	Always/Very Often	183	.69		
3. Make students aware of the content areas that the test will cover.	Sometimes/Never	77	.29	.50	.000
	Always/Very Often	188	.71		
4. Make students aware of the test formats.	Sometimes/Never	44	.17	.50	.000
	Always/Very Often	221	.83		
5. Make students aware of the weighting of content areas.	Sometimes/Never	147	.55	.50	.085
	Always/Very Often	118	.45		
6. Make students aware of how the test will be scored and graded.	Sometimes/Never	113	.43	.50	.020
	Always/Very Often	152	.57		

Table 8 (continued)

Principle	Category	N	Observed Prop.	Test Prop.	P-value
7. Make students aware of the rules governing the conduct of the test.	Sometimes/Never	47	.18	.50	.000
	Always/Very Often	218	.82		
8. Make students aware of the importance of the results of the test.	Sometimes/Never	48	.18	.50	.000
	Always/Very Often	217	.82		
9. Do not give tests immediately before or just after a long vacation or other key events.	Sometimes/Never	164	.62	.50	.000
	Always/Very Often	101	.38		
10. Ensure that the sitting layout allows enough space.	Sometimes/Never	37	.14	.50	.000
	Always/Very Often	228	.86		
11.Ensure adequate ventilation and lighting in the testing room.	Sometimes/Never	44	.17	.50	.000
	Always/Very Often	221	.83		
12. Use “Do Not Disturb. Examinations in Progress” sign.	Sometimes/Never	218	.82	.50	.000
	Always/Very Often	47	.18		

Table 8 (continued)

Principle	Category	N	Observed Prop.	Test Prop.	P-value
13. Expect and cater for all possible emergencies during examinations.	Sometimes/Never	135	.51	.50	.806
	Always/Very Often	130	.49		
14. Announce the remaining Time at regular intervals.	Sometimes/Never	21	.08	.50	.000
	Always/Very Often	244	.92		
15. Stand where all students can be viewed.	Sometimes/Never	10	.04	.50	.000
	Always/Very Often	255	.96		
16. Keep all remarks in the testing room to a minimum and make sure they relate to the test.	Sometimes/Never	76	.29	.50	.000
	Always/Very Often	189	.71		
17. Ask students to work faster in order to finish on time.	Sometimes/Never	131	.49	.50	.902
	Always/Very Often	134	.51		
18. Tell students the dire consequences of failure in the test they are taking	Sometimes/Never	137	.52	.50	.623
	Always/Very Often	128	.48		

There is a general indication that majority of the respondents in the study practised most of the test administration principles outlined either always or very often (Please refer to Appendix I).

With the principle of preparing the students in advance for the test, the first step outlined was to make the students aware of when (date & time) the test would be given. As many as 90.6% of the respondents indicated they did this always or very often. Only 9.4% of the respondents indicated they did it sometimes or never. Table 8 shows that the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). Teachers in the study, therefore, practised this principle.

On making students aware of the conditions (number of items, place of test, closed or open book) under which the test will be given, the majority (68.8%) of the respondents indicated they did it either always or very often. Only 31.3% did it sometimes or never did it at all. Teachers in the study applied this principle. The observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). (Refer to Table 8).

Next on preparation, was to make the students aware of the content areas (study questions, list of topics or learning targets) that the test woud cover. Here, 70.9% of the respondents pointed out they did it always or very often while 29.1% did it sometimes or never did it at all. Table 8 indicates that the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). Hence, teachers in the study applied this principle.

Next, was making students aware of the test format (objective-type or essay-type). The response pattern of the respondents remained the same. Most of the respondents (83.4%) pointed out that they did it always or very often

while 16.6% did it sometimes or never did it at all. From Table 8, the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). Teachers in the study, therefore, practised this principle.

With regard to making students aware of the emphasis or weighting of content areas (i.e., value in points or content areas with higher marks), the response pattern of the respondents was reversed. Less than half of the respondents (44.5%) indicated they did it always or often. The rest (55.5%) of the respondents did it only sometimes or never did it at all. The indication from Table 8 is that teachers in the study did not apply this principle to any appreciable level since the observed proportion for always and very often is lower than and significantly different from the hypothesised proportion of 50% ($p > 0.05$).

Next, was making the students aware of how the test will be scored and graded. On this, the response pattern resumed the initial trend. The majority (57.0%) indicated they did it always or very often while the minority (43.0%) indicated they did it sometimes or never at all. This principle is practised by respondents. The observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). (Refer to Table 8).

On making the students aware of the rules and regulations governing the conduct of the test, 56.2% of the respondents pointed out that they did it always, 26.0% did it very often, 14.3% did it sometimes, and while 3.4% never did it at all. The indication from Table 8 is that teachers in the study applied this principle since the observed proportion for always and very often

is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). The last principle on preparing the students in advance for the test is making them aware of the importance of the results of the test. With this, most of the respondents (82.2%) indicated they did it always or very often, with only 17.7% indicating that they did it only sometimes or never at all. The respondents practised this principle since the observed proportion for always and very often is relatively higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). (Refer to Table 8).

On actual test administration and invigilation, the first principle outlined in the questionnaire was the issue of the inappropriateness of giving tests immediately before or after a long vacation or any other important event where students will be either psychologically or emotionally involved. There was a reverse of the initial response pattern. Only 38.4% of the respondents indicated they always or very often did not give tests immediately before or after important events. As many as 42.3% of the respondents indicated that sometimes, they did not give tests immediately before or after important events. The proportion that responded never to this principle was 17.0%, implying they always gave tests even immediately before or after important events. Table 8 shows that the observed proportion for always and very often is lower than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). Hence, respondents did not apply this principle to any appreciable level. Analysis of the data obtained from the observations carried out of the conditions under which students are tested in the schools supported this finding. In seven of the 10 schools observed, the day of vacation was the same day students wrote their last examinations.

With regard to the time (period) of testing, however, in the 20 observations undertaken, the times were appropriate. What constituted an appropriate time of testing was a time that students would not be doing something pleasant such as having breakfast, lunch, sports and games. Concerning the starting and stopping of tests, the teachers who were observed performed quite creditably. Of the 20 observations undertaken, there were 15 cases of starting promptly and stopping on time, three cases of not starting promptly and not stopping on time, one case of starting promptly and not stopping on time and one case of not starting promptly and stopping on time.

The next test administration principle was ensuring that the sitting arrangement allows enough space so that students would not copy from each other. There was a resumption of the initial response pattern. Most respondents (85.6%) indicated they did this always or very often. Only 14.4% indicated they did this sometimes or never did this at all. It could be seen from Table 8 that respondents practised this principle since the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). The observational data fully supported this finding. In all the 20 observations made in the 10 schools, there were only three cases of inappropriate arrangement of tables and chairs. In addition to this, there was only one case of the tables and chairs being unsuitable. In this particular school, the desks were generally old with a few of them broken down.

Next, was to ensure adequate ventilation and lighting in the testing room. As many as 83.1% of the respondents pointed out they did it always or very often while only 16.9% of the respondents indicated they did it

sometimes or never did it at all. The indication from Table 8 is that teachers in the study applied this principle since the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). In the 20 observations conducted, there was adequate ventilation and lighting in all the testing rooms. The observational data therefore, supported this finding.

On the use of a "Do Not Disturb. Examinations in Progress" sign when students are taking tests and examinations, there was a total reverse of the initial response pattern. Only 17.8% of the respondents pointed out they used this sign always or very often, with 82.2% using this sign sometimes or never using it at all. Respondents did not apply this principle to any appreciable level since Table 8 shows that the observed proportion for always and very often is lower than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). Of the 20 observations done, there were 16 cases where quietness in the vicinity was at an appropriate (acceptable) level while only four cases were inappropriate. This might have accounted for the reason why teachers did not use the "Do Not Disturb. Examinations in Progress" sign to ward off intruders. There seemed to be no need for the use of this sign. The inappropriate cases stemmed mainly from the sound of moving vehicles particularly with schools very close to highways and noise from dormitories where classroom blocks are very close to dormitories.

On the principle of expecting and taking care of all possible emergencies during examinations, the distribution of the responses was near even, from always to sometimes. About half of the respondents (49.2%) indicated they did it always or very often, while the other half, 50.8%

indicated they did it sometimes or never at all. Here, the observed proportion for always and very often was not higher or lower than and significantly different from the hypothesised proportion of 50% ($P > 0.05$). Hence, it could be concluded that while half of the respondents did expect and cater for all possible emergencies during testing, half of them did not. The observational data fully supported this finding. Of the 20 observations undertaken, there were nine and 11 cases, respectively, of adequately and inadequately expecting and taking care of all possible emergencies. The key items observed in the schools were a first aid box to at least give first aid treatment to any student who might suddenly fall sick during an examination and an available school car to transport any examination casualty to the clinic or hospital.

Next, was the principle of announcing the remaining time (time to complete test) at regular intervals during examinations. As many as 91.7% of the respondents indicated they did it always or very often and only 8.3% pointed out they did this sometimes or never did this at all. Table 8 indicates that this principle is applied by respondents since the observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). The observational data agreed with this finding that the respondents announced the remaining time (time left to complete test) at regular intervals. In all, a total of 83 announcements about time were made in the 20 observations. Of these, 64 were done at regular intervals and 19 at irregular intervals. There was one particular case in a school where no announcements were made at all about the time from the beginning of the test to the end.

On the principle of the invigilator standing where he could view all students and move among the students once a while to check malpractices, the majority (96.2%) of the respondents pointed out they did this always or very often, while the minority (3.8%) indicated they did this sometimes or never at all. From Table 8, this principle is practised by teachers in the study. The observed proportion for always and very often is higher than and significantly different from the hypothesised proportion of 50% ($p < 0.05$). The data from the 20 observations supported this finding but not so strongly. There were 12 cases of appropriate positions of the invigilator and eight cases of inappropriate positions of the invigilator. The appropriate positions were mainly cases of invigilators sitting or standing either in front or at the back of the students from where every student could be seen. The inappropriate positions were mainly invigilators who walked around and looked over students' shoulders during testing. The observational data gave 26 cases of invigilators looking over students' shoulders in the 20 observations carried out.

Next, was keeping all remarks in the testing room to a minimum and making sure they are related to the test. Here, 71.3% of the respondents pointed out they applied this principle always or very often while 28.6% indicated they applied this principle sometimes or never at all. Table 8 shows respondents practised this principle ($p < 0.05$). Analysis of the observational data showed that with interruption to give instructions in the course of examinations, 78 cases were recorded in the 20 observations done. Of these, a total of 74 cases were to correct typographical errors. In some of the 10 schools observed, their printed materials were full of errors and the quality of

print also poor. In the four remaining cases, there were three cases of distribution of bills, and one case of announcement about payment of school fees. Thus, the observational data to a large extent supported the initial finding from the respondents since 94.9% of the interruptions to give instructions were related to the test.

Next, was item 29 on the questionnaire which sought to find out whether teachers/invigilators asked students to work faster during testing in order to finish on time and their level of doing this. With this, 24.2% of the respondents did it always, 26.4% did it very often, 31.4% did it sometimes while18.0% never did it at all. The indication from Table 8 is that half of the respondents asked their students to work faster during testing in order to finish on time while the other half did not. The observed proportion for always and very often was not higher or lower than, and not significantly different from the hypothesised proportion of 50% ($p > 0.05$). In the 20 observations carried out, there were 15 cases in four of the 10 schools observed that teachers asked their students to work faster in order to finish on time. Thus, the observational data confirms this finding.

Item 30 on the questionnaire also sought to find out whether teachers/invigilators told students the dire consequences of failure in the tests they take and their level of doing this. A proportion of 48.3% of the respondents indicated they did this always or very often and 51.7% indicated they did this sometimes or never did this at all. The observed proportion for always and very often was not higher or lower than, and not significantly different from the hypothesised proportion of 50% ($p > 0.05$). (Refer to Table 8). Hence, half of the respondents told their students the dire consequences of

failure in the tests they take while the other half did not. The observational data did not confirm this finding fully. In the 20 observations done, there were only three cases that occurred in only one of the 10 schools by one particular teacher.

Other physical conditions observed were the provision of extra sheets and other writing materials, and other invigilator activities that are detrimental to the conduct of examinations. On the provision of extra sheets and other writing materials, in all the 20 observations made, there was adequate provision. Concerning other invigilator activities, there were 19 cases in the 20 observations done. Some of these activities observed were marking of answer scripts, conversing with a friend, listing to and making of telephone calls, reading, and dozing off while invigilating.

One other psychological condition observed was indication of cheating by students. A total of 128 cases were recorded in the 20 observations undertaken. Actions that constituted cheating or attempts to cheat were stretching of necks to look at other students' work, talking or whispering something to a friend, exchange of items such as calculators and erasers without the prior knowledge of the invigilator, and going out to urinate and overstaying unjustifiably. It was largely observed that some invigilators were not committed to their work and this gave the students the room to cheat. One invigilator in one of the schools left the examination room for a duration of about 15 minutes before coming back.

In conclusion, the test administration principles teachers indicated that they applied are:

i. making students aware of when (date & time) the test will be given,

ii. making students aware of the conditions (number of items, place of test, open book or closed) under which the test will be given,

iii. making students aware of the content areas (study questions, list of topics or learning targets) that the test will cover,

iv. making students aware of the test formats (objective- or essay-type),

v. making students aware of how the test will be scored and graded,

vi. making students aware of the rules and regulations governing the conduct of the test,

vii. making students aware of the importance of the results of the test,

viii. ensuring that the sitting arrangement allows enough space so that students do not copy from each other,

ix. ensuring adequate ventilation and lighting in the testing room,

x. announcing the remaining time (time left to complete test) at regular intervals,

xi. standing where all students can be viewed and moving among the students once a while to check malpractices during examinations, and

xii. keeping all remarks in the testing room to a minimum and making sure they are related to the test.

The above number of test administration principles practised by teachers gives the number of principles not practised by teachers to any appreciable level as six. These are:

i. making students aware of the emphasis or weighting of content areas (i.e., value in points or content areas with higher marks),

ii. not giving tests immediately before or just after important events,

iii. the use of a "Do Not Disturb. Examination in Progress" sign when

students are taking tests and examinations,

iv. expecting and catering for all possible emergencies during examinations,

v. not asking students to work faster during testing in order to finish on time, and

vi. not telling students the dire consequences of failure in the tests they take.

Research Question 4

What differences exist between SSS teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement in terms of the principles used in test administration?

To answer this question, the responses to the items on the principles of test administration (see Appendix B) were coded and scored. The mean scores, standard deviation and t-values of respondents who received instruction in testing and those who did not receive instruction in testing for each principle are shown in Table 9. The level of significance was 0.05.

Table 9

Results of t-test of Independence for Application of Test Administration Principles by Respondents

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
1. Make students aware of when (date & time) the test will be given.	Not trained in testing.	83	2.570	.666	-.308	.758
	Trained in testing.	182	2.590	.664		
2. Make students aware of the conditions under which the test will be given.	Not trained in testing.	83	2.010	.943	-.263	.793
	Trained in testing.	182	2.040	.903		
3. Make students aware of the content areas that the test will cover.	Not trained in testing.	83	2.100	.919	.204	.839
	Trained in testing.	182	2.070	.929		
4. Make students aware of the test formats.	Not trained in testing.	83	2.220	.856	-1.445	.150
	Trained in testing.	182	2.380	.844		

Table 9 (continued)

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
5. Make students aware of the weighting of content areas.	Not trained in testing.	83	1.420	1.952	-.849	.397
	Trained in testing.	182	1.530	1.006		
6. Make students aware of how the test will be scored and graded.	Not trained in testing.	83	1.780	.938	-.475	.635
	Trained in testing.	182	1.850	1.029		
7. Make students aware of the rules governing the conduct of the test.	Not trained in testing.	83	2.240	.905	-1.330	.185
	Trained in testing.	182	2.390	.819		
8. Make students aware of the importance of the results of the test.	Not trained in testing.	83	2.190	.917	-1.323	.188
	Trained in testing.	182	2.350	.777		
9. Do not give tests immediately before or after a long vacation.	Not trained in testing.	83	1.270	.951	-.942	.209
	Trained in testing.	182	1.380	.961		

Table 9 (continued)

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
10. Ensure that the sitting layout allows enough space.	Not trained in testing.	83	2.340	.816	-1.259	.209
	Trained in testing.	182	2.460	.710		
11. Ensure adequate ventilation and lighting in the testing room.	Not trained in testing.	83	2.130	1.009	-3.417	.001
	Trained in testing.	182	2.550	.739		
12. Use “Do Not Disturb. Examinations in Progress” sign.	Not trained in testing.	83	.530	.721	-2.851	.005
	Trained in testing.	182	.840	1.009		
13. Expect and cater for possible emergencies during examinations.	Not trained in testing.	83	1.580	1.014	-.507	.613
	Trained in testing.	182	1.660	1.285		
14. Announce the remaining time (time left to complete test) at regular intervals.	Not trained in testing.	83	2.610	.621	-1.000	.318
	Trained in testing.	182	2.700	.632		

Table 9 (continued)

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
15. Stand where all students can be viewed.	Not trained in testing	83	2.720	.548	-1.349	.180
	Trained in testing	182	2.820	.510		
16. Keep all remarks to a minimum and make sure they relate to the test.	Not trained in testing.	83	1.990	.930	-.660	.510
	Trained in testing.	182	2.130	1.803		
17. Ask students to work faster during testing in order to finish on time.	Not trained in testing.	83	1.480	.980	.837	
	Trained in testing.	182	1.370	1.047		
18. Tell students the dire consequences of failure in the test they are taking.	Not trained in testing.	83	1.480	.980	.837	.403
	Trained in testing.	182	1.370	1.047		

From Table 9, only two out of the 18 test administration principles outlined in the questionnaire indicated statistically significant differences between respondents who received instruction in testing and those who did not receive instruction in testing. These are:

i. ‘Ensure adequate ventilation and lighting in the testing room’: not trained in testing (M = 2.130, SD = 1.009) and trained in testing (M = 2.550, SD = 0.739), $t_{(263)} = -3.417$, $p < 0.05$.

ii. ‘Use “Do Not Disturb. Examinations in Progress” sign when students are taking examinations’: not trained in testing (M = 0.530, SD = 0.721) and trained in testing (M = 0.840, SD = 1.009), $t_{(263)} = -2.851$, $p < 0.05$.

Inferring from the arithmetic means of the two groups for each principle, it could be concluded that respondents who received instruction in testing indicated that they did better in ensuring adequate ventilation and lighting in the testing room during testing and also in the use of the “Do Not Disturb. Examinations in Progress” sign than their counterparts who did not receive instruction in testing. The other 16 items on the questionnaire on test administration principles showed no statistically significant differences between the two groups of respondents.

Research Question 5

Which principles do SSS teachers use in the scoring of their classroom achievement essay-type tests?

Nine items on the questionnaire (items 40−48, Section D) addressed this research question. Item 39, however, was on the method that respondents used in scoring their essay-type tests. Table 10 shows the data on this item.

Table 10

Frequency Distribution of the Method Teachers Used in Scoring Essay Tests

Method	Frequency	Percent (%)
Analytic Method	245	92.5
Holistic Method	20	7.5
Total	265	100

From Table 10, it could be observed that 245 (92.5%) of the respondents used the analytic method and 18 (7.5%) used the holistic method in scoring their essays.

Table 11 shows the z-test of proportions of the observed proportion for analytic method of 92.5% against a hypothesised (test) proportion of 50%.

Table 11

Binomial Test of Proportions for Method Used in Scoring Essay-Type Tests

Method	N	Observed Proportion	Test Proportion	P-value
Holistic	20	0.08		
Analytic	245	0.92	0.50	0.000

From Table 11, the observed proportion for analytic method was found to be higher than and significantly different from the test proportion of 50%.

The result indicates that, mostly, the teachers in the study used the analytic method in scoring their essay-type tests.

Item 40 on the questionnaire asked respondents to indicate the procedure (i.e., whether script by script or item by item) they used in scoring their essay-type tests. Table 12 shows the data on this item.

Table 12

Frequency Distribution of Procedure Teachers Used in Scoring Essay Tests

Procedure	Frequency	Percent (%)
Script by Script	173	65.3
Item by Item	92	34.8
Total	265	100

From Table 12, it could be observed that 173 (65.3%) of the respondents scored their essay-type tests script by script. The rest, 92 (34.8%) scored their essay-type tests item by item.

Table 13 shows the binomial (z-test) test of proportions of the observed proportion for item by item of 34.8% against a hypothesised proportion of 50%.

Table 13

Binomial Test of Proportions for Procedure Used in Scoring Essay Tests

Procedure	N	Observed Proportion	Test Proportion	P-value
Item by item	93	35		
Script by Script	172	65	0.50	0.000

From Table 13, the observed proportion for item by item of 34.8% was found to be lower than and significantly different from the hypothesised proportion of 50%. The result indicates that on the whole, the respondents in the study used the principle (procedure) of script by script in scoring their essay-type tests. This is not an acceptable principle.

Items 41 to 48 on the questionnaire sought to find out whether or not respondents practised eight other essay-type test scoring principles. Table 14 shows the data on the items.

Table 14

Other Principles Used by Teachers in Scoring Essay-Type Tests

Principle	Yes (%)	No (%)	Total
1. Constantly follow the scoring key when scoring.	89.1	10.9	100
2. Randomly reshuffle answer scripts of essay-type tests after scoring each set of items.	26.0	74.0	100
3. Score all responses to a particular item at a Sitting without interruption.	42.6	56.2	100
4. Score essay-type tests only when you are physically sound and mentally alert.	93.6	6.5	100
5. Keep previously scored items out of sight When scoring the rest of the items.	57.7	42.2	100
6. Provide comments on the answer scripts for students to aid learning.	89.4	10.5	100
7. Score answer scripts of essay-type tests with the names of the students unknown.	74.3	25.7	100
8. Score the mechanics of expressions such as penmanship, general neatness and spelling, separately from subject matter correctness.	61.9	38.1	100

From Table 14, with regard to the principle of 'constantly following the marking scheme when scoring', 89.1% of the respondents indicated that they did it while 10.9% indicated they did not do it. On 'randomly reshuffling answer scripts after scoring each set of items', 26.0% of the respondents indicated they applied this principle whereas 74.0% indicated they did not

apply the principle. On scoring all responses to a particular item at a sitting without interruption, the response pattern remained unchanged. The majority (56.2%) pointed out they were not able to do this while the minority (42.6%) pointed out they were able to do it. With the principle of 'scoring essay-type tests only when the scorer is physically sound and mentally alert', there was a reverse of the response pattern. As many as 93.6% indicated they applied this principle whereas only 6.5% indicated they did not apply this principle.

On the principle of 'keeping previously scored items out of sight when scoring the rest of the items', the response pattern remained as before. The majority (57.7%) of the respondents pointed out they complied with this principle while the minority (42.2%) pointed out they did not comply with this principle. Next, was the provision of 'comments on answer scripts for students to aid learning.' A large proportion (89.4%) of the respondents indicated they did this while 10.5% indicated they did not do this. When the respondents were asked whether they scored essay-type tests with the names of the students known, the response pattern resumed the initial trend. A larger proportion (74.3%) of the respondents indicated they scored with the names of the students unknown while 25.7% showed they scored with the names of the students known.

Last on the other principles of scoring essay-type tests was 'scoring the mechanics of expressions separately from subject matter correctness.' On this, 61.9% of the respondents indicated they scored the mechanics of expressions separately from subject matter correctness while 38.1% indicated they did not score the mechanics of expressions such as penmanship, general neatness and spelling separately from subject matter correctness.

In conclusion, teachers in the study generally indicated that they practised six of the other eight essay-test scoring principles. These are:

i. Following the marking scheme constantly when scoring essay-type tests.

ii. Scoring essay-type tests only when physically sound and mentally alert.

iii. Keeping previously scored items out of sight when scoring the rest of the items.

iv. Providing comments on the answer scripts for students to aid learning.

v. Scoring the answer scripts of essay-type tests with the names of the students unknown.

vi. Scoring the mechanics of expressions such as penmanship, general neatness and spelling separately from subject matter correctness.

It can, therefore, be concluded that teachers in the study generally indicated they did not use three of the essay-type test scoring principles. These are:

i. Scoring of essay-type tests item by item.

ii. Randomly reshuffling the answer scripts after scoring each set of items

iii. Scoring all responses to a particular item at a sitting without interruption.

Research Question 6

What differences exist between SSS teachers who received instruction in educational measurement and those who did not receive instruction in educational measurement in terms of the principles used in the scoring of essay-type tests?

To answer this question, the responses to the items on the principles of essay-type test scoring (see Appendix B) were coded and scored. The mean scores, standard deviations and t-values of respondents who received instruction in testing (i.e., educational measurement) and those who did not receive instruction in testing for each principle are shown in Table 15. The level of significance was 0.05.

Table 15

Results of t-test of Independence for Application of Essay-Type Test Scoring Principles by Respondents

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
1. Score essay-type tests item by item	Not trained in testing.	83	.280	.450	-1.755	.081
	Trained in testing.	182	.380	.488		
2. Constantly follow the marking scheme when scoring essay tests	Not trained in testing.	83	.380	.480	-.976	.311
	Trained in testing.	182	.440	.489		
3. Randomly reshuffle the answer scripts of essay tests after scoring each item.	Not trained in testing	83	.180	.387	-2.036	.043
	Trained in testing	182	.290	.456		
4.Score all responses to a particular item at a sitting without interruption.	Not trained in testing.	83	.390	.490	-.996	.321
	Trained in testing.	182	.450	.499		

Table 15 (continued)

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
5. Score essay-type tests only when physically sound and mentally alert.	Not trained in testing.	83	.920	.280	-1.179	.241
	Trained in testing.	182	.960	.206		
6. Keep previously scored items out of sight when scoring the rest of the items.	Not trained in testing.	83	.510	.503	-1.657	.100
	Trained in testing.	182	.620	.488		
7. Provide comments on the answer scripts for students to aid learning.	Not trained in testing.	83	.830	.377	-2.134	.035
	Trained in testing.	182	.930	.258		
8. Score the answer scripts of essay-type tests with names of students unknown.	Not trained in testing.	83	.710	.456	-.817	.414
	Trained in testing.	182	.760	.429		

Table 15 (continued)

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
9. Score the mechanics of expressions such as penmanship, general neatness and spelling, separately from subject matter correctness.	Not trained in testing.	83	.610	.490	-.185	.853
	Trained in testing.	182	.630	.485		

The t-test results indicated statistically significant differences in two of the nine test scoring principles between the two groups of respondents. These are:

i. 'Randomly reshuffle the answer scripts after scoring each set of items': not trained in testing ($\underline{M}$ = 0.180, $\underline{SD}$ = 0.387) and trained in testing ($\underline{M}$ = 0.290, $\underline{SD}$ = 0.456), $\underline{t}_{(263)}$ = -2.036, $\underline{P} < 0.05$.

ii. 'Provide comments on the answer scripts when scoring for students to aid learning': not trained in testing ($\underline{M}$ = 0.830, $\underline{SD}$ = 0.377) and trained in testing ($\underline{M}$ = 0.930, $\underline{SD}$ = 0.258), $\underline{t}_{(263)}$ = -2.134, $\underline{p} < 0.05$.

The arithmetic means of the two groups of respondents in each case above, shows that respondents who received instruction in testing indicated that they applied the two principles better than their counterparts who did not

receive instruction in testing. The results of the t-test further indicated that there were no statistically significant differences between the two groups of respondents with respect to the other seven essay test scoring principles (Please refer to Table 15).

Research Question 7

How do SSS teachers interpret the results of their classroom achievement tests?

This question sought to find out how SSS teachers go about the interpretation of the results of their achievement tests. Four items on the questionnaire (Items 49–52, Section D) addressed this research question.

Item 49 sought to find out the methods that SSS teachers used in their test-score interpretation. Table 16 shows the frequency and percentage distribution of the participants' responses on item 49.

Table 16

Frequency Distribution of Method(s) Teachers Used in the Interpretation of Test Results

Method of Interpretation	Frequency	Percent (%)
Norm-referenced Interpretation	99	37.5
Criterion-referenced Interpretation	25	9.4
Norm-referenced and Criterion-referenced Interpretations	141	53.1
Total	265	100

The responses in Table 16 show that the majority (53.1%) of the respondents used a combination of the norm-referenced and criterion-referenced approaches in interpreting the results of their achievement tests. Table 17 shows the frequency and percentage distributions of the method(s) employed by teachers who used norm-referenced interpretation. Table 17 is the analysis of the responses to item 50.

Table 17

Frequency Distribution of Method(s) Employed by Teachers who Used Norm-referenced Interpretation

Method	Frequency	Percent (%)
Simple Ranking of Raw Scores	177	73.7
Percentile Ranks	27	11.1
Simple Ranking of Raw Scores and Percentile Ranks	36	15.2
Total	240	100

Table 17 indicates that of the 240 teachers who used norm-referenced interpretation, 73.7% of them used simple ranking of raw scores. It could, therefore, be concluded that most teachers used simple ranking of raw scores in norm-referenced interpretation.

Table 18 shows the frequency and percentage distributions of performance standard or cut-off score setting methods employed by teachers who used criterion-referenced interpretation. This is analysis of responses to item 51.

Table 18

Frequency Distribution of Performance Standard Setting Methods Employed by Teachers who Used Criterion-referenced Interpretation

Performance Standard	Frequency	Percent (%)
Fifty percent correct score.	105	63.3
Arbitrary standard, e.g., 75% correct, adjustable up or down as various conditions and experiences are considered.	28	16.9
Others (Forty percent correct score)	33	19.9
Total	166	100

From Table 18, of the 166 teachers who indicated they used criterion-referenced interpretation, the majority (63.3%) pointed out that in setting a standard of performance or cut-off score, they used '50 percent correct score.' This is followed by 19.9% who indicated that they used '40 percent correct score' as the standard of performance or cut-off score. In conclusion, in setting a standard of performance or cut-off score in criterion-referenced interpretation, teachers used mostly '50 percent correct score.'

Respondents were asked to indicate how they handled the students who failed to master stated instructional objectives in their classrooms based on their test results. This was an open ended item (item 52, Appendix B), where respondents provided their own responses. The responses are shown in Table 19.

Table 19

Frequency Distribution of how Low-Achieving Students are Handled.

	Frequency	Percent (%)
Remedial Teaching	91	34.4
Extra attention during instructional hours	95	35.9
Others	79	29.7
Total	265	100

Table 19 indicates that the largest proportion of 35.9% which represents 95 of the 265 respondents gave extra attention to their low-achieving students during normal instructional hours mainly in the form of letting their teaching centre on them. The main reason of these teachers was that the SSS syllabus is very extensive and so if they set time aside for remedial teaching, they would not be able to complete the syllabus. Some of the responses they gave are:

i. "I let my teaching centre on such students when teaching since there is no time for remedial teaching".

ii. "I give them special attention in the classroom by helping them to overcome their difficulties".

iii. "Since there is no time for remedial teaching, I sometimes give them more attention when teaching".

Table 19 also indicates that 91 (34.4%) of the respondents organised remedial teaching for their low-achieving students. Some of these teachers indicated that they are able to meet such students during the afternoons after classes and during the weekends. Others also pointed out that they

recommended extra tuition with fees to the parents of such students for their wards and they embarked on such tuition only when the parents agreed to pay.

Table 19 again reveals that 79 (29.7%) of the respondents handled their low-achieving students in other ways. Some of the responses they gave are:

i. "I do normally set different questions which are lower in difficulty than what they answered and let them try their hands on the them. I also advise them to read many textbooks and attempt some of the questions in them".

ii. "I counsel and encourage them to work harder".

iii. "They cannot be given individual attention because of large class size, so I encourage them to learn outside class hours and also join study groups".

It could be concluded from the analysis above that teachers gave extra attention to or organised remedial teaching for their low-achieving students. Teachers also handled their low-achieving students in other ways in their bid to help them.

Research Question 8

What differences exist between SSS teachers who received instruction in testing (i.e. educational measurement) and those who did not receive instruction in testing in terms of how they interpret the results of their classroom achievement tests?

To answer this question, the responses to the items on interpretation of the results of classroom achievement tests (see Appendix B) were coded and

scored. The mean scores, standard deviations and t-values of respondents who received instruction in testing and those who did not receive instruction in testing for each item on the interpretation of the results of tests are shown in Table 20. The level of significance was 0.05.

Table 20

Results of t-test of Independence on Interpretation of the Results of Classroom Achievement Tests by Respondents

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
1. Method(s) used in the interpretation of test results.	Not trained in testing.	83	1.370	0.487	-2.942	0.004
	Trained in testing.	182	1.570	0.497		
2. Method(s) employed in norm-referenced interpretation.	Not trained in testing.	61	1.36	0.484	-1.603	0.113
	Trained in testing.	179	1.24	0.429		

Table 20 (continued)

Principle	Teacher Status	N	Mean	Std Deviation	t-value	P-value
3. Method of setting performance standard (cut-off score) in criterion-referenced interpretation.	Not trained in testing.	30	0.80	0.407	1.331	0.192
	Trained in testing.	136	0.90	0.295		
4. How teachers handled their low-achieving students.	Not trained in testing.	83	0.950	0.215	-4.003	0.000
	Trained in testing.	182	1.120	0.450		

The t-test results indicated statistically significant differences in two of the four items on test-score interpretation procedures between teachers who received instruction in testing and those who did not. These were:

i. 'Method(s) used in the interpretation of test results': not trained in testing ($\underline{M}$ = 1.370, $\underline{SD}$ = 0.847) and trained in testing ($\underline{M}$ = 1.570, $\underline{SD}$ = 0.497), $\underline{t}_{(263)} = -2.942$, $\underline{p} < 0.05$.

ii. 'How teachers handle their low-achieving students': not trained in testing ($\underline{M}$ = 0.950, $\underline{SD}$ = 0.215) and trained in testing ($\underline{M}$ = 1.120, $\underline{SD}$ = 0.450), $\underline{t}_{(263)} = -4.003$, $\underline{p} < 0.05$.

The arithmetic means of the two groups in each case above indicate that respondents who received training in testing indicated that they did better

in the matters of test-score interpretation than their counterparts who did not receive training in testing. The results of the t-test further indicated there are no statistically significant differences between the two groups of respondents with respect to the method(s) employed in norm-referenced interpretation and method of setting performance standard (cut-off score) in criterion-referenced interpretation. (Please refer to Table 20).

Discussion of Research Findings

In this section, the findings are discussed in relation to:

i. Test construction principles used by teachers

ii. Test administration principles used by teachers

iii. Essay-type test scoring principles used by teachers

iv. Methods of test-score interpretation used by teachers

Test Construction Principles Used by Teachers

The first research question sought to find out the kind of principles that SSS teachers used in the construction of their classroom achievement tests. The findings indicated that, in general, teachers in the study practised seven out of 10 principles in test construction:

In all the principles that teachers practised, they pointed out they applied them either always or very often. These results are consistent with six of the eight general principles of achievement test construction put forward by Tamakloe et al. (1996), Amedahe and Gyimah (2003) and Etsey (2004). These findings are in the right direction for classroom test construction practices.

The findings of the study showed that teachers did not practise three test construction principles to any appreciable extent. The contributions of these three principles to the quality of any achievement test cannot be underestimated. For instance, the table of specifications "makes sure that justice is done to all the topics covered in the course, helps the teacher to determine the content validity of the test, helps to weigh the score distribution fairly, avoids overlapping in the construction of test items, and helps students to determine the content and behavioural areas where they have difficulty" (Etsey, 2004, p.21). Ebel and Frisbie (1991) summed the importance of a table of specifications by pointing out that, a table of specifications is a planning guide for ensuring adequate representation of content and abilities in a test. Although the table of specifications is not useful with all item formats, especially the essay-type test format, the objective-type tests are used in all the SSS course subjects and as such, teachers in the study must make use of it.

Preparing an initial draft of more items than needed in the test ensures that the teacher has an adequate replacement of test items after the items have been reviewed and edited and the defective ones discarded. This avoids time wastage in going back to start all over again to construct new test items to replace defective ones. Hence, since teachers in the study indicated that they did not practise this principle, they would in no doubt be wasting precious time to construct new items whenever the items are reviewed and edited and defective ones discarded.

The final evaluation of the test on the other hand, clears the test of all inadequacies and certifies it as good for submission to be processed for subsequent administration. It could therefore be deduced that failure to apply

these three principles has very grievous implications for classroom achievement test development.

Violations of item writing rules such as the three found in the study would certainly result in faulty items. This was the case in the study by McMorris and Boothroyd (1992) in the USA which reported that faults were found in 35% of completion items and 20% of multiple-choice items on teachers' tests. The impacts of item faults on teacher-made tests are diverse. Items may be made easier by faults (Dunn & Goldstein, 1959; Haladyna & Downing, 1989a; 1989b; McMorris, Brown, Snyder & Pruzek, 1972, cited in McMorris & Boothroyd, 1992). Tests containing item faults are inconsistent with the principle that test items should elicit only the behaviours which the test developer desires to observe. Faulty items would obviously introduce extraneous variance, which would in turn, reduce somewhat the validity of descriptions and decisions based on the test (McMorris & Boothroyd, 1992).

Mention should be made of the fact that the findings above do not wholly confirm the finding of the study of Amedahe (1989) that to a great extent, secondary school teachers did not follow the basic prescribed principles of classroom test construction. This difference in findings could be attributed to the difference in the percentage of participants with training in testing in the two studies. This present study had 68.7% of the participants with training in testing while Amedahe (1989) had 62% of the participants with training in testing. On the assumption that training leads to increased competence, this factor could be responsible for the difference in findings in the two studies.

Mention should also be made of the fact that the findings that teachers write their test items in advance of the test date in order to permit reviews and editing and also review their test items after they have been set aside for a few days, are at variance with research findings in England reviewed in the articles by Crooks (1988) and Black (1993b). As stated in the literature, Crooks and Black wrote that teachers in England generally do not review the assessment questions that they use and do not discuss them critically with peers, so there is a little reflection on what is assessed.

The finding that teachers in the Ashanti Region of Ghana, generally, do not use a table of specifications to determine the items on their tests is, however, consistent with the findings of the studies in the USA by Marso and Pigge (1992) and Oescher and Kirby (1998). They concluded that teachers generally lack competence in the use of the table of specifications.

One would have expected that receipt of instruction or training would be a significant factor in increasing competence and that teachers who received instruction in educational measurement would perform far better in the degree of application of test construction principles than their counterparts who did not receive instruction in educational measurement. The second main finding of the study, however, was that there were statistically significant differences between teachers who received instruction in educational measurement during their pre-service training and those who did not receive any instruction in educational measurement, in terms of their level of application of five of the 10 test construction principles.

The results of the t-test of independence (Please refer to Table 7) showed that teachers who had training in testing indicated they did better in

terms of their level of application of the five principles than teachers who did not receive any training in testing. These findings indicate that the test construction practices of teachers who received training in testing and their counterparts who did not receive any training in testing are not much different. However, there was an indication, at least, of a bearing of theory on practice. It is apparent that the presence of teachers who had training in testing is being felt on the test construction scene, but not to a greater extent. This is because it was only in five (half of the number of principles listed) out of the 10 principles that teachers who received training in testing indicated they did better in terms of their application than their counterparts who did not receive training in testing.

These findings partly support the assertion of Quaigrain (1992) that even teachers who have gone through formal training in testing techniques may not adopt the ideas they have learnt from their professional training but rather may be inclined towards the techniques they were exposed to when they were students. This is supported by Ort (cited in Quaigrain, 1992) who saw classroom testing as being debased because teachers tend to repeat the testing techniques and ideas they experienced during their own school days.

The findings also support the finding of Quaigrain (1992) that there was a significant positive relationship between pre-service training in measurement and evaluation and actual testing practices in the field. Quaigrain (1992) obtained a point-biserial correlation coefficient (r_{pbis}) of 0.43, which was a weak positive relationship. The findings also partly support the finding of Amedahe (1989) that there was no significant difference between the procedure used in constructing classroom achievement tests by teachers who

received instruction in testing and those who did not, in terms of the accuracy in following prescribed test construction principles.

Test Administration Principles Used by Teachers

The third main finding of the study was an answer to research question three which sought to find out the kind of principles that SSS teachers used in the administration of their classroom achievement tests. The results showed that teachers indicated they practised a total of 12 out of the 18 principles outlined in the questionnaire.

With regard to the violations of the psychological conditions indicated by the results (the fact that teachers gave tests to their students irrespective of whether or not it was just before or after an important event, and also did not minimise test anxiety in students during testing), (Refer to Table 8), it could be said that, generally, teachers in the study did not observe good psychological conditions when testing their students. This is actually a disturbing phenomenon since psychometricians such as Nunnally (1972) and Gronlund (1985) have asserted that poor psychological conditions such as asking students to hurry up in order to complete the test on time and other threatening behaviours of invigilators affect students' performance in one way or the other, negatively.

The findings in terms of the physical conditions are by and large in line with the test administration guidelines proposed by Tamakloe et al. (1996) and Etsey (2004). These are, basically, with the intent of providing examinees with a fair chance to demonstrate their ability on what is being measured. There is, however, one issue of teachers, generally, not using the 'Do Not Disturb.

Examinations in Progress' sign when students are taking examinations to ensure total silence in the vicinity. From the researcher's point of view, however, this situation did not have any adverse effect on the conduction of examinations in the schools. This is because in the 20 observations carried out on the conditions under which students take their examinations, there were only four cases of intermittent noisy environments.

Finally, the findings here did not wholly confirm the finding of Amedahe (1989) that teachers generally observed good physical and psychological conditions when administering classroom achievement tests. This is because this study clearly indicates that teachers observed good physical conditions but did not observe a considerable number of good psychological conditions.

On whether there were any differences between teachers who had training in testing and those who did not have training in testing with respect to the degree of application of test administration principles, the t-test results showed statistically significant differences in only two of the 18 principles in favour of teachers who had training in testing. These were the principles of ensuring adequate ventilation and lighting in the testing room, and the use of a 'Do Not Disturb. Examinations in Progress' sign during testing. It could be deduced from this finding that, there is some level of impact of theory on practice, even though, the level of impact is quite subtle.

Essay-Type Test Scoring Principles Used by Teachers

The fifth main finding of the study was that teachers practised six of the nine test scoring principles outlined in the questionnaire. The findings are

consistent with six of the ten general principles of test scoring proposed by Mehrens and Lehmann (1991), Tamakloe et al. (1996), Amedahe and Gyimah (2003), and Etsey (2004). Mention should also be made of the fact that the finding that teachers provided comments on the answer scripts of essay-type tests to facilitate learning is a very positive step in formative assessment. This is, however, at variance with the finding of the study in England reviewed in the articles by Crooks (1988) and Black (1993b) that teachers over-emphasised the grading function while they under-emphasised the learning function.

It is an encouraging news that teachers in the study applied a total of six out of nine test scoring principles. Nevertheless, due to the complexities involved in the scoring of the essay-type tests, even an omission of one of the scoring guidelines has the potential of causing inconsistencies in the test scores and thereby render them unreliable. The magnitude of teacher activities, inactions and characteristics in influencing score reliability and grading of students was confirmed in a study by Ashburn and cited by Quaigrain (1992). It was found that:

> The passing or failing of 40 percent of students depends not on what they know or do not know, but on who reads the papers. The passing or failing of about 10 percent depends on when the papers are read (p. 103).

The fact that teachers in the study did not score their essay-type tests, item by item, meant that, there was a high possibility of the carryover effect on the scores derived from such tests. This would in no doubt render the scores inconsistent. Again, the fact that teachers in the study did not randomly

reshuffle the answer scripts after scoring each item meant the possibility of the introduction of bias into the scoring process as a result of the position of one's scripts. According to Mehrens and Lehman (1991), this is especially significant when teachers are working with high- and low-level classes and read the best scripts first or last. Lastly, the inability of teachers to score all responses to a particular item at a sitting without interruption meant a possible variation of the scorer's standards due to excessive interruptions in the course of scoring. This would render the scores unreliable.

The results of the study further indicated that teachers generally used the analytic method in scoring their essay-type tests. This was in response to item 39 of the questionnaire. As many as 245 (92.5%) of the teachers pointed out they used the analytic method while only 18 (6.5%) of the teachers pointed out they used the holistic method. This is a very positive indication for achievement test scoring since the analytic method ensures objectivity and consistency in scoring and higher reliability of test scores (Amedahe & Gyimah, 2003; Mehrens & Lehmann, 1984; Tamakloe et al, 1996). This finding positively confirms the findings of Amedahe (1989) and Quaigrain (1992) that teachers in the schools used the analytic method in scoring their essay-type tests. In the study of Amedahe (1989), most teachers preferred the analytic method to the holistic method for reasons they assigned as follows:

a) It ensures uniform scoring criteria and minimised subjectivity.

b) Fairness in scoring is maintained.

c) It avoids biases in the test scores.

d) It gives accurate assessment and facilitates easy marking.

A possible reason for some teachers using the holistic method in scoring their essay-type tests is that the procedure entails less time in the rating of scripts.

The sixth main finding of the study was that teachers who received instruction in testing indicated that they applied two of the nine test scoring principles more frequently than their counterparts who did not receive any instruction in testing. There is again an indication of the influence of theory on practice. This is quite insignificant since it is only two out of the nine test scoring principles that produced statistically significant differences in favour of teachers who had training in testing. This finding rightly confirms the finding of Quaigrain (1992) that there was a bearing of pre-service training in measurement and evaluation on competence in using essay-type tests.

Methods of Test Score Interpretation Used by Teachers

The most common approach to the problem of interpreting test scores of teacher-made tests is norm-referencing (Amedahe & Gyimah, 2003). Amedahe and Gyimah (2003), however, point out that the approach has the major limitation of not giving any indication of how well a student performed in terms of mastering what was taught. The fact that teachers employed both norm-referenced and criterion-referenced approaches in the interpretation of their tests, therefore, comes as welcoming news for classroom achievement test development in Ghanaian schools. From Table 16, more than half of the teachers 141(53.1%) indicated they used both approaches of test score interpretation. From the literature, as cautioned by Gronlund (1988), if the two interpretation approaches are to be combined for a single test, then it is most

likely to be effective where norm-referenced interpretation is added to the performance description of a criterion-referenced test.

The second finding indicated the predominance of the use of simple ranking of raw scores by teachers who employed norm-referenced interpretation. This means that the use of percentile ranks and the stanine system of standard scores is not popular with teachers. This implies that teachers are not able to describe a student's relative position in a group in terms of the percentage of group members scoring at or below the student's score, so as to give a better idea of the quality of a student's performance relative to that of other members of the class. It also implies that teachers are not able to compare a student's relative achievement on different tests so as to be able to determine the particular subjects in which a student is doing well. This is a setback in the area of test-score interpretation that needs to be addressed.

The third finding under test-score interpretation is that, teachers who used criterion-referenced interpretation used mostly 50 percent correct score as the performance standard or cut-off score. The basis for the 50 percent correct score might be that, on a scale of 0% to 100 %, 50% is half of whatever task that is given to students. This is, however, not consistent with the position of a measurement expert such as Gronlund (1988) who proposed that for formative assessment, a relatively simple and practical procedure for setting standard of performance or cut-off score is to arbitrarily set a standard and then adjust it up or down as various conditions and experiences are considered. He gave an example of setting the cut-off score of a multiple-choice test at 85 percent correct score. This is high enough on a scale of 0% to

100% and can be used as a yardstick to certify students as having attained learning targets or stated instructional objectives.

On the fourth finding in this section, the means through which teachers handled their low-achieving students based on their test results to a large extent were in the right direction. Organising remedial teaching for low-achieving students is a very concrete means of helping them. It is worth noting that in order for the remedial teaching to meet the needs of individual students, an item-by-item analysis must be done for the unmastered objectives to pinpoint students' errors (Gronlund, 1988).

Giving low-achieving students extra attention during instructional hours can be said to be good. However, it has the problem of disadvantaging the higher-achievers in the class. To them, it might seem a boring experience and a feeling of a sense of neglect. It would, therefore, be best to isolate or group low-achieving students and give them remedial teaching at their ability levels.

Lastly, helping low-achieving students through means such as counselling, advising and encouragement is good, but in addition, these students could be taken through concrete measures such as extra teaching.

The eighth finding of the study was that teachers who received instruction in testing indicated that they did better in two of the four issues on test-score interpretation than their counterparts who did not receive any instruction in testing. The indication that teachers trained in testing did better than their counterparts who were not trained in testing in two out of four issues on test- score interpretation gives a confirmation of the earlier findings in this study of the bearing that training has on practice.

On the first finding which is on the method(s) used in test-score interpretation, teachers who had no training in testing had a mean of 1.370 while teachers with training in testing had a mean of 1.570. This means that teachers with training in testing scored higher on item 49 of the questionnaire which means that majority of them indicated that they used both norm-referenced and criterion-referenced approaches in their test-score interpretation. This is in the right direction. It is in agreement with the assertion of Gronlund (1988) and Nitko (1996), that both methods of test-score interpretation are important to understand how well students are learning. NRI tells how an individual's test performance compares with that of others while CRI tells in specific performance terms what an individual can do without reference to the performance of others, so that if necessary, remedial work can be planned.

On the second finding of how teachers handled their low-achieving students, most teachers with training in testing gave positive, concrete and constructive means of helping low-achieving students in their classes such as organising remedial teaching alongside normal teaching for them, letting the teaching centre on them during instructional hours and giving them extra work to do in addition to normal classroom work. Most teachers with no training in testing, on the other hand, gave means of helping low-achieving students that were not readily concrete such as recommending them to be repeated in class, warning them to work harder, and leaving them to their fate because there is no time. There was one respondent who indicated he would recommend the withdrawal of such students from the school. The account above resulted in teachers with training in testing indicating that they did better in the ways they

handled their low-achieving students than their counterparts with no training in testing.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Overview

The study was a descriptive survey that investigated the testing practices of SSS teachers of English Language, Core Mathematics and Integrated Science with respect to the construction, administration and scoring of their classroom achievement tests and the interpretation of the results of these tests. The study was primarily aimed at finding out whether the procedures used by teachers in the construction, administration and scoring of classroom achievement tests and the interpretation of the results of these tests were in line with the principles and guidelines prescribed by measurement specialists. The study also sought to find out whether any differences existed between teachers who received instruction in educational measurement and those who did not, in terms of their testing practices.

The study was conducted in the Ashanti Region of Ghana. A sample of 10 districts was randomly selected from the 21 districts in the Ashanti Region. The cluster sampling method was used to select 26 SSSs from a total of 56 SSSs in the 10 sampled districts. The sample for the study comprised 265 teachers teaching the three aforementioned subjects.

A 52-item questionnaire and a 17-item observation guide were the main instruments for data collection. The data collected were analysed mainly

by frequency and percentage tables, the t-test for independent samples, and the binomial (z-test) test.

Summary of Findings

The following are the main findings from the data analysis.

Under test construction, teachers indicated that they practiced seven out of 10 principles. Also, teachers who had training in testing indicated that they used five out of 10 principles more frequently than their counterparts who had no training in testing.

Under test administration, teachers indicated they applied 12 out of 18 principles. Furthermore, teachers who had training in testing indicated that they used two out of 18 principles more frequently than their counterparts who had no training in testing.

Under scoring of essay-type tests, teachers indicated that they applied six out of nine principles. In addition, teachers who had training in testing indicated that they used two out of nine principles more frequently than their counterparts who had no training in testing.

Under test-score interpretation, teachers indicated that they used: both norm-referenced and criterion-referenced approaches; simple ranking of raw scores in norm-referenced interpretation; and 50 percent correct score as the cut-off score in criterion-referenced interpretation. Additionally, teachers helped their low-achieving students in the form of organising remedial teaching for them, giving them extra attention during instructional hours, and counselling or encouraging them.

Lastly, teachers who received instruction in testing indicated that they did better by way of applying all the two methods of test-score interpretation, and also, handled their low-achieving students better than their counterparts who had no training in testing.

Conclusions

The results of the study indicated that on test construction, administration and scoring, teachers generally reported they applied a considerable number (more than half) of the principles outlined under each of them in the questionnaire. It could therefore, be concluded that, to a great extent, teachers in the Ashanti Region of Ghana followed the basic principles prescribed by testing experts in the construction, administration and scoring of their classroom achievement tests.

On test-score interpretation, since the type of interpretation used depends on the uses to which the test results would be put, it could be concluded that, to a great extent, teachers in the Ashanti Region of Ghana followed the techniques prescribed by testing experts in their test-score interpretation.

The results of the study further indicated that under the construction, administration and scoring of test and the interpretation of the test results, teachers with pre-service training in educational measurement reported they did better in the application of some of the principles and issues that were raised, than their counterparts who did not receive training in educational measurement. It could therefore, be concluded that, pre-service training in educational measurement has an impact on actual testing practice. This gives

an indication that all is not lost with respect to the attempts being made by teacher training universities in Ghana to train pre-service teachers to become competent test developers and users.

Limitations of the Study

The study is an exploratory one which only gives the state of affairs concerning the construction, administration and scoring of teacher-made tests and the interpretation of the test results. It, therefore, does not establish any cause-effect relationship in any of the four aspects of the problem under study.

In the ideal situation, a nationwide study is required. This would have given much credence to any generalisations made. The time for the study and the resources available, however, made this impracticable. Hence, the selection of the Ashanti Region and even a sample drawn from the SSS teacher population.

Not all the subjects taught in the SSS were included in the study. This is because the subjects are so many and that time and resources available would have been a hindrance to the inclusion of all of them in the study. In view of this, Core Mathematics, English Language and Integrated Science were considered for the study, for the reason that these are core subjects that all students offer. The fewer number of subjects taught at the SSS that were used for the study, therefore, might affect generalisation to the whole senior secondary school system.

Lastly, a major limitation of the study was the unenthusiastic attitude of teachers toward research work and especially completing of questionnaires. This resulted in 265 of the questionnaires being retrieved which represented

only 52.06% of the 509 questionnaires distributed. This affects the generalisability of the study.

Recommendations

In view of the above research findings and the conclusions arrived at, the following recommendations are made.

1. Teaching and testing are systematically interwoven. According to Stiggins (1999), a typical teacher can spend a third to a half of his professional time on assessment related activities. Therefore, on the finding that teachers did not apply all the basic principles and techniques in their test development, it is recommended that every teacher should be given formal training in educational measurement and evaluation during pre-service training to equip him/her for the tasks and demands on the job. In the Ghanaian situation where non-professional teachers are employed to teach, frequent in-service training is recommended for all SSS teachers to train the non-professional ones and to sharpen the skills of the already trained ones. This would in no doubt improve both the quality of the tests constructed by teachers and the information derived from the use of such tests.
2. The results of the study indicated a bearing of training on practice. The researcher, therefore, recommends that the University of Cape Coast and the University of Education, Winneba, in their undergraduate courses in educational measurement and evaluation should always stress on the practical application of the theoretical knowledge students

acquire. In this regard, there could be assignments or end of semester papers where student teachers construct, administer and score tests and interpret the results of such tests for lecturers to assess and evaluate the quality of these tests and the interpretation of the test scores. This, however, calls for more lecturers to be employed by the universities to handle the increasing number of students admitted every year.

3. In test administration, the observations done indicated that one principle that some teachers and school authorities violated to a great extent was keeping all remarks in the testing room to a minimum and making sure they are related to the test. The researcher recommends that all announcements that are not related to the test, such as those about school fees and students who owe individual teachers in purchase of books and in extra classes, must be made at other places other than in the testing room.

Suggestions for Further Research

The following are recommended for future research.

1. The study was exploratory in nature. In order to accept or refute the findings of the study and generalise them for the whole of the country, it is suggested that the study is replicated in other regions of the country at the basic and secondary levels.
2. It is also suggested that research that will centre on the characteristics of teacher-made tests be undertaken. This is because this study was on principles teachers used in their test development without touching on the main determinants of the quality of these tests, which are how

reliable are the scores from these tests, and how valid are the uses to which the scores from these tests are put.

REFERENCES

Amedahe, F. K. (2004). *Notes on educational research.* Unpublished document, University of Cape Coast, Ghana.

Amedahe, F . K. (1989). *Testing practices of secondary schools in the Central Region of Ghana*. Unpublished master's thesis, University of Cape Coast, Cape Coast, Ghana.

Amedahe, F. K. (2000). *Continuous assessment.* Unpublished paper, University of Cape Coast, Ghana.

Amedahe, F. K. & Gyimah, K. A. (2003). *Measurement and evaluation,* Cape Coast, Ghana: Centre for Continuing Education.

American Association for the Advancement of Science (AAAS). (1998). *Blueprint online—project 2061*. Available: http://www.project 2061.org. online.

Anastasi, A. (1982). *Psychological testing*. New York: Macmillan Publishing Company.

Ary, D., Jacobs, L. C., & Razavieh, A. (1990). *Introduction to research in education*. (4th ed.). Forth Worth: Holt, Rinehart & Winston Inc.

Barton, P. E. & Coley, R. J. (1994). *Testing in America's schools. Policy information report*. Princeton, New Jersey: Educational Testing Service.

Black, P. J. (1993b) Formative and summative assessment by teachers. *Studies in Science Education.* 21, 49 – 97.

Black, P. & William, D. (1998). Teachers' practices in formative evaluation. *Assessment in Education: Principles , policy and practice*. 5 (1), 7—68.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harpers and Brothers Publishers.

Crooks, T. J. (1998). The impact of classroom evaluation practices on students. *Review of Educational Research*. 58, 438– 481.

Cunningham, G. K. (1986). *Educational and psychological measurement.* New York: Macmillan Publishing Company.

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

Etsey, Y. K. A. (2004). *Educational measurement and evaluation.* Lecture notes on EPS 203. Unpublished document, University of Cape Coast, Ghana.

Flanagan, D, Genshaft, J. L. & Harrison, P. L. (1997). *Intellectual assessment, tests, and issues.* New York: The Guilford Press.

Fraenkel, J. R. & Wallen, N. E. (2000). *How to design and evaluate research in education.* (4th ed.). New York: McGraw-Hill, Inc.

Gay, R. L. (1992). *Educational research: Competencies for analysis and application.* (4th ed.). New York: Macmillan Publishing Company.

Gronlund, N. E. (1985). *How to construct achievement tests.* (3rd ed.) New Jersey: Prentice-Hall, Inc.

Gronlund, N. E. (1988). *How to construct achievement tests.* (4th ed.) New Jersey: Prentice-Hall, Inc.

Guilford, J. P. & Fruchter, B. (1983). *Fundamental statistics in psychology and education* .(7th). Tokyo: McGraw-Hill.

Gyimah, K. A. (2002). *An evaluation of the practice of continuous assessment in the secondary schools in the Ashanti Region of Ghana*. Unpublished master's thesis, University of Cape Coast, Ghana.

Joint Committee on Standards for Educational and Psychological Testing (JCSEPT). (1999). *Standards for educational and psychological testing.* Washington D.C: American Educational Research Association.

Kubiszyn, T. & orich, G. (1984). *Educational testing and measurement: Classroom application and practice.* New Jersey: Scott, Foresman and Company.

Linn, R. L. & Gronlund, N. E. (1995). *Measurement and assessment in teaching.*(7th ed.). New Jersey: Merrill, Prentice-Hall.

Marso, R. N. & Pigge F. L. (1989). *Staff development implications from a state-wide assessment of classroom teachers' testing skills and practice.* (ED312309). Available: http://eric.ed.gov/ERICWebPortal/home. on line.

Marso, R. N. & Pigge F. L. (1989). *A summary of published research: Classroom teachers' knowledge and skills related to the development and use of teacher-made tests.* (ED346148). Available: http://eric.ed.gov/ERICWebPortal/home. on line.

McBride, M. (1999). *Letting students shine: Assessment to promote students learning.* Available: www.ets.org/letstalk. on line.

McDaniel, E. (1994). *Understanding educational measurement*. Madison: Brown & Benchmark Publishers.

McMorris, R. F &.Boothroyd, R. A. (1992). *Tests that teachers build: An analysis of classroom tests in Science and Mathematics.* (ED350348). Available: http://eric.ed.gov/ERICWebPortal/home. on line.

Mehrens, W. A. & Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology*. (3rd ed.) New York: Holt, Rinehart and Winston Inc.

Mehrens, W. A. & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. (4th ed.) New York: Holt, Rinehart and Winston Inc.

Nitko, A. J. (1996). *Educational assessment of students*. (3rd ed.). New Jersey: Prentice-Hall, Inc.

Nunnally, J. C. (1964). *Educational measurement and Evaluation.* New York: McGraw-Hill Inc.

Oescher, J. & Kirby P. C. (1998). *Assessing teacher-made tests in secondary mathematics and science classrooms*. (ED 322169). Available: http://eric.ed.gov/ERICWebPortal/home. on line.

Osuola, E. C. (1991). *Introduction to research methodology*. (2nd ed.). Onitsha, Nigeria: Africana F.E.P Publishers Ltd.

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

Pecku, N. K. (2000, April). *Formal assessment in the classroom: The Ghana Education Service termly assessment plan*. Paper presented to the Quality Improvement in Primary Schools (QUIPS) Project. Funded by the USAID.

Quaigrain, A. K. (1992). *Teacher competence in the use of essay tests: A study of secondary schools in the western region of Ghana*. Unpublished Master's Thesis, University of Cape Coast, Ghana.

Sarantakos, S. (1998). *Social Research*. Hound mills: McMillan Press Ltd.

Stiggins, R. (1999). *Assessment without victims*. Available: htpp://www.nsdc.org. on line.

Spiegel, M. R. (1994). *Theory and problems of statistics*. (2nd ed.). New York: McGraw-Hill Inc.

Tamakloe, E. K., Atta, E. T. & Amedahe, F. K. (1996). *Principles and methods of teaching.* Accra: Black Mask Ltd.

APPENDIX A

SENIOR SECONDARY SCHOOLS AND DISTRIBUTION OF TEACHERS SAMPLED

School	Frequency	Percent (%)
1. Aduman Secondary School	6	2.3
2. Adu Gyamfi Secondary School	11	4.2
3. Adventist Day Secondary School	17	6.4
4. Agogo State Secondary School	10	3.8
5. Agona Secondary Technical School	5	1.9
6. Amaniampong Secondary School	15	5.7
7. Beposo Secondary School	5	1.9
8. Effiduase Secondary Commercial School	7	2.6
9. Ejisu Secondary Technical School	7	2.6
10. Ejisuman Secondary School	5	1.9
11. Konongo Odumasi Secondary School	20	7.5
12. Kofi Adjei Secondary School	12	4.5
13. Kumasi Girls Secondary School	17	6.4
14. Obuasi Secondary Technical School	10	3.8
15. Osei Kyeretwie Secondary School	20	7.5
16. Prempeh College	9	3.4
17. St Louis Secondary School	10	3.8
18. St Monica's Secondary School	10	3.8
19. S. D. A Secondary School, Bekwai	10	3.8
20. S. D. A Secondary School, Agona	8	3.0

21. Simms Secondary Commercial School	11	4.2
22. T. I. Ahmadiyya Secondary School, Asokore	9	3.4
23. Technology Secondary School	5	1.9
24. Wesley High School, Bekwai	4	1.5
25. Wesley Girls' High School, Kumasi	16	6.0
26. Yaa Asantewaa Girls' Secondary School	6	2.3
Total	265	100.0

APPENDIX B

UNIVERSITY OF CAPE COAST
TEACHERS' QUESTIONNAIRE

The purpose of this questionnaire is to collect information on how teachers construct, administer, score and interpret the results of their classroom achievement tests. This will help determine whether the testing practices of teachers in Senior Secondary Schools are following the established principles of testing which can promote learning.

It would therefore be appreciated if you could provide frank answers to the questionnaire items. You are assured of complete confidentiality and anonymity of every information provided.

SECTION A
BACKGROUND INFORMATION

DIRECTIONS: Please tick (√) the box that best describes your response(s) where applicable or write in the space provided.

1. Gender: Female [ ] Male [ ]

2. Highest educational qualification.
   [ ] Teacher's Diploma
   [ ] HND
   [ ] University Degree (B.A., B.Sc.)B.Ed., M.Ed., M.A, M.Sc.)
   [ ] University Degree (B.Ed)
   [ ] Master's Degree (M.Ed., M.A., M.Sc., M.Phil.)
   Other (Please specify)……

3. How many years (approximate) have you taught in the Senior Secondary School?
   [ ] Less than five (5) years.
   [ ] Five (5) years or more.

4. Complete the following table to indicate the Form(s) and Subject(s) you teach.

Form (s)	Subject(s)
I	
II	
III	

5. Have you ever taken a course in testing (i.e., educational measurement)?
   [ ] No
   [ ] Yes

SECTION B

CONSTRUCTION OF CLASSROOM ACHIEVEMENT TESTS

DIRECTIONS: Please tick (√) the cell that indicates closely how frequently you practice the following **test construction principles.**

In the construction of classroom achievement test:				
	Always	Very Often	Sometimes	Never
6. I define the purpose of the test.				
7. I relate the instructional objectives of the subject matter to the test.				
8. I select the test format suitable for testing stated objectives.				
9. In determining the items on the test, I use a table of specifications or test blueprint.				
10. I prepare more items than needed in the test or examination				
11. I write the test items in advance (at least two weeks) of the test date to permit reviews and editing.				
12. I prepare the marking scheme as soon as the test items are written.				
13. I review the test items after they have been set aside for a few days by reading over the items.				
14. I write clear and concise directions for the entire test and sections of the test.				
15. I evaluate the test as a whole to find out whether: the test items are simple and clear; the test is a representative sampling of the material taught; the students will have enough time to complete the test; the test is the best instrument to assess the desired knowledge; the students have been prepared adequately for the test, etc, before submitting it for typing.				

DIRECTIONS: Please tick (√) the box that best describes your response(s), where applicable or write your answer in the space provided.

16. Do you usually do distracter analysis of your multiple-choice test items?
    [ ] No
    [ ] Yes
    [ ] Not familiar with the term distracter analysis.

17. Do you usually estimate the difficulty level of your objective test items?
    [ ] No
    [ ] Yes
    [ ] Not familiar with the term item difficulty.

18.Have you ever computed the discrimination index of your objective type tests?
    [ ] No
    [ ] Yes
    [ ] Not familiar with the term discrimination index.

19. Which of the following do you consider in the arrangement of your objective
    test items?
    [ ] The type of item used.
    [ ] The difficulty level of the items (arranged in order of increasing difficulty).
    [ ] The learning outcomes being measured.
    [ ] The subject matter being measured.
    [ ] None of these.

20. How do you know that the time you allot to your essay tests is adequate?
    [ ] By using the number of items on the test to estimate the time.
    [ ] By using the time that you (the teacher) can take to complete the test.
    [ ] By using the time that about 90% of the students can take to complete the test.
    [ ] None of these.
    Other (Please specify)……

SECTION C

ADMINISTRATION OF CLASSROOM ACHIEVEMENT TESTS

DIRECTIONS: Please tick (√) the cell that indicates how frequently you engage your students in each of the following steps when **preparing** the students in **advance** for the **test**.

In preparing students in advance for the test, I make them aware of :				
	Always	Very Often	Sometimes	Never
21. When (date & time) the test will be given.				
22. The conditions (number of items, place of test) under which the test will be given.				
23. The content areas (study questions, list of topics or learning targets) that the test will cover.				
24. The test formats (objective type or essay-type tests).				
25. The emphasis or weighting of content areas (i.e. value in points or content areas with higher marks)				
26. How the test will be scored and graded.				
27. The rules and regulations governing the conduct of the test				
28. The importance of the results of the test.				

DIRECTIONS: Please tick (√) the cell that closely indicates how frequently you practice the following **test administration principles**.

	Always	Very Often	Sometimes	Never
29. I do not give tests immediately before or just after a long vacation, or other important events.				
30. I ensure that the sitting arrangement allows enough space so that students will not copy from each other.				
31. I ensure adequate ventilation and lighting in the testing room.				

	Always	Very Often	Sometimes	Never
32. I use "Do Not Disturb. Examinations In Progress" sign when students are taking tests and examinations.				
33. During examinations, I expect and cater for all possible emergencies.				
34. I announce the remaining time (time left to complete test) at regular intervals.				
35. During examinations, I stand where I can view all students and move among the students once a while to check malpractices.				
36. I keep all remarks in the testing room to a minimum and make sure they are related to the test.				
37. I ask students to work faster during the time of testing in order to finish on time.				
38. I tell students the dire consequences of failure in the test they are taking.				

SECTION D

SCORING OF TEST AND INTERPRETATION OF THE TEST RESULTS

DIRECTIONS: Please tick (√) the box that best describes your response(s) where applicable or write your answer in the space provided.

39. Which method do you use in scoring your essay tests?
    [ ] Reading the whole essay through and based on your impression about the quality of the essay you award the marks (Holistic Method).
    [ ] Using a marking scheme in which the ideas and points clearly stated are awarded the marks (Analytic Method).
    Other (Please specify)……
40. How do you score your essay tests?
    [ ] Scoring all the items answered by each student before proceeding to the next student (i.e., student by student).
    [ ] Scoring one item for all students before proceeding to the next item (i.e.item by item).
    Other (Please specify)……

41. In scoring essay tests, do you constantly follow the marking scheme as you score?
    [ ] No
    [ ] Yes

42. In scoring essay tests, do you randomly reshuffle the answer scripts after scoring each item?
    [ ] No
    [ ] Yes

43. In scoring essay tests, do you score all responses to a particular question at a sitting without interruption?
    [ ] No
    [ ] Yes

44. Do you score essay tests only when you are physically sound and mentally alert?
    [ ] No
    [ ] Yes

45. Do you keep previously scored items out of sight when scoring the rest of the items?
    [ ] No
    [ ] Yes

46. In scoring essay tests, do you provide comments on the answer scripts for students to aid learning?
    [ ] No
    [ ] Yes

47. Do you score the answer scripts of essay tests with the names of the students known to you?
    [ ] No
    [ ] Yes

48. In scoring essay tests, do you score the mechanics of expressions such as penmanship, general neatness, spelling etc, separately from subject matter correctness?
    [ ] No
    [ ] Yes

49. Which method(s) do you use in interpreting the results of your tests? (You may indicate more than one)
    [ ] Describing a student's level of performance in relation with that of other members of the class (Norm-referenced interpretation). E.g. A student places 12th out of 50 students in a class.
    [ ]Describing a student's level of performance in terms of the learning tasks he can do. (Criterion-referenced interpretation). E.g. A student can solve 14 out of 20 problems in Algebra.
    Other (Please specify)……

50. In norm-referenced interpretation of test scores, i.e. describing a student's level of performance in relation with that of other members of the class, which method(s) do you employ? (You may indicate more than one).
    [ ] Simple ranking of raw scores
    [ ] Percentile ranks
    [ ] Stanine system of standard scores
    Other (Please specify)……

51. In criterion-referenced interpretation of test scores, i.e. describing a student's level of performance in terms of the learning tasks he can do, how do you set the performance standard (cut-off score)?
    [ ] By using a standard of 50 percent correct score.
    [ ] By setting an arbitrary standard of say, 85 percent correct, and then adjust it up or down as various conditions and experiences are considered.
    [ ] None of the above.
    Other (Please specify)……

52. How do you handle the low-achieving students (i.e. students who fail to master stated instructional objectives) in your class based on their test results?

……

……

……

THANK YOU VERY MUCH

APPENDIX C

OBSERVATION GUIDE FOR TEST ADMINISTRATION CONDITION

OBSERVER:……. SCHOOL:……

SUBJECT : … …….. DATE ……. …

A. PHYSICAL CONDITTIONS

CONDITION	JUDGEMENT OF OBSERVER (Tick)	REMARKS
1. Tables and chairs.	Suitable / Unsuitable	
2. Arrangement of tables and chairs.	Appropriate / Inappropriate	
3. Lighting	Good / Poor	
4. Ventilation	Good / Poor	
5. Quietness in the vicinity.	Appropriate / Inappropriate	
6. Provision of extra sheets and other writing materials.	Adequate / Inadequate	
7. Catering for emergencies	Adequate / Inadequate	

CONDITION	FREQUENCY OF OCCURRENCE (TALLY)	TOTAL OBSERVATION	REMARKS
8. Invigilator activities. E.g., reading, chatting, making phone calls, etc.			

B. PSYCHOLOGICAL CONDITIONS

CONDITION	FREQUENCY OF OCCURRENCE (TALLY)	TOTAL OBSERVATION	REMARKS
1. Asking students to hurry up and finish in time.			
2. Interruption to give instruction.			
3. Telling students the dire consequences of failure in the test they are taking			
4. Making announcements about time during test.			
5.Indication of cheating			
6. Walking around and looking over students' shoulders during testing			

7. Position of teacher / invigilator.
   [ ] Appropriate
   [ ] Inappropriate

8. Time of testing (when students would not be doing something pleasant. E.g., having lunch, breakfast, sports and games, etc).
   [ ] Appropriate
   [ ] Inappropriate

9. Starting of test and stopping of test.
   [ ] Starting promptly and stopping on time.
   [ ] Not starting promptly and not stopping on time.
   [ ] Starting promptly and not stopping on time.
   [ ] Not starting promptly and not stopping on time.

APPENDIX D

DATES ON WHICH QUESTIONNAIRES WERE ADMINISTERED IN THE SAMPLED SCHOOLS

Date	School
28th June, 2006	Aduman Secondary School
	S. D. A Secondary School, Agona
	Agona Secondary Technical School
29th June, 2006	Osei Kyeretwie Secondary School
	Yaa Asantewaa Girls Secondary School
30th June, 2006	S. D. A Secondary School, Bekwai
	Wesley High School, Bekwai
3rd July, 2006	St Monica's Secondary School
	Amaniampong Secondary School
	Adu – Gyamfi Secondary School
	Simms Secondary Commercial School
4th July, 2006	Adventist Day Secondary School
	Agogo State Secondary School
5th July, 2006	Konongo Odumasi Secondary School
	Ejisuman Secondary School
6th July, 2006	Kumasi Wesley Girls High School
	Kumasi Girls Secondary School
	Prempeh College
7th July, 2006	Effiduase Secondary Commercial School
	T. I Ahmadiyya Secondary School, Asokore
9th July, 2006	Technology Secondary School

	Obuasi Secondary Technical School
11th July, 2006	Ejisu Secondary Technical School
	Beposo Secondary School
12th July, 2006	Kofi Adjei Secondary Technical School
	St Louis Secondary School

APPENDIX E

DATES ON WHICH QUESTIONNAIRES WERE COLECTED BACK FROM THE SCHOOLS

Dates	School
13th July, 2006	Aduman Secondary School
	S. D. A Secondary School, Agona
	Agona Secondary Technical School
14th July, 2006	Yaa Asantewaa Girls Secondary School
	Osei Kyeretwie Secondary School
	S. D. A Secondary School, Bekwai
	Wesley High School, Bekwai
17th July, 2006	St Monica's Secondary School
	Amaniampong Secondary School
	Adu–Gyamfi Secondary School
	Simms Secondary Commercial School
19th July, 2006	Adventist Day Secondary School
	Agogo State Secondary School
20th July, 2006	Konongo Odumasi Secondary School
	Ejisuman Secondary School
21st July, 2006	Kumasi Wesley Girls High School
	Kumasi Girls Secondary School
	Prempeh College
24th July, 2006	Effiduase Secondary Commercial School
	T. I Ahmadiyya Secondary School, Asokore
25th July, 2006	Technology Secondary School

	Obuasi Secondary Technical School
27th July, 2006	Ejisu Secondary Technical School
	Beposo Secondary School
	St Louis Secondary School
31st July, 2006	Kofi Adjei Secondary Technical School

APPENDIX F

OBSERVATION IN SELLECTED SCHOOLS WITH THEIR DATES

School	Date(s) of Observation	Number of Observations
1.Kumasi Girls Secondary School	14th July, 2006	2
2. Wesley High School, Ashanti Bekwai	17th July, 2006	2
3. Obuasi Secondary Technical School	17th and 24th July,2006	2
4. Simms Secondary Commercial School	18th and 19th July, 2006	2
5. Kofi Adjei Secondary Technical School	20th and 21st July, 2006	2
6. Prempeh College	21st and 24th July, 2006	2
7. Osei Kyeretwie Secondary School	21st and 24th July, 2006	2
8. Kumasi Wesley Girls' High School	24th July, 2006	2
9. Adventist Day Secondary School	25th July, 2006	2
10. Aduman Secondary School	26th July, 2006	2
Total		20

APPENDIX G

TOWN IN WHICH SCHOOL IS LOCATED

Town	Number of Teachers Sampled	Percent (%)
1. Aduman	6	2.3
2. Agona Ashanti	13	4.9
3. Agogo	10	3.8
4. Ashanti Bekwai	14	5.3
5. Asokore	9	3.4
6. Bampenase	12	4.5
7. Beposo	5	1.9
8. Ejisu	12	4.5
9. Effiduase	7	2.6
10. Fawade, Kumasi	11	4.2
11. Jamasi	11	4.2
12. Kumasi	100	37.7
13. Konongo Odumasi	20	7.5
14. Mampong	25	9.4
15. Obuasi	10	3.8
Total	265	100.0

APPENDIX H

Principles used in the Construction of Classroom Achievement Tests

Items	Response in Percent (%)				
	Always	Very Often	Sometimes	Never	Total
1. Define the purpose of the test.	41.9	32.8	23.8	1.5	100
2. Relate the instructional objectives of the subject matter to the test.	50.0	38.4	10.8	0.8	100
3. Select the test format suitable for testing the stated objectives.	46.0	37.0	14.7	2.3	100
4. Use table of specifications to determine the items on the test.	9.1	23.8	47.5	19.6	100
5. Prepare more items than needed in the test or examination.	9.8	23.4	42.6	24.2	100

6. Write the test items in advance (at least two weeks) of the test date to permit reviews and editing	30.6	28.3	32.5	8.3	100
7. Prepare the marking scheme as soon as the test items are written.	57.0	25.3	41.0	3.8	100
8. Review the test items after they have been set aside for a days by reading over the items.	34.0	26.0	36.2	3.4	100
9. Write clear and concise directions for the entire test and sections of the test.	75.5	17.7	4.9	1.9	100
10. Evaluate the test as a whole on the criteria of clarity, validity practicality, efficiency and fairness.	9.8	13.4	40.6	36.2	100

APPENDIX I

Principles used in the Administration of Classroom Achievement

Item	Response in Percent (%)				
	Always	Very Often	Sometimes	Never	Total
1. Make students aware of when (date & time) the test will be given.	67.2	23.4	9.0	0.4	100
2. Make students aware of the conditions (number of items, place of test) under which the test will be given.	38.5	29.8	27.2	4.6	100
3. Make students aware of the content areas the test will cover.	41.5	29.4	24.5	4.5	100
4. Make students aware of the test format (objective-type or essay-type tests).	54.0	29.4	12.8	3.8	100
5. Make students aware of the emphasis or weighting of content areas (i.e. content areas with higher marks).	21.1	23.4	40.0	15.5	100

6. Make students aware of how the test will be scored and graded.	35.1	21.9	34.7	8.3	100
7. Make students aware of the rules and regulations governing the conduct of the test.	56.2	26.0	14.3	3.4	100
8. Make students aware of the importance of the test results.	49.4	32.8	15.5	2.3	100
9. Do not give tests immediately before or just after a long vacation or other important events.	15.8	22.6	42.3	19.3	100
10. Ensure that the sitting arrangement allows enough space so that students do not copy from each other.	56.6	29.1	13.6	0.8	100
11. Ensure adequate ventilation and lighting in the testing room.	62.6	21.5	12.5	4.4	100
12. Use “Do Not Disturb. Examinations in Progress” sign during examinations.	9.1	8.7	31.3	51.8	100

13. Expect and cater for all possible emergencies during examinations.	26.0	23.2	36.2	14.6	100
14. Announce the remaining time (time left to complete test) at regular intervals.	75.1	16.6	7.9	0.4	100
15. Stand where all students can be viewed and move among the students once a while to check malpractices during invigilation.	82.6	13.6	3.0	1.2	100
16. Keep all remarks in the testing room to a minimum and make sure they are related to the test.	35.8	35.5	23.0	8.2	100
17. Ask students to work faster during the time of testing in order to finish on time.	24.2	26.4	31.4	18.0	100
18. Tell students the dire consequences of failure in the test they are taking.	17.0	31.3	31.5	20.1	100