UNIVERSITY OF CAPE COAST

FAIR TESTING PRACTICES IN DISTRICT-MANDATED TESTING PROGRAMME IN THE ASHANTI REGION OF GHANA

EMMANUEL MENSAH BOAKYE

2016

UNIVERSITY OF CAPE COAST

FAIR TESTING PRACTICES IN DISTRICT-MANDATED TESTING PROGRAMME IN THE ASHANTI REGION OF GHANA

BY

EMMANUEL MENSAH BOAKYE

Thesis submitted to the Department of Education and Psychology, Faculty of Educational Foundations of the College of Education Studies, University of Cape Coast, in partial fulfillment of the requirements for the award of Master of Philosophy degree in Measurement and Evaluation

JULY 2016

DECLARATION

Candidate's Declaration

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

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

Name: Emmanuel Mensah Boakye

Supervisors' Declaration

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

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

Name: Prof. Young Kafui Abel Etsey

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

Name: Dr. Kenneth Asamoah-Gyimah

ABSTRACT

The study's main purpose was to investigate the fair testing practices of a test developer and teachers in a district-mandated testing programme in the Ashanti Region of Ghana. The research design adopted for this study was the multilevel mixed methods triangulation design. Critical case sampling technique was utilized in selecting 3 key informants and 9 test questions. Also, the two-stage cluster sampling technique was adopted in selecting 251 teachers from 162 public JHSs. The main instruments used for this study included interview protocol and document review for the test developer's strand and a 72-item questionnaire for the teachers' strand. The reliability coefficient of the questionnaire was 0.82. The data was analysed using qualitative content analysis, frequency and percentages, one-sample t-test, and independent-samples t-test. The study's results showed that the practices of the test developer, to a large extent, were fair to all test takers. The results of the study further showed that in terms of test preparation, administration, interpretation, reporting and uses made of test results, the practices of teachers, to a large extent, were fair to all test takers. However, in terms of grading students' test performance, the study's results showed that student' grades were influenced by factors such as students' comportment, which are irrelevant to the constructs as measured by the district-mandated test and thus, are unfair. It was recommended that the test developer develops a comprehensive test manual and share with all participating schools. The test developer should also engage the services of experienced test administrators as chief proctors who could serve as test monitoring teams.

KEY WORDS

Ashanti Region

District-mandated test

Fair testing practices

Multilevel mixed methods triangulation design

Teachers

Test developer

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; Prof. Y. K. A. Etsey of the Department of Educational Foundations, UCC, who offered an immeasurable guidance and support throughout the research and the entire MPhil programme. It was in indeed a great pleasure working with such an astute statistician and researcher.

I am also indebted to Dr. K. Asamoah-Gyimah, my co-supervisor, for his expert advice and suggestions during this research work and the entire MPhil programme. Through his intellectual and personal interactions, a lot have been learnt.

I also express my sincere appreciation to all the lecturers of the Department of Educational Foundations, UCC, especially Prof. F. K. Amedahe for the diverse ways they contributed to the successful completion of the MPhil programme. I will also like to register my sincere thanks to Prof. I. R. Amuah and Mr. E. Oko, all of CePME.

My appreciation also extends to all my course mates, especially, Aaron Adusei, Abraham Yeboah, Abraham Gyamfi and Mrs. Stella E. Komasi, for their love and encouragement in diverse ways. My final appreciation goes to all participants selected for the study.

DEDICATION

To Naomi Tiwaa, Isaac Nyamekye Sarpong, Lucy Achiaa, Belinda Kuffuor,

Michael Duodu, Michael Mpianim, and Florence Ofosua Kwarteng

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AERA: American Educational Research Association

APA: American Psychological Association

CePME: Center for Performance Monitoring and Evaluation

CIV: Construct Irrelevant Variance

CU: Construct Underrepresentation

DIF: Differential Item Functioning

ETS: Educational Testing Service

GES: Ghana Education Service

JHSs: Junior High Schools

NCME: National Council on Measurement in Education

OTL: Opportunity to Learn

UDA: Universally Designed Assessments

CHAPTER ONE

INTRODUCTION

The ideal of test fairness has been pursued since the Imperial Examinations in the 15th century, during which rigorous measures such as double marking, obscuring test takers' names, transcribing answer sheets, and proofreading were used in ensuring that test results are not influenced by factors that are irrelevant to constructs purported to be measured (Cheng, 2010). However, students of the 20th and 21st centuries are more diverse in their characteristics than ever before, and many of them have special needs (Redfield, 2001). Therefore, inquiries into test fairness helps direct efforts to reduce bias against certain test takers or groups of test takers, and create equal opportunities for all test takers to demonstrate their knowledge and skills, and promote social justice (Xiaomei, 2014).

Background to the Study

Nature of classroom testing

Tracing the concept through the Chinese Imperial Examination System in the 15th century, the notion of testing seems to have proved to be one of the indispensable tools in the educational enterprise (Khalanyane & Hala-hala, 2014). According to Nitko (2001, p. 5), a test is defined as an "instrument or systematic procedure for observing and describing one or more characteristics of a student using either numerical scale or classification scheme." Testing in Ghana's

educational institutions is designed to assess either curriculum-based (classroom instructional) achievement or a variety of student traits other than curriculum-based achievement.

Tests that are curriculum-based assess the goals or objectives of the curriculum that a student is mastering. In the Ghanaian context, curriculum-based assessment (CBA) used to be mainly teacher-made and state-mandated tests (Anhwere, 2009). "State assessments are usually based on a state's curriculum framework or standards" (Nitko, 2001, p. 381). In Ghana, national mandated tests include Basic Education Certificate Examination (BECE) and West African Secondary Schools Certificate Examinations (WASSCE). Classroom teacher-made tests, on the other hand, are tests constructed by the classroom teacher for the purposes of making classroom decisions based on information obtained on students in a particular class or school. According to Asamoah-Gyimah (2002), "classroom or teacher-made tests are frequently used as a major evaluating device of students' progress in schools" (p. 2).

In recent years, however, there have been an increasing concern about the nature of the most widely used form of student assessment (i.e., classroom teacher-made tests) and uses that are made of its' results.

> 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. (Oduro-Okyireh, 2008, p. 4).

Such concerns are probably as a result of the Ghanaian teacher's expertise in, and perception about assessment practices. It has been reported that some Ghanaian teachers perceive the management of assessment practices as extra load to their teaching activities (Anhwere, 2009). In reaction to such concerns, some educators around the world have advocated for the use of published achievement tests.

Published achievement tests

Published achievement tests are constructed and managed by individuals and institutions that are believed to be experts in educational measurement in a specific subject area. Therefore, these tests are assumed to be of much higher quality than teacher-made tests. Published achievement tests could be standardized or non-standardized tests (Nitko, 2001). One form of published achievement tests are those tests developed by contractual test developers and these kinds of tests form the focus of this study.

Private testing companies are contracted by District Directors of Education for the provision of assessment materials for public schools within their jurisdiction, and therefore, these tests can best be described as district-mandated tests (Nitko, 2001). In Ghana, district-mandated tests mostly take the form of terminal examinations, promotion examinations and mock examinations, and serve as summative assessments. The diverse nature and large number of test-takers in large-scale assessments such as district-mandated tests, and high-stakes decisions based on summative assessment results make the concept of test

fairness an important criterion for improving the validity of scores from district-mandated assessments (Nitko).

Nature, and need for test fairness in large-scale assessments

Fairness implies that every test taker has the opportunity to prepare for the test and is informed about the general nature and content of the test, as appropriate to the purpose of the test (Joint Committee on Testing Practices [JCTP], 2004). Rodabaugh (1991) made an assertion that three factors help a test appear fair to students:

1. All the materials on the test are relevant to the courses' objectives and were covered in lectures, readings or both.
2. The test is appropriate in difficulty for the course.
3. The test is well designed, with clearly phrased questions and unambiguous multiple-choice response options.

Felder (2002) asserted that an examination, according to students, is deemed unfair in the following scenarios: (1) Problems on content not covered in lectures or homework assignments; (2) Problems the students consider tricky, with unfamiliar twists that must be worked out on the spur of the moment; (3) Excessive length, so that only the best students can finish in the allotted time; (4) Excessively harsh grading, with little distinction being made between major conceptual errors and minor calculation mistakes; and (5) Inconsistent grading, so that two students who make identical mistakes lose different points.

Lack of fairness in educational testing can result in serious consequences. What students hate more than anything else are examinations that they perceived

as unfair (Felder, 2002). Rodabaugh (1991) stated that if deprived of the grades they think they deserve; students might be tempted to cheat. Also, lack of fairness in testing practices may trigger undesirable behaviours from students, in and outside the classroom. Wankat and Oreovicz (1993) stated that unfair and poorly graded examinations cause student resentment.

To better assure high degree of professionalism in assessment practices, many associations have prepared guidelines to assist testing professionals in maintaining a professional level of quality in classroom assessments. For example, a working group of the Joint Committee on Testing Practices (JCTP) prepared the Code of Fair Testing Practices in Education, which provides parallel statements concerning the roles of test developers and test users in order to achieve test fairness (JCTP, 2004).

The Government of Ghana, since September 1987, has embarked upon a new educational programme geared strategically at making education more accessible to all children of school-going age, improving equity and the quality of education as a whole, and making education more relevant to the socio-economic needs of the country (Ghana Ministry of Education [GMOE], 2002). In this time of educational reform, measurement experts have asserted that to have accurate and fair measures of progress, all students must be included in accountability systems such as mandated tests. Beyond simply including all students in assessments, there is a need to have their test performance to be a valid and fair measure of their knowledge and skills (Thurlow, Quenemoen, Thompson, & Lehr, 2001). Therefore, ongoing research is essential to address many unanswered

questions about the fairness of mandated assessments in Ghana. This study, thus, in using multiple sources of relevant information, investigates the roles of a test developer and teachers in providing good quality tests that are fair to all test takers in a district-mandated testing programme in the Ashanti Region of Ghana.

Statement of the Problem

Globally, investigations of test fairness of large-scale assessments have primarily involved teat users such as teachers and students, and statistical approaches that rely on actual scores of test takers. Such procedures for investigating test fairness exclude the test developer, who is considered an important stakeholder in the actual testing process. Moreover, many researchers (Amedahe, 1989; Anamuah-Mensah & Quagrain, 1998; Anhwere, 2009; Quagrain, 1992; Oduro-Okyireh, 2008) have conducted studies on the practice of testing in educational institutions in Ghana. However, these studies are delimited to classroom teacher-made tests, and therefore, the findings of such studies, although very useful, cannot be generalized to the practice of testing in large-scale assessments of basic schools at the district level. This situation does not give a holistic picture of assessment practices in Ghanaian schools, and thus has created a research gab that needs to be filled.

In recent times, the use of achievement tests constructed by private testing organisations, in large-scale district-mandated testing programmes, is in ascendency to the extent that these tests are relied upon to assess students even at the kindergarten level. It could be estimated that over sixty (60) Districts of Education and hundreds of private schools in Ghana are involved in assessment

materials, prepared by private testing organisations, for a particular academic term. Due to its' summative nature, it could also be said that high-stakes decisions in the classroom are based on district-mandated test results. However, test users make inferences about the knowledge and skills of students when decisions about the students are made on the basis of the test scores. The extent to which those inferences are appropriate for different groups of test takers is an important aspect of test fairness (Educational Testing Services [ETS], 2009).

In Ghana, increasing incidents of cheating behaviours in mandated and high-stakes examinations, and the GES's move towards a more inclusive system of education as stated in the Education Strategic Plan 2003-2015 (GMOE, 1999), has made an investigation into test fairness even more relevant. For instance, as a result of the strategic plan, the diversity of students' characteristics at the basic school level in Ghana is expected to broaden significantly. In order to have assessments that are truly valid for such a wide range of learners, they should be fair and accessible to all students assessed (Darling-Hammond et al., 2013). However, limited research on the concept makes it difficult to make claims as to the extent to which test fairness has evolved in assessment practices of Ghana's basic schools (Xiaomei, 2014).

This study, therefore, sought to investigate the separate roles played by a test developer and teachers in achieving fairness in district-mandated testing programme in the Ashanti Region. Stated in question form, the main research problems are (1) how are fair testing practices ensured by test developers in the Ashanti Region of Ghana, (2) what are the fair testing practices of teachers in the

Ashanti Region of Ghana, and (3) what significant differences exist in teachers' fair testing practices in terms of their levels of training in educational measurement?

Purpose of the Study

This study sought to address the problem of test fairness in district-mandated testing programme of public JHSs in the Ashanti Region of Ghana. The study's main purpose was to investigate the fair testing practices of a test developer and teachers in the Ashanti Region in terms of the standard approved practices for developing tests, test preparation, administering tests, grading and interpreting students' test performance, reporting test results, and uses made of students' test results. Specifically, this study sought to assess:

1. A test developer's standard approved practices of ensuring test fairness.
2. Teachers' fair testing practices in preparing students for district-mandated test, and the differences in test preparation practices that exist among them in terms of their levels of training.
3. Teachers' fair testing practices in administering district-mandated tests, and the differences in test administration practices that exist among them in terms of their levels of training.
4. Teachers' fair testing practices in grading students' test performance, and the differences in grading practices that exist among them in terms of their levels of training.
5. Fair testing practices of teachers in interpreting students' test performance.
6. Fair testing practices of teachers in reporting district-mandated test results.

7. Teachers' fair uses of district-mandated test results.

Research Questions

The above listed objectives of the study were investigated by addressing the following research questions:

1. In what ways does a test developer follow the standard approved practices of test fairness?
2. In what ways do teachers adhere to fair test preparation practices?
3. What significant difference in fair test preparation practices exists between teachers who have received training in educational measurement, and those who have received no training in educational measurement?
4. In what ways do teachers adhere to fair test administration practices?
5. What significant difference in fair test administration practices exists between teachers who have received training in educational measurement, and those who have received no training in educational measurement?
6. What factors influence students' grades on the district-mandated test?
7. What significant difference exists between teachers who have received training in educational measurement, and those who have received no training in educational measurement in terms of factors influencing students' grades?
8. What factors are considered by teachers in interpreting students' performance on the district-mandated test?
9. What are teachers' fair reporting practices of district-mandated test results?
10. What are teachers' fair uses of district-mandated test results?

Significance of the Study

The rate at which district-mandated tests are relied upon to assess students' achievement of some specific contents of the curriculum, and the high-stakes nature of the uses that are made of the test results, necessitated a research into the validity of such uses of district-mandated test results. The findings would inform the Centre for Performance Monitoring and Evaluation (CePME) and the Ghana Education Service (GES) about the lapses associated with the district-mandated testing programme and thereby, help plan appropriately towards it.

The GES, in collaboration with the Ministry of Education and other test stakeholders, could develop testing policies such as a Code of Fair Testing Practices in Education in Ghana. Such a policy document would guide the activities of test developers and test users in order to achieve test fairness, and thereby, better meeting the needs of the country's increasingly diverse student population. The findings of this study would serve as an important source of reference for such an important policy document. Moreover, CePME, in collaboration with the district directorate of education could organize training sessions on fair testing practices for teachers in the participating districts.

The findings of this study would help create a holistic picture of classroom achievement testing practices in Ghanaian schools. Also, findings of this study would serve students, educationists, and experts in measurement as an important reference source for further studies. For instance, recommendations made in this study would be a good source of research problems for further studies on the concept of test fairness.

Delimitations

The study was confined to public JHSs in the Ashanti Region of Ghana who are involved in district-mandated test materials, prepared by CePME. The study was delimited to CePME based on a formal case study screening procedure I conducted. Yin (2011) stated that screening criteria include the willingness of key persons in the case to participate in your study, the likely richness of the available data, and preliminary evidence that the case has had the experience that you are seeking to study. CePME is a Ghanaian-based educational consultancy, which is duly registered under the companies' code, 1963, and came into existence in August, 2004. CePME has been approved and recommended by the GES, and its core activities include (1) providing in service training for teachers in participating districts and private schools, and (2) providing standardized assessment materials for evaluation of pupils' performance at the end of every term.

Assessment materials prepared by CePME are administered in both public and private schools in the Ashanti Region. However, due to feasibility constraints, the study was confined to only public schools in the region. The choice of JHSs was also a delimitation. District-mandated test, prepared by CePME, are mostly administered at the basic school level, which includes Primary Schools and JHSs. Moreover, the study was confined to test users such as teachers at the public JHSs. Other test users such as parents and students were excluded.

Critical views on test fairness indicate two broad conceptual perspectives: (1) Views that focus on testing process; and (2) Views that focus on socio-cultural

context (Xiaomei, 2014). However, this study was delimited to general views that focus on testing process such as absence of bias, equitable treatment, and opportunity to learn standards.

Limitations

The target population for this study was 8,424 teachers in the 162 public JHSs in the Ashanti Region of Ghana. Therefore, a sample size of 251 teachers, representing 2.97% of the target population was relatively small.

Data on roles played by the test developer in achieving test fairness was collected and analysed qualitatively, and thus, do not meet the conditions for statistical generalization. Moreover, I cannot judge the honesty and truthfulness of such responses made by respondents in an interview or on a questionnaire.

Lastly, a major limitation of this study was very limited local sources of literature on test fairness. I, therefore, depended on other foreign works that have been conducted on test fairness.

Definition of Terms

Absence of bias: This implies that the content of a test does not discriminate against any student or groups of students.

Construct-irrelevant variance: Refers to differences in the test performance of students, caused by factors that are irrelevant to the purpose of an assessment.

Equitable treatment: This implies that students are assessed using appropriate methods and procedures, which may vary from one student to the next.

External tests: Assessment instruments that are developed and/or graded by people who are not associated with the schools providing the students' learning.

Large-scale assessments: Testing programmes that test relatively large numbers of students, such as those in a country, state or district.

Mandated tests: Tests that are administered because they are required by school district policy, and are compulsory for all students in that district.

Opportunity to learn: The extent to which students have had exposure to instruction or knowledge that affords them the opportunity to learn the content targeted by a test.

Test developers: These are people and organizations that construct tests, as well as those that set policies for testing programmes.

Test fairness: It is the extent to which students are given equal opportunity to demonstrate their knowledge and skills on a test.

Organisation of the Study

The study was organized into five chapters. The first chapter discussed the Introduction, which highlighted the study's background, statement of the problem, purpose, research questions, significance, delimitations, limitations, and definitions of terms. Chapter Two reviews the literature related to this study. A conceptual framework and an empirical model were adopted to review literature. Chapter Three discusses the Research Methods in terms of the research design, study area, population, sampling procedure, data collection instruments, data collection procedure, and data processing and analysis techniques. In Chapter Four, the results are presented and discussed, while Chapter Five, which is the final chapter, summarizes the main findings, and provide conclusions, recommendations, and suggestions for further research.

CHAPTER TWO

LITERATURE REVIEW

Introduction

I sought to investigate the fair testing practices of a test developer and teachers in a district-mandated testing programme in the Ashanti Region of Ghana. Thus, this chapter reviewed literature in the following areas relating to test fairness in large-scale assessments:

1. Conceptual Review:
    a. The Concept of Achievement Testing.
    b. The Concept of Large-scale Assessments.
    c. The Concept of Test Fairness.
    d. Testing Standards, Guidelines and Codes of Practices.
    e. Fair Test Development Practices.
    f. Fair Test Preparation Practices.
    g. Fair Test Administration Practices.
    h. Fairness in Grading Students' Test Performance.
    i. Fairness in Interpreting Students' Test Performance.
    j. Fair Test Reporting Practices.
    k. Uses of Mandated-test Results.
2. Empirical review.

The Concept of Achievement Testing

Testing in educational institutions is designed to assess either curriculum-based (classroom instructional) achievement or a variety of student traits other than curriculum-based achievement. Tests such as career interest, attitudes, and personality tests assess a variety of student's traits other than curriculum-based achievement (Nitko, 2001). Stainback and Stainback (1996) argued that depending on how it is interpreted, assessing almost any student performance deriving or related to the classroom curriculum, including achievement testing could be an example of curriculum-based assessment (CBA). It must be emphasized that achievement testing is concerned with assessing students based on the domain of content areas they have studied, which are drawn from the school curriculum.

Etsey (2012) stated that achievement test "measures the extent of present knowledge and skills. In achievement testing, test takers are given the opportunity to demonstrate their acquired knowledge and skills in specific learning situations" (p. 41). An extensive review of the literature posits two main types of achievement tests. These are teacher-made tests and external tests (Nitko, 2001). Assessment made by teachers of students' attainment, knowledge and understanding is called variously as teacher-made tests. Teachers construct these tests to assess the amount of learning done by students (Amedahe, 1989).

External tests or "extra-classroom assessments" (Nitko, 2001, p. 43), on the other hand, include assessment instruments that are developed and/or graded by people who are not associated with the schools providing the students' learning

(Lissitz & Schafer, 2002). Commercial test publishers, departments of education, and local school jurisdictions, usually develop external test (Reeves, 2003). According to the National Association of School Psychologists (NASP, 2002), external tests are usually mandated by core components of standard based reform, which includes (1) content and performance standards set for all students, (2) development of tools to measure the progress of all students toward the standards, and (3) accountability systems that require continuous improvement of student achievement. External test can take the form of textbook accompaniments, survey tests and mandated tests (Munson & Parton, 2013; Nitko, 2001; Zucker, 2004).

Mandated tests are tests that are administered because they are required by school district policy. Mandated testing programmes are mandatory for students in a particular district or state, and are best described as state-mandated or district-mandated tests (Nitko, 2001). Mandated testing programmes are also referred to as large-scale assessments due to the large number of students in a state or district taking the test. Mandated testing programmes ensure all public school students, no matter where they go to school, receive quality education (Munson & Parton, 2013).

The Concept of Large-scale Assessments

Large-scale testing programmes are those that test relatively large number of students, such as those in a state or district. The advent of large-scale assessments in education evolved in response to a perceived need. In large part, large-scale assessment expanded to fill the assessment and accountability void left by teacher-made and internal assessments. Popham (2001) provided the following

response in describing the "primary measurement mission" of large-scale assessment programmes:

> It's all about accountability. Large-scale assessment programmes [*sic*], the bulk of which are of the high-stakes variety, are in place chiefly because someone believes that the annual collection of students' achievement scores will allow the public and educational policymakers (such as state school board members or state legislators) to see if educators are performing satisfactorily. (p. 34).

A second part of the need for large-scale assessments lies in the deep void that existed in teachers' training in and understanding of assessment design and use (DePascale, 2003; Popham, 2002; Webber, Aitken, Lupart, & Scott, 2009). The emphasis on accountability, combined with a lack of confidence in local educators' ability to assess students resulted in large-scale testing (1) becoming the primary vehicle to assess all students, and (2) serving as a model for internal assessment (DePascale).

Many large-scale assessment programmes in Ghana and West Africa as a whole take the form of summative assessments as these tests are usually tailored at the end of an academic term, year or an instructional programme. At the national level, large-scale assessment programmes in the West Africa sub-region serve certification purposes and are thus regarded as summative assessments. This includes the West Africa Secondary Schools Certificate Examination (WASSCE). Similarly, large-scale assessment programmes at the district level in Ghana is an accountability measure that is generally used as part of the grading process and

thus serves summative purposes (Scottish Qualifications Authority [SQA], 2014). Summative assessments provide critical information about students' learning, as well as an indication of the quality of classroom instruction, especially when they are accompanied by other sources of information (SQA).

Standard-based assessments and alignment

Large-scale assessment programmes vary in different ways and purpose. For instance, some large-scale assessments compare individual student performance to a national group while others compare individual student performance to established performance standards (Gichuru, 2014). The latter are known as standards-based systems of assessment, which include criterion-referenced tests. Linn (2008) stated that large-scale testing programmes have moved away from a reliance on norm-referenced tests, and have embraced standards-based assessments. In such systems, test items reflect a pre-established set of content standards that specify the knowledge and skills students are expected to acquire as a function of schooling. Therefore, a key component of standard-based assessments' validation is alignment of the test to both curriculum and instruction.

> The logic of criterion-referenced assessment is say what you want students to be able to do (see learning objectives), teach them to do it (through lectures, tutorials, and learning activities), and then see if they can do it. Thus, it is about alignment. (Briggs, 2009, p. 144).

Alignment refers to the degree to which the items on a test match the structure and intent of the curriculum and instruction (Bunch, 2012). Fairness of

the standard-based assessment process is indicated by careful alignment of standards, curriculum and instruction, assessment, and opportunity to learn (NASP, 2002). According to Webb (2006), La Marca, Redfield, Winter, Bailey, and Despriet (2000), and Ananda (2003), there are three traditional methodologies for systematically evaluating and documenting the alignment between standards and assessments, and they include (1) sequential development, in which the standards and assessments are developed in a serial manner, (2) expert review, which relies on the opinions of specialists, and is used to analyse the alignment between assessments and standards when both have already been developed, and (3) document analyses, which involves the analysis of standards and assessment documents, using a system for encoding their content and structure.

If the assessments in a system do not adequately represent the depth and breadth of the standards upon which it is based, then the system is not aligned (Redfield, 2001). Redfield further stated that assessments and content standards could be misaligned in a number of ways, including:

1. The tests may include only some items or tasks that can be directly aligned with the standards.
2. The tests may include items that fully align with the content standards, but in combination only cover a few of the parts covered by the content standards.
3. The situation wherein the content standards are much narrower than the knowledge and skills required answering the test items.

The first situation poses a problem whereby students would not have had adequate opportunity to learn the knowledge and skills assessed by the test while the

second situation renders the test results an inaccurate reflection of students' mastery of most of the domain covered in the content standards (Redfield).

Degree of standardisation

Airasian (1999) defined standardized assessment procedures as those that are intend to be administered, scored, and interpreted in the same way for all examinees, regardless of where or when they are assessed. "But standardization is an ideal, and so we speak of degree of standardization" (Nitko, 2001, p. 15). Standardization attempts to control for external factors to the greatest degree possible so that the assessment tasks are a valid measurement tool that produces meaningful results. "A major reason for standardizing an assessment procedure is to permit fair comparisons of different students' performance or of the same student on different occasions" (Nitko, p. 15).

Standardization of assessment systems includes test manual, which provide detailed directions to ensure consistent administration and scoring procedures. The provision of scoring keys and guides are important features of standardization because they help to reduce errors when student responses are hand-scored. According to Illinois State Board of Education (ISBD, 2014), mandated tests are an important and required tool used to monitor state, district, school, and student achievement. For tests to yield fair and equitable results, they must be given under standardized conditions. Only then will the results for students, schools, and districts are comparable across the state or district and from year to year.

The Concept of Test Fairness

An examination should be appropriate for all qualified examinees irrespective of ethnicity, religion, gender, socio-economic status, and age. Xi (2010) defined fairness as "comparable validity for all the identifiable and relevant groups across all stages of assessment, from assessment conceptualization to the use of assessment results" (p. 154). A number of researchers (Bouville, 2008; Davies, 2010) have raised arguments against the pursuit of test fairness in classroom testing. Bouville argued that, "if people do not agree on what fair means then the consensus that exams [*sic*] must be fair is illusory" (p. 1). Davies also noted that the pursuit of fairness is in vain "first because it is unattainable and second because it is unnecessary" (p. 171).

However, it should be noted that fairness, like validity, is a matter of degree (Cole & Zieky, 2001), and that the aim of fairness studies is to investigate the extent to which assessment practices provides equal opportunities for all test takers.

> There is no such thing as a fair test, nor could there be [*sic*]: the situation is too complex and the notion simplistic. However, by paying attention to what we know about factors in assessment and their administration and scoring we can begin to work towards tests that are more fair to all the groups likely to be taking them, and this is particularly important for assessment used for summative and accountability purposes. (Gipps, 1995, p. 83).

General views of test fairness

Test fairness is a multi-faceted issue. Fairness has been conceptualized in various ways, which result in different approaches of viewing fairness (Baharloo, 2013). An extensive review of the literature on test fairness reveals discussions on the concept around five interwoven and related concepts, including validity, absence of bias, equitable treatment, equality of testing outcomes, and opportunity to learn (OTL) standards. However, equality of testing outcomes has been, unanimously, rejected in the field of research in educational measurement. Students come to school with different experiences and they do not have identical experiences at school either. We cannot, therefore, expect assessments to have the same meaning for all students (Gipps, 1995). According to AERA, APA and NCME (2014), the focus of fairness discussions should be delineated to aspects of test, testing, and test use that relate to fairness, which are the responsibility of those who develop, use, and interpret the results of tests, and upon which there is general professional and technical agreement.

The relationship between the concepts of test fairness and validity of assessment results provides better insights into the conceptualization of test fairness and its practical investigation (Baharloo, 2013). According to Camilli (2006) and Stobart (2005), measurement experts generally accept that fairness is an important quality that is distinct from but related to validity. This view is clearly represented by the Code of Fair Testing Practices in Education [JCTP, 1998; 2004] (Baharloo). Thus, the concept of fairness in assessment is impossible to divorce from the concept of validity because the two share a mutuality of

meaning and import. “Fairness is essential for valid measurement, and validity is essential for fair measurement” (Educational Testing Service [ETS], 2012, p. 19).

Absence of bias refers to tests that are free of contents and contexts that cause differences in students’ performance on the test, but are extraneous to the purpose of the test. Thus, test bias is defined as a technical term reflecting either of two situations in which examinees from protected populations’ score differently because of deficiencies in the test itself or are offended by an assessment (Lang & Wilkerson, 2008). Sources of bias in assessment include factors related to context, instrument, scorers, and students.

> To ensure that the results of assessments adequately reflect what candidates know and can do, it is important to remove any contextual distractions and/or problems with the assessment instruments that introduce sources of bias. Contextual distractions include inappropriate noise, poor lighting, discomfort, and the lack of proper equipment. Problems with assessments include missing or vague instructions, poorly worded questions, and poorly reproduced copies that make reading difficult. (Lang & Wilkerson, p. 15).

The elimination of bias also means that an assessment is free of poorly conceived language and task situations that might interfere with candidate performance and unintentionally favour some candidates over others. Furthermore, the elimination of bias includes consistent scoring of an assessment and vigilant efforts not to discriminate against groups of test takers (Lang & Wilkerson).

Equitable treatment means that students are assessed using appropriate methods and procedures, which may vary from one student to the next (Gipps, 1995). Lam (1995) stated that a fair assessment is one in which students are given equitable opportunities to demonstrate what they know. Equitable treatment does not necessarily mean that all students should be treated exactly the same, but rather all students should be assessed using methods and procedures most appropriate to them (Suskie, 2000).

AERA et al. (2014) defined opportunity to learn (OTL) standards as "the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test" (p. 56). In connection with assessment, Winfield (as cited in CRDD, 2005) stated that opportunity to learn may be measured by the amount and depth of content covered with particular groups of students. OTL researchers typically have distinguished three overlapping categories of concern: content coverage, instructional strategies, and instructional resources (Bachman, 1990; Buren, Ziker, Brashear, & Crosswell, 2006). Instructional resources are defined as anything that is read, listened to, manipulated or experienced by students as part of the instructional process while instructional strategies are the method adopted in teaching to enhance students' learning of course content (Porter, 1993). However, instructional resources and instructional strategies are not considered as part of the achievement testing process, and therefore, cannot be considered in fairness discussions within the measurement field.

Content coverage, on the other hand, refers to the extent to which students have been exposed to the specific topics that are essential to attaining particular standards and/or that are directly assessed (QAA, 2012), and thus, content covered is a basic consideration in the development of a test specification, which is considered as part of the achievement testing process. "Assessments are fair when they assess what has been taught" (Lang & Wilkerson, 2008, p. 13). An achievement test that assumes a particular syllabus would not be a fair test for students who did not follow that curriculum (Willingham & Cole, 1997). It is also generally accepted that concerns about students' opportunity to learn do necessarily apply to situations where the same individual is responsible for the delivery of instruction and the testing and/or interpretation of test results (AERA et al., 2014).

Threats to test fairness

Construct-irrelevant variance (CIV)

Threats to test fairness are well documented in the literature, and are usually discussed around construct-irrelevant variance (AERA et al., 2014). CIV refers to factors that cause differences in students' scores, but are not attributable to the construct that the test is designed to measure. It occurs when test scores are influenced by factors irrelevant to the construct measured, such as an individual's background knowledge, personality characteristics, test-taking strategies, and general intellectual or cognitive ability (Bachman, 1990). According to Schouwstra (2000), construct-irrelevant variance represents systematic interference in the measurement data, often associated with the scores of some,

but not all, examinees. These factors, incorrectly and systematically increase or decrease test scores for some students (Haladyna & Downing, 2004).

Construct-irrelevant variance may be introduced by inappropriate sampling of test content, lack of clarity in test instructions, item complexities that are unrelated to the construct being measured, and/or test scoring criteria that may favour one group over another (AERA et al., 2014). Abedi, Leon, and Mirocha (as cited in Thompson, Johnstone, Anderson, & Miller, 2005) noted that features of large-scale assessments might underestimate the achievement of certain group of students. For example, some poorly designed item formats could make it more difficult for some students to give a correct answer (Bachman, 1995). Moreover, test takers get distracted when a test advocates positions counter to their strongly held beliefs, and therefore, may respond emotionally rather than logically to controversial materials (ETS, 2009). Assessments should work equally well for all students regardless of their construct-irrelevant characteristics.

Construct under-representation (CU)

In addition to CIV as a major threat to test validity is construct under-representation (CU), which refers to the under-sampling or biased sampling of the content domain by the assessment instrument (Schouwstra, 2000). According to Messick (1989), construct under-representation refers to the imperfectness of tests in accessing all features of the construct, hence leaving out some important features that should have been included. CU occurs when a test fails to capture important aspects of the construct that it is intended to measure so that CU leads to underperformance on the part of some test takers.

Teaching the test

When teachers can study a test and identify the content and specific objectives that each test item measures, there is a strong temptation to teach content that will directly affect test performance. This practice is known pejoratively as 'teaching to the test' (Haladyna & Downing, 2004). Redfield (2001) argued that teaching to the test increases the probability of students' success relative to any assessment based on the standards, not just the items on a particular form of a particular test, and therefore, it is useful to distinguish between 'teaching to the test' and 'teaching the test'.

Teaching the test implies teaching students the actual, or nearly identical, questions that will appear on an external test. This practice is also referred to as 'item teaching' (Popham, 2001). Item teaching constitutes cheating, and confines instruction to a mere sample of the knowledge and skill domain represented by the test (Redfield, 2001). Activities that matched this term include (1) going over the actual test or questions from the test with students, (2) using modified versions of test questions as practice in class, and (3) taking older tests and giving them as practice (Thompson et al., 2005). According to Lane (2014), practicing drilling items can increase students' scores, but unlikely to develop general understanding, which defeats the purpose of mandated testing. Cheng (1998) stated that teaching the test requests students to cram for the examination rather than prepare for a broad curriculum.

Stereotype threat

When a widely known poor intellectual ability exists about a group, it creates for its members a burden of suspicion that acts as a threat (Steele, 1997). Stereotype threat is the threat that members of a stigmatized group experience when they believe that they may, by virtue of their performance on a task, confirm a negative stereotype about themselves and members of their group (Kellow & Jones, 2008). Negative stereotypes about intellectual abilities can act as a threat that disrupts the performance of students targeted by bad reputations. It has been established that group members perform poorer on a particular task if they have been confronted with a negative stereotype towards their group with respect to attainment in certain activities.

Students experiencing threat consequently perform more poorly because they have fewer cognitive resources to devote to tasks than do their peers who are not experiencing threat (Alter, Aronson, Darley, Rodriguez & Ruble, 2009, p. 166). This is explained by an anxiety that one will confirm the stereotype, which puts additional pressures on the member of the targeted group (Wright & Taylor, 2003). All individuals have knowledge of various stereotypes, and it is likely that teachers themselves often unintentionally reinforce stereotypes. Fairness is threatened when assessment decisions are influenced by stereotypes, but these are often so entrenched that they are overlooked (Tierney, 2013).

Minimizing threats to test fairness

The goal of fairness in assessment can be approached by ensuring that test materials are as free as possible of unnecessary barriers to the success of a diverse

group of students. According to Messick (1989), test developers have a major obligation to minimize construct-irrelevant test variance. Careful editing of test content, adequate testing time, and the use of standardized testing procedures diminish construct-irrelevant test variance and yield more accurate test scores.

The threat of teaching the test can be minimized through the provision of practice test questions. Also, by employing different set of test questions for different examinations, teachers could be discouraged from practicing drilling items. The consensual knowledge of group ability may stem from communicative processes that play a central role in the acquisition of stereotype (Croizet, Désert, Dutrévis, & Leyens, 2001). Therefore, directions and/or instructions for tests should not make any reference to potential stereotype-relevant information (Kellow & Jones, 2008). Croizet et al. found that when instructions accompanying the test did not create stereotype threat, stigmatized group members' performance was equal to that of other participants.

Testing accommodations

Today's students are more diverse in their characteristics than ever before, and many of them have 'special needs' (Redfield, 2001). Students with special needs might need to be accommodated on district assessments in order for them to demonstrate their knowledge. Accommodations refer to changes in the administration of an assessment, which do not change the construct, intended as measured by the assessment. Accommodations must not provide advantage to students eligible to receive them, but rather be used for purposes of equity in assessment (Redfield).

Testing accommodations are grouped in the following four categories (Virginia Department of Education [VDE], 2015):

1. Time/scheduling accommodations address adjustments in the tests' schedule;
2. Setting accommodations address adjustments to the physical environment where the test would normally be administered to the student;
3. Presentation accommodations include adjustments in how test items are presented to the student; and
4. Response accommodations address how the student answers or completes the test items.

Disallowing appropriate, or valid, accommodations prevents students with special needs and/or injuries from demonstrating their competence. Moreover, certain accommodations have been found beneficial to both students with and without disabilities. For example, reading test directions aloud helps to ensure that all test takers, wherever they are seated, have access to the same information (Redfield, 2001).

Universally designed assessments

In an effort to increase accessibility to structures, architects have developed a term called universal design. The idea behind universal design is to consider access of structures from their initial development, so that they become accessible to all people, including those with disabilities (Johnstone, Altman & Thurlow, 2006). Therefore, the goal of universally designed assessments is to provide the most valid assessment possible for the greatest number of students, including students with disabilities. It is an approach to test design that seeks to

maximize accessibility for all intended test-takers (AERA et al., 2014). Universal design makes tests better for every student (Johnstone et al.).

The elements of universal design, according to Thompson, Johnstone, and Thurlow (2002), are varied and include maximum legibility. Legibility refers to the capacity with which items can be deciphered with ease. According to Thompson et al., the following recommendations increase the legibility of test:

1. Contrast: White or glossy paper should be avoided to reduce glare. Black type on matte pastel or off-white paper is most favourable.
2. Type Size: 12-point type increases readability and can increase test scores for both students with and without disabilities, compared to 11 and 10-point type.
3. Spacing: Letters that are too close together are difficult for partially sighted readers. Spacing needs to be wide between both letters and words.
4. Leading: Leading should be 25-30 percent of the point (font) size for maximum readability.
5. Typeface: Standard serif or sans serif fonts with easily recognizable characters are recommended. Text printed completely in capital letters is less legible than text printed completely in lower-case, or normal mixed-case text. Italic is far less legible than regular lower case.
6. Justification: Staggered right margins are easier to see and scan than uniform or block style right justified margins. Text that is flush to the left margin is easiest to read.
7. Line Length: Lines of text should be about 40-70 characters, or roughly eight to twelve words per line.

8. Blank Space: A general rule is to allow text to occupy only about half of a page. Too many test items per page can make items difficult to read.

Testing Standards, Guidelines and Codes of Practices

Assessments depend on professional judgment. "Testing standards, guidelines, and codes of practices are developed by large committees or testing publishers to provide guidance on fairness practices for the broader educational communities" (Xiaomei, 2014, p. 51). Standards, guidelines, and codes of practices identify issues to consider in exercising professional judgment and in striving for the fair and equitable assessment of all students (JCTP, 2004).

However, not all of such documents are useful and relevant to all testing purposes. Gipps and Stobart (2009) noted that fairness considerations in large-scale high-stakes testing might be different from fairness considerations in classroom teacher-made testing. Therefore, for the purposes of usefulness and relevance, I considered only standards, guidelines and codes that pertain to large-scale testing, and these include:

1. The Standards for Educational and Psychological Tests (AERA et al., 1999; 2014), which is geared primarily for test developers, researchers, and psychometricians.
2. Responsibilities of Users of Standardized Test (JCTP, 2000), which provides a concise statement useful in the ethical practice of testing.
3. ETS Standards for Quality and Fairness (ETS, 2014), which helps to design, develop, and deliver technically sound, fair, accessible, and useful products and services.

4. The Principles (Joint Advisory Committee on Testing Practices, 1993), which was developed primarily in response to inappropriate use of large-scale assessment results in Canada.
5. Code of Professional Responsibilities in Educational Measurement (NCME, 1995), which serves as a statement of professional responsibilities for stakeholders in testing.

Code of fair testing practices in education (JCTP, 2004)

The technical nature of the standards and other guidelines makes it difficult to be easily interpretable by educational practitioners such as teachers and researchers. To assist the stakeholder, a working group of the Joint Committee on Testing Practices (JCTP) prepared the Code of Fair Testing Practices in Education. The Code represents selected portions of the Standards and other guidelines in a way that is relevant and meaningful to different stakeholders (JCTP, 2004). Thus, the Code serves as an appropriate framework for conducting research on fair testing practices in education.

The Code has a number of advantages over other frameworks, such as Kunnan's (2004) test fairness framework. According to Xi (2010), Kunnan's framework fails to provide "practical guidance on how to go about developing the relevant evidence to support fairness" and thus, does not "offer a means to plan fairness research" (p. 148). The Code specifically defines qualities devoted to the responsibilities of test developers and test users regarding the importance of their roles (McNamara & Roever, 2006). In addition, the Code acknowledges intra-

group differences or individual differences, such as test taking strategies, regarding the ability being tested.

The Code (JCTP, 2004) attempts to condense the most salient statements concerning the responsibilities of test developers and test users from existing codes and standards in four areas: (1) Development and selection of tests; (2) Administration and scoring of tests: (3) Reporting and interpretation of test results; and (4) Informing test takers. However, for the purposes of this study, statements concerning the selection of test were not considered. Moreover, statements on fair uses of district test results were treated as a theme on its own. "Fairness is a fundamental validity issue and requires attention throughout all stages of test development and use" (AERA et al., 2014, p. 49). Hence adaptation is needed in using this framework for this study.

To guide my fairness investigations, I developed an operational model that focused on seven phases of fair testing practices, which were defined by different roles of the test developer and teachers in this study. The roles of test developers and teachers during the seven phases of fair testing practices would ensure absence of bias against any student, equitable treatment of all test takers, and opportunity for students to learn the content of the district-mandated test. The following model (See Figure 1) provides a visual representation of this study's conceptual framework. One of the important contributions of this model is that it provides practical guidance for fairness investigations. Through the accounts of the test developer and teachers, it is possible to identify issues such as

misalignments and inhibiting classroom practices that might be undetected by statistical approaches such as DIF.

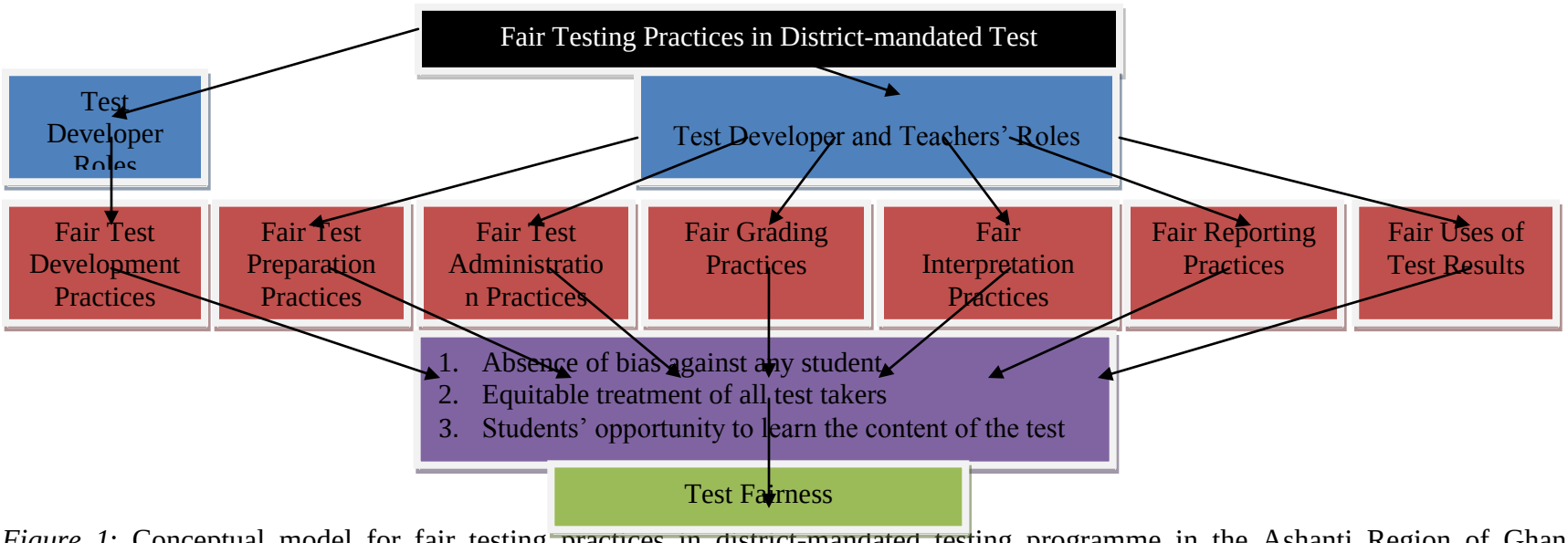

Figure 1: Conceptual model for fair testing practices in district-mandated testing programme in the Ashanti Region of Ghana

Fair Test Development Practices

The work of ensuring test fairness starts with the design of the test and test development (Tierney, 2013). Test development is the process of creating all aspects of a test and putting them together in a purposeful way designed to accomplish the overall aims of the test. AERA et al. (1999) list four phases of test development:

1. Delineation of the purpose(s) of the test and scope of the construct or extent of the domain to be measured;
2. Development and evaluation of test specifications;
3. Development, field testing, evaluation and selection of the items, scoring guides and procedures; and
4. Assembling and evaluation of the test for operational use.

Purpose(s) of test

Good assessment requires clarity of purpose, goals, standards and criteria. The initial step in constructing an examination is to delineate the assessment objective. Once established, the purpose of the examination provides a foundation for subsequent test construction (Plake & Jones, 2002). The purposes and intended uses of assessment systems are central in making decisions about the assessment instruments and procedures to be included in the system.

Evaluation, accountability, and improvement have been identified as the three primary objectives of large-scale assessments (Redfield, 2001). Each possible goal has implications for the design of an assessment system. For instance, an assessment needs to align with content standards if it intends to serve the purpose of improving student learning. Test items used for the

purpose of promotion must necessarily be the same for every test taker while test items used for pupil diagnosis may differ from one test taker to the other (Redfield). An assessment system serves the purpose of increasing and sustaining teachers' motivation to teach well when teachers participate in the development process. This can range from participation in the drafting of content standards, to review of test items, and professional development programmes for teachers (Redfield). In many cases, assessment is designed to serve more than one of these purposes (Rahn & Stecher, 1997).

Characteristics of test takers

Henning (as cited in Jaturapitakkul, 2013) argued that a test should never be developed without due consideration of the characteristics of the intended examinees. By doing so, all test takers would be given a fair chance to be tested with an appropriate test to reflect the test takers' actual ability. Items that respect the diversity of the assessment population are sensitive to test taker characteristics and experiences such as gender, age, ethnicity, socioeconomic status, region, religion, disability, and language (Klinger & Luce-Kapler, 2007). Item writers, therefore, need a description of the diverse needs of the population of students tested within a particular district (Willingham & Cole, 1997).

According to Redfield (2001), students' characteristics must be taken into account during the development of an assessment instrument. Factors to consider might include:

1. Coverage of the full range of student knowledge and skills in the content areas tested. This will help ensure that students at all levels of learning have an opportunity to demonstrate their knowledge and skills.

2. Providing students, the opportunities to demonstrate knowledge and skills in a variety of ways.
3. Appropriate assessment techniques for students with disabilities.
4. Content and contexts appropriate for the cultural and ethnic diversity in the student population.

Another important characteristic to be considered is the test takers' experience in taking such test as the format of a district-mandated test's items. Students' familiarity with test items' format helps to reduce measurement errors (Redfield).

Test specification

According to Popham and Lindheim (1980), "a test development project begins with a careful consideration of the skills or attitudinal characteristics proposed for measurement" (p. 3). The easiest way to ensure a representative sample of content and cognitive objectives on a test is to prepare a test specification (Suskie, 2000). Popham (2002) referred to test specification as "rules to be followed in constituting the overall nature of the test" (p. 138).

Adequately defining the knowledge and skill domain to be encompassed by the content standards requires input from knowledgeable experts, including content specialists, and especially, teachers experienced with the wide variety of students enrolled in a district's schools (Redfield, 2001). According to Etsey (as cited in Oduro-Okyireh, 2008), the table of specification makes sure that justice is done to all the topics covered in a course.

Item writing

Once the test specifications are complete, item development can begin. Item specifications indicate the defining characteristics of the item class, the rationale for the class, and the required characteristics for each item component (Popham, 2002). Specifications for item writers might include specific information on fairness concerns, item content and scope, item types, item skill levels, expected item difficulties, non-discriminatory subject matter, and language usage (Tierney, 2013). Tierney further stated that universal design principles must be followed in the development of district-mandated assessments.

If mandated tests are to do minimum harm and maximum good, it follows that item writers should include well-qualified teachers who have extensive content knowledge, and who represent genders as well as diverse ethnic backgrounds and geographic locations (Downing, 2004). Active schoolteachers from a variety of schools, small to large, public and private, should be included (Tierney, 2013). Effective item writers are trained, not born (Haladyna, 2004). Item writing usually takes place during item writing workshops where item writers are trained (Oregon Department of Education [ODE], 2011). Without specific training, most novice item writers tend to create poor-quality, flawed test questions that test trivial content (Haladyna). Thus, one of the more important validity issues associated with test development concerns is the selection and training of item writers (Haladyna & Downing, 2004).

Test content is less secure when teachers participate in the writing of the test items. Ideally, item writers are given an assignment to produce a small

number of items. The small number of items assigned to each item writer ensures that materials on a wide variety of topics would be produced and that the security of the testing programme would be maintained, since any item writer would have knowledge of only a very small proportion of the items produced (Tierney, 2013). Alternatively, the security of district-mandated test content can be maximized by creating a secure bank of test items, store the items in a library with highly restricted access, and then have designated staff generate assessments by selecting a subset of the items in the bank. To the extent that item writers are unaware which specific items would appear on a given test, the security of the test would be maximized.

Item review

All items or tasks eligible to be selected for a test must first be reviewed and evaluated both qualitatively and quantitatively. According to Haladyna (1999), all items must go under review for content, item writing violations, and grammatical errors. Messick (1989) emphasized that all items must be reviewed for factors that would impact the degree of difficulty of the test, or test biases. Generally, qualitative reviews pre-empt differential item functioning in test items. It is partially aimed at avoiding construct irrelevant variance (Rudner, 1994). Qualitative reviews usually can take two forms: Content review and Fairness review. Content review panels appraise the technical quality of items, looking for items that are free from such flaws as inappropriate readability level, ambiguity, incorrectly keyed answers and distracters, unclear instructions, and factual inaccuracy (ODE, 2011).

There are many factors to consider when evaluating the quality of test items. For example, one might examine the extent to which items conform to

widely accepted item-writing guidelines (Haladyna, Downing, & Rodriguez, 2002). A number of scholars (Burton, Sudweeks, Case & Swanson, 2002; Collins, 2006; Etsey, 2012; Haladyna et al., 2002; McDonald, 2008; Merrill & Wood, 1991; Penn, 2009; Sudweeks, Merrill, & Wood, 1991) have suggested the following rules for writing multiple-choice items:

1. Provide clear directions at the beginning of each section of the test.
2. Begin the test with a few quick and easy questions.
3. Avoid changing pages in the middle of an item.
4. Base each item on a specific problem stated clearly in the stem.
5. Use proper grammar, punctuation, and spelling.
6. Avoid using unnecessarily difficult vocabulary.
7. State the stem in positive form (in general). Whenever negative wording is used in the stem, emphasize it.
8. Keep each item independent from other items.
9. Include as much of the item as possible in the stem, but do not include irrelevant material.
10. Keep the alternatives homogeneous in content.
11. Keep the alternatives parallel in form.
12. Keep the alternatives similar in length.
13. Avoid the use of specific determiners.
14. Use plausible distracters.
15. Avoid the alternatives “all of the above” and “none of the above” (in general).
16. Present the answer in each of the alternative positions approximately an equal number of times, in a random order.

17. Use a vertical format for presenting alternatives.
18. The responses/options in agreement must be in alphabetical/sequential order.
19. The expected response should not be put at the beginning of the stem.
20. Be consistent in the number of options used.

According to Reiner, Bothell, Sudweeks, and Wood (2002), the following guidelines also help to eliminate deficiencies in essay items that serve as CIV:

1. Allow adequate time to answer the questions.
2. Essay questions should be used only to evaluate higher-order outcomes.
3. Avoid the use of optional questions.
4. Specify the relative point value and the approximate time limit in clear directions.
5. Make sure questions are sharply focused on a single issue.
6. Use several relatively short essay questions rather than one long one.
7. Improve the essay question through preview and review.

Flawed test items result from the violation of one or more of these standard item-writing principles (Case & Swanson, 2002; Haladyna, 2004). Deviating from established guidelines for writing test items can be problematic because it can detract from the quality of individual items and of the test as a whole (Downing, 2002; Tarrant & Ware, 2008).

Sensitivity review panel, on the other hand, reviews items for bias, controversial content and overly emotional issues. Grade level experts, representatives of major cultural and disability groups, researchers and teaching professionals all make up an effective review team.

Quantitative reviews consider statistical indices that are computed on results of a trial test or field test of items. The purpose of pretesting items is to determine whether the items are technically sound and at the appropriate level of difficulty for the examinee population. Statistical indices of item difficulty and discrimination, among other statistics, can be compiled on the basis of pre-test results (Tierney, 2013).

Assembling test items for operational use

The last phase of test development involves assembling and evaluation of the test for operational use. In this phase, the test developer identifies and selects the best items or tasks based on test specifications and psychometric properties. Assembling of the test according to test specifications should be documented as part of the total test system (AERA et al., 2014). Considerations at this stage include, how items should be ordered and grouped; how items will look on a page; how the test should be printed; and how test security should be maintained during storage and transport of test materials (Chan, 2009).

Practice test

When developing classroom assessments, opportunities should be provided for students to practice with new assessment formats and item types (JCSEE, 2003). According to the Standards (AERA et al., 1999):

> When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identified with each major area in the test's classification or domain should be provided to the test takers prior to the administration of the test or included in the

testing material as part of the standard administration instructions. (p. 47).

Practice test provide clear instructions and, support students' understanding of what will be required during the assessment. Testing instructions are often supplemented with practice exercises for test takers prior to the operational testing period as a means of reducing construct-irrelevant test variance (Plake & Jones, 2002). Practice or sample questions, activities, or tasks support students' understanding of what will be required during the assessment (AERA et al., 2014).

Fair Test Preparation Practices

Test takers have the right to be informed, prior to testing, about the test's purposes, the nature of the test, whether test results will be reported to the test takers, and the planned uses of the results (Camara, 2007). Assessment specialists (Guskey & Jung, 2009; McMillan, 2011; Suskie, 2000) have recommended for decades that explicit learning expectations and assessment criteria be shared with students. Test preparation activities for students must have two goals (ISBE, 2014):

1. To ensure that all students have the opportunity to learn in accordance with an appropriate curriculum and to become knowledgeable about the content covered by the tests; and
2. To give all students occasion to become familiar with the types of questions used on the test (multiple-choice, extended-response, and short-response questions) so that students are tested for their knowledge and ability, not their test-taking skills.

The practice of preparing students for mandated tests tends toward the use of ethically acceptable strategies. A number of researchers (Camara, 2007; Nitko, 2004; Plake & Jones, 2002) have recommended that testing professionals should provide the following information to test takers:

1. Give or provide test takers with access to a brief description about the test purpose and the kind(s) of tests and formats that will be used.
2. The content and abilities that will be assessed.
3. When the assessment will be administered.
4. The planned use(s) of the test results.
5. Information about how the test will be scored and in what detail.
6. How to request appropriate accommodations or modifications.
7. Which materials or personal possessions are required for the tests, which are permitted, and which are prohibited.
8. Appropriate test-taking strategies or skills.
9. Actions that constitute misconduct during the test administration, and the consequences of such misconduct.
10. Motivating students to do their best.

Individuals perform best when they understand the purpose of what they are doing. Common sense and research show that being well grounded in the test content and form can improve test performance (ISBE, 2014). According to Dodeen (2015), test-taking strategies affect other factors such as reducing test anxiety and improving students' attitudes toward tests. Training students in test-taking skills improves students' chances of showing their actual knowledge on the content of a test (Maxwell, Cumming, Wyatt-Smith, & Colbert, 2012). However, Khoii and Shamsi (2012) suggested that secure

test questions or questions that are similar or altered versions of secure test questions for practice constitute unethical test preparation practices.

Whether or not students received test preparation can be a source of CIV. If some students in a school have received test preparation, and another group of these students have not, differences in the performance of these groups of students might be attributable to the fact that some students received test preparation and others did not. There should be some evidence that all students received uniform and ethical test preparation (Gipps, 1995).

Fair Test Administration Practices

Pre-administration procedures/practices

Large-scale programmes, like district testing, require adherence to testing directions. A comprehensive test manual is essential to properly administer the test. "Test developers promote fair testing practices for test takers by establishing standardized testing and scoring procedures, and communicating these procedures to test sponsors, administrators, and test takers" (Plake & Jones, 2002, p. 6). For paper-and-pencil tests, the pre-administration procedures include verifying the integrity of materials prior to the test date and maintaining their security and confidentiality (Plake & Jones).

Cizek (1999) suggested that states and school districts should adopt policies, which clearly forbid school personnel to look at 'high stakes' test questions except as needed during administration. He further argued that Shrink-wrapping of materials, providing only the exact number of tests required, specifying accounting procedures for test materials, and instituting test monitoring teams have also helped reduce test security problems.

Actual administration procedures/practices

Etsey (2012) presented a number of standardized procedures for the actual administration of an achievement test, including:

1. Testing environment should be free from distractions.
2. The sitting arrangement must allow enough space so that pupils will not copy each other's work.
3. Pupils should start the test promptly and stop on time.
4. Announcements must be made about the time at regular intervals.
5. Invigilators are expected to stand at a point where they could view all students.
6. Invigilators should not be allowed to read novels, newspapers or grade papers.
7. Invigilators should avoid threatening behaviours.
8. Students should not be warned to do their best.
9. Students should not be told to work faster or threatened dire consequences if they fail.
10. When an item is ambiguous, it should be clarified for the entire class of test takers.

Actual administration practices should also protect the security of a test to help maintain the meaning and integrity of each test score (Xiaomei, 2014). McCabe, Trevino, and Butterfield (2006) advocated communicating clear expectations for honest behaviour, including clear consequences for dishonest behaviour, and enforcing those consequences. Teachers can show students that they expect honest academic behaviour by appearing aware of the class, noticing and acting on questionable behaviour, and moving through

the classroom during the test (Cizek, 1999). Herman and Golan (1991) also noted that assigning seating has an added benefit of preventing students from choosing cheating partners. A number of researchers (ISBE, 2014; McMillan, 2011; Wormeli, 2006; Zucker, 2004) have suggested actions that violate test security, and they include:

1. Testing students out of sequence from the district's testing schedule.
2. Making remarks about quality or quantity of student work.
3. Providing answers to a student.
4. Leaving the room while administering a test.

The administration of high stakes test to an individual is a potentially stressful event. Administrators should therefore avoid any action or behaviour that makes students anxious. According to Gordon and Fay (2010), things that create anxiety are (1) warning students to do their best because the test is important, (2) telling students that they must work fast in order to finish on time, (3) threatening dire consequences if they fail, and (4) threatening students with tests if they do not behave. It is also important to note that in certain cases changes to the administration of a standardized assessment may be appropriate in order to provide some students a comfortable environment to demonstrate their knowledge. There can be a change in format, response, setting, or timing that does not alter in any significant way what the test measures. Accommodations that are appropriate are often detailed in the Individualized Education Plan (IEP) for a student with a disability (Nitko, 2004). However, students with an injury that would make it difficult to participate may use, as appropriate, any of the universal test accommodations conditions (Alberta Education, 2009).

At the end of a test sitting, it is also important to inventory all secure test materials, and include notes on irregularities encountered, and accommodations used, during test administration (JCTP, 2000). For large-scale testing programmes, testing agencies typically engage the services of well-trained test administrators to supervise test administration. These professional proctors (i.e., test administrators) are key to successful, secure, well-organized and well-administered large-scale examinations (Downing, 2004). It is generally preferable to designate a single, highly experienced test administrator as 'chief proctor' who assumes full responsibility for all aspects of the secure test, including supervision of other proctors (Cheng, 1998).

Fairness in Grading Students' Test Performance

Grading is the most common method of communicating whether a student has learned something or not (Allen, 2005). Grading refers primarily to the process of using a system of symbols for reporting various types of students' progress. Grading for summative purposes provides a report about how well a student has achieved the curriculum learning targets (Nitko, 2001). Grades summarize assessments, made by teachers, of students at the end of a specified time (Allen). This is done through the use of a letter code or percentage that represents the overall quality of student work (Green & Emerson, 2007).

A fair grade should be based on the student's competence in the academic content of the course (Close, 2009). It reflects an expert assessment of the student's actual achievement. A grade "provides an accurate undiluted indicator of a student's mastery of learning standards" (Wormeli, 2006, p. 18). Therefore, care should be taken to ensure that results are not influenced by

factors that are not relevant to the purpose of an assessment (JCTP, 1993). If grades are intended to measure student achievement, then they likely should not take into account school behaviours such as participation in class and comportment (McMillan, 2001).

Comportment concerns student behaviour that does not conform to school rules, and includes absence from school, tardiness, fighting, inappropriate language, defiant behaviour, and dress/grooming code violations. Attendance is probably the most common form of comportment that finds its way into course grades (Close, 2009). Regardless of its effectiveness, a grading practice that considers students' comportment is unfair. A student's grade should never be affected by virtues such as his or her cheerfulness, helpfulness, dedication, sensitivity, and other moral virtues (Close). Younger, Warrington, and Jaquetta (1999) claimed that boys get worse marks as a consequence of their uncooperative classroom behaviour. Moreover, when the intent of a written assessment is to assess content and thinking alone, stylistic factors such as handwriting, vocabulary, or sentence structure, should not form part of students' grade (Wormeli, 2006).

Brookhart (as cited in Nitko, 2001, p. 338) stated that "what teachers seem to intend when they add nonachievement [*sic*] factors to grades is to mitigate negative social consequences, but grades are not the appropriate tool for social engineering." Teachers may value both social behaviour and achievement, but if the grade they report intertwines the two, they are communicating poorly and encouraging confusion (Nitko), which is considered unfair to both students and parents.

Fairness in Interpreting Students' Test Performance

Interpreting students' results refers to the procedures used to combine assessment results in the form of summary comments and grades, which indicate (1) a student's level of performance, and (2) the valuing of that performance (JCTP, 1993). According to Ebel and Frisbie (1991), test scores can be made meaningful by referencing them separately to some expected score level; to scores of other individuals; or to scores that represent different performance levels. Thus, there are two popular ways of interpreting test scores: (1) Norm-referenced interpretation, which describes test performance in terms of a student's position in a referenced group that has been administered the assessment; and (2) Criterion-referenced interpretation, which describe test performance in terms of the kinds of tasks a person with a given score can do (Etsey, 2012).

However, the interpretation of scores on any test should take place along with a thorough knowledge of the technical aspects of the test, student's personal and social context, and limitations in the assessment methods used. Many factors can impact the valid and useful interpretations of test scores. These can be grouped into the following categories (JCTP, 1993):

1. Psychometric Factors. Factors such as the reliability, norms, standard error of measurement, and validity are important when interpreting test results.
2. Test Taker Factors. Specifically, the test user should evaluate how the test taker's gender, age, ability, motivation, opportunity to learn, self-esteem, socio-economic background, special interests, special needs, and test-taking skills, impact on the individual's results.

3. Contextual Factors. The relationship of the test to the instructional programme, quality of the educational programme, home environment, and other factors that would assist in understanding the test results are useful in interpreting test results.

According to Etsey (2012), factors affecting validity are several, and includes (1) too difficult reading vocabulary and sentence structure, (2) ambiguous statements in assessment tasks and items, (3) inadequate time limits, (4) inappropriate level of difficulty of the test items, (5) poorly constructed test items, (6) cheating, (7) unreliable scoring, (8) fear of the assessment situation, and (9) testing conditions.

Fair Test Reporting Practices

Following administration of a large-scale achievement test, students' scores must be reported in a timely and confidential manner, and provided in a format that is clear, relevant, and useful to each intended audience. Examinees have a right to a precise, timely, meaningful, and useful report of their performance on a district-mandated test. Score reports must be written in language that is understandable to all recipients (Cheng, 1998).

Grades are often not detailed enough to give parents or students a thorough understanding of the overall gains made by students in the classroom. Thus, teacher comments often convey whatever information has not been completely explained by the grade (Alberta et al., 2006). Cizek, Germuth, and Schmid (2011) provided the following criteria to guide practices related to reporting students' results:

1. Share specific examples of what a student knows and can do in relation to the student learning outcomes.

2. Communicate student progress and identify next steps for learning.
5. Identify strategies currently being employed at school and suggest, where appropriate, strategies for how the parent can support the student's learning at home.
6. Report both strengths and weaknesses of students so that strengths can be built upon and problem areas addressed.
7. Report achievement, effort, attitude, and other behaviours separately.
8. Modify reporting procedures for students with special needs based on their individual education plans.

Uses of Mandated-test Results

Assessment can be a powerful tool in education when used appropriately. However, it can also be used to the detriment of students, teachers and school districts when it is used in capacities beyond what it was designed for. District-mandated assessments are given in schools for different reasons including, identifying students at-risk for academic failure; identifying teachers in need of support; ranking students by achievement level; and comparing students' ability levels with their achievement. All of these are valid reasons to use assessment results (Johnson, 2008). Johnson further stated that assessments are necessary to gauge learning, to monitor student progress and to identify students who may need extra support. ITC (2001) suggested four defensible uses of mandated tests:

1. Informing parents about their children's relative achievements.
2. Informing teachers about their students' relative achievements.
3. Selecting students for special programmes.
4. Allocating supplemental resources.

According to Tunks (2001), the use of tests for purposes other than diagnosing learner needs and measuring student progress in instructional content constitute a concern. The following has been identified as inappropriate uses of external achievement test results: (1) Evaluating schools; (2) Evaluating teachers; (3) Promoting or grading students; and (4) Making classroom instructional decisions (ITC, 2001). This is mainly because measurement experts agree that it is inappropriate to use performance on a single test for making high-stakes decisions for individuals and schools (Rudner & Schafer, 2002). Multiple indicators are essential so that students who are disadvantaged on one assessment have an opportunity to offer alternative evidence of their expertise (Linn, 2003). Moreover, to assign a rating of teacher effectiveness may not be a fair idea if mandated assessments are not being maximally aligned to standards (Johnson, 2008). AERA et al. (1999) indicated that when district or other authorities mandate educational testing programme, the intended uses of test results should be clearly described.

Empirical Review

Findings on fair test development practices

Yip and Cheung (2005) suggested that confusion about assessment purposes threatens fairness. This is particularly important because no single test can serve all purposes of testing. In a final report for a multi-stage study of student assessment in Alberta, Canada, Webber et al. (2009) provided an overview of responses from educators, secondary students, and parents as to the importance of a number of assessment purposes. All categories of respondents assigned very high levels of importance to the following purposes

of assessment: (1) Promote high standards; (2) Focus on provincial curriculum; (3) Promote improvement; (4) Inform students and parents; (5) Inform teachers; (6) Inform school district staff; and (7) Identify professional development needs of teachers.

Testing professionals recognize the need to align the curriculum with test content. However, research findings (Polikoff, Porter, & Smithson, 2011; National Board on Educational Testing and Public Policy [NBETPP], 2003; Webber et al., 2009) on the alignment of mandated tests to State's curriculum or standards have yielded mixed results. Polikoff et al. inspected the content of mandated assessments and State's standards of multiple (31) States. They found that there were significant mismatches between the mandated assessments and the State's standards that guide teachers' instruction. This implies that students in these States were not given adequate opportunity to learn the content of the mandated assessments. Students' performance on these test, therefore, do not reflect how well they have mastered the objectives of the State's curriculum. However, in contrast to the previous finding, NBTEPP (2003) found a vast majority of teachers indicating that a mandated test is aligned to the State's curriculum.

Downing (2002) evaluated the effect of sets of flawed items on the quality indices of an educational achievement test and found that flawed items were generally more difficult and failed more students than comparable standard items. Flawed item formats were more difficult than standard, non-flawed item formats for students in three of four examinations studied. Passing rates tended to be negatively impacted by flawed items. Poorly crafted, flawed test questions tended to present more of a passing challenge for the test takers,

and thus, students' scores on the three examinations studied could be said to be biased because passing challenge, due to flawed test questions, is irrelevant to the constructs as measured on the three examinations.

Findings on fair test preparation practices

In order to facilitate the Basic Education Comprehensive Assessment System (BECAS), the Ghana Education Service's Curriculum Research and Development Division (CRDD) authorized a team of researchers to investigate students' opportunity to learn English and Mathematics in primary school (CRDD, 2005). A major finding of the study was that opportunity to learn standards for most schools was very low and that the majority of teachers completed only 60% of the content of the Mathematics and English syllabuses. This result raises test fairness concerns, especially, when students are assessed externally. This is because external test covers the full range of the syllabus or curriculum on which it is based, and thus, students are likely to be tested on contents that they had no opportunity to learn.

Traub and MacRury (as cited in Struyven, Dochy, & Janssens, 2005) reviewed the empirical research on multiple-choice (MC) and free-response (FR) tests since 1970, and concluded that students are influenced by the expectation that a test will be in MC or FR format. Students expecting MC format reported more positive attitudes towards the tests. However, the performance of students expecting MC test format was not significantly different from that of students told to expect a FR test format, but students expecting a FR test format performed significantly better on the FR test than students told to expect a MC test. Thus, the assessment expectation of students

on a district-mandated test would seem to prepare them in a distinctive way, which will reflect their true standings on the constructs as measured.

The effect of teaching test-taking strategies on students' performance has also been investigated. Gallagher (1998) examined the hypothesis that differences in test performance on math between males and females are the results of differences in the students' strategies for solving math problems. Results showed significant differences in strategies used by males and females in math tests. Differences in test performance as a result of differences in test takers' test-taking strategies introduce construct-irrelevant test variance into students' achievement test scores.

Findings on fair test administration practices

In a survey of state test policies, Cannell (1989) found that most district test policies do not address issues of test security, and therefore open boxes of unsealed tests are delivered to schools, weeks before the test is to be given. Teachers are often given unsealed test booklets days before they are scheduled to administer the test. Such lapses in test security policies can influence teachers to teach the test, which create a biased test score.

Nicole (2013) tested the effects of stereotype threat by altering the test administration of two groups of test-takers. One group of test takers were told that the test had shown gender differences in the past, while the other group of test-takers were informed of no such gender differences in the past. The results showed that women under performed in relation to men when test-takers were informed that the test had gender differences but when they were told that gender differences did not exist, women performed at the same level as men. The differences in test performance could be explained by differences

in females and males' ability to deal with test anxiety, resulting from stereotype threat. However, one's ability to deal with test anxiety is irrelevant to the purpose of classroom achievement testing.

On testing accommodations, research findings indicate that teachers, in general, agree that students with special needs should be accommodated on mandated tests. For example, the quantitative findings of Webber et al. (2009) revealed that teachers show high levels of agreement (94.8% agreed or strongly agreed) that students with special needs should have access to accommodations for assessments. In a similar vein, the perception that teachers do change assessments for students with special needs was strongest for educators at 87.4%. Moreover, Elhoweris and Alsheikh (2010) reported that UAE teachers as a group considered testing adaptations as helpful for students with disabilities. The provision of testing accommodations ensures that achievement test results are not influenced by test takers' irrelevant characteristics such as special needs or injuries.

Findings on factors that influence students' grades

Webber et al., (2009) reported that elementary students in their study agreed to the assertion that grades on report cards are influenced by student's good/naughty behaviour. An elementary teacher added some support to these perceptions by noting that "assessment can be used for behavioural control, whether intentional or not" (p. 132). In contrast to this finding, Tierney, Simon, and Charland (2011) reported that teachers in their study did not consider students' attitude, motivation, or participation in calculating grades. There were, however, fewer consensuses among teachers about students' effort, with one-third reporting that they considered students' effort in

calculating grades. This finding is supported by the findings of Green, Johnson, Kim, and Pope (2007), and Zoeckler (2005), as teachers in these studies also continue to weigh student effort in grading.

Dee (2007) found that teachers in the study give better grades to students of their own gender. In England, Gibbons and Chevalier (2007) also found teacher biases depending on students' race and gender. Grades that are influenced by teachers' predispositions reduce the degree of soundness of the interpretations and uses of the assessment results. However, in Sweden, Hinnerich, Hoglin and Johanneson (2011) also investigated teacher biases in grading practices and found significant teacher biases by student ethnicity but not by student gender. This is supported by the findings of Anhwere (2009) who reported that the scoring practices of tutors in teacher training colleges in Ghana are never influenced by the students' gender.

Findings on fair reporting practices

Webber et al. reported 95.3% of educators indicating agreement or strong agreement to the item; "Teachers regularly discuss with students, ways of improving their grades". However, there was some doubt that student achievement was being reported accurately. Parents in this study identified deficiencies in reporting formats, citing problematic use of educational jargons. These sentiments by parents are supported by the findings of Lekoko and Koloi (2007), which asserted that when teachers grade students' work, they did not provide adequate comments that could help students understand where they went wrong.

Lack of specific comments does not encourage positive communication between teachers, parents and students. Also, students would

not be motivated to do their best on subsequent examinations when adequate comments are not provided.

Findings on uses made of mandated tests results

According to NBETPP (2003), decisions about individual student's placement or grades are clearly beyond the scope of what most teachers see as appropriate uses of mandated test results, whereas decisions about global planning of instruction are viewed as appropriate. Based on the responses of all teachers in the study, NBETPP reported top three uses of state test results: (1) "assess my teaching effectiveness" (38%); (2) "give feedback to parents" (35%) and "give feedback to students" (29%); and (3) "evaluate student progress" (20%). Less than 10% of teachers indicated that they used the results to "group students within my class" or "determine student grades in whole or in part".

According to the findings of a survey conducted by Herman and Golan (1991), teachers managed the sequence of presenting their teaching materials based on what was included in external assessment. They found that external examinations substantially affect teachers' instructional planning. On the contrary, Valazza (2008) found relatively little effect of standardized tests on teacher decision-making such as placing students, planning instruction, or grading.

The findings that teachers plan instruction based on the content of mandated assessments leads to *teaching to the test*. Such a practice denies students an opportunity to learn contents not covered by mandated assessments, and therefore, defeats the purpose of mandated assessments. Also, the finding that mandated test results have little effect on teachers

grading practices provides an opportunity for students to demonstrate their knowledge in other similar settings, and thus, considered as fair testing practices.

Summary of Literature Review

The review of literature obviously disclosed that not much study has been conducted on the topic within the Ghanaian context. In conclusion, this literature review synthesizes various conceptualizations of test fairness from three broad views: (1) Absence of bias; (2) Equitable treatment; and (3) Students' opportunity to learn. The literature review identified relevant practices that ensured and/or violates these three broad views. Previous findings revealed that test items that violate standard item-writing rules serve as CIV, and create a biased score for students (Downing, 2002). However, previous findings that mandated test results have little effect on teachers grading practices ensures absence of bias (NBETPP, 2003; Valazza, 2008).

Moreover, the provision of testing accommodations by teachers, as indicated by previous findings, ensures that students with special needs and/or injuries are given equitable treatment in the administration of mandated assessments (Elhoweris & Alsheikh, 2010; Webber et al., 2009). Nevertheless, lack of test security policies gives some students an unfair treatment over others (Cannell, 1989).

Finally, previous findings about misalignment between State's curriculum and mandated assessments do not provide students an opportunity to learn the content of mandated assessments (Polikoff et al., 2011; Webber et al., 2009). Nonetheless, informing students about the format of mandated assessments seem to prepare them in a distinctive way.

CHAPTER THREE

RESEARCH METHODS

Introduction

I sought to determine the degree of test fairness in a district-mandated testing programme in the Ashanti Region of Ghana by investigating the fair testing practices of a test developer and teachers. This chapter describes the procedures adopted in conducting this study. It embraces the research design, study area, population, sampling procedure, and data collection instruments. The procedures for data collection and methods of data processing and analysis are also discussed.

Research Design

The research objectives and questions posed necessitated the collection and analysis of qualitative and quantitative data at different levels of the participants of the study. Thus, the research design adopted for this study was a multilevel mixed-methods triangulation design. Mixed-method studies, based on pragmatic worldview, according to Onwuegbuzie and Leech (2007), involve the collection, analysis and interpretation of both qualitative and quantitative data. In a multilevel design, quantitative and qualitative methods are used to address different levels within a study in order to address a research problem. The main purposes for the multilevel design were to seek convergence and corroboration of results from different methods studying the same phenomenon, and expand the breadth and range of inquiry by using

different methods for different inquiry components (Johnson, Onwuegbuzie, & Turner, 2007).

A qualitative method: instrumental case study was employed at the test developer's level. Instrumental case study is used to gain insight and understanding of a particular situation or phenomenon (Baxter & Mislevy, 2005). On the other hand, a quantitative method, cross-sectional survey, was adopted in order to generalize statements about the roles of teachers, and the differences in fair testing practices that exist among them in terms of their levels of training. "A cross-sectional survey is one in which data are collected from selected individuals at a single point in time" (Gay et al., 2009, p. 176). The point of interface occurred at the discussion of results where I merged both qualitative and quantitative findings. Equal weight was given to findings from each level in addressing the research problem.

A number of researchers (Creswell & Plano-Clark, 2010; Teddlie & Tashakkori, 2009) have highlighted the strengths of mixed-methods design, and it includes (1) allowing the researcher to both generate and confirm theory by answering confirmatory and exploratory research questions, (2) allowing researchers to address research problems that cannot be answered by a mono-method, and (3) mixed methods research is practical in the sense that the researcher is free to use all methods possible to address a research problem. However, mixed-methods designs are not without challenges. I have to learn about multiple methods and approaches and understand how to appropriately mix them. It was also more expensive and time consuming. More specifically, the multilevel design requires much expertise and effort particularly because of the concurrent data collection and analysis (Teddlie & Tashakkori, 2009).

In order to address such challenges, the supervisors' experience and expertise in both qualitative and quantitative research was relied upon in the analysis of both data.

Study Area

I conducted the study in three municipal districts in the Ashanti Region of Ghana, namely, Atwima Nwabiagya; Asante-Akim Central; and Mampong. The Ashanti Region is located in southern part of Ghana, and it is the third largest of the 10 administrative regions, occupying one-tenth (10.2%) of the total land area of Ghana. In terms of population, however, it is the most populated region with a population of 4,780,380 in 2010. Aside its indigenous habitants (i.e., the Ashanti people), many people from other ethnic groups, regions, and countries have migrated to the Ashanti Region due to the region's arterial routes linking it to other parts of the country, and also, by the fact that it is an educational centre with a significant number of primary, secondary, and tertiary educational institutions. Almost all other ethnic groups in Ghana are represented in the region (GhanaWebb, 2016).

The region comprised both urban and rural settlements, with more than half of the population residing in urban areas (Ghana Statistical Service, 2012). The dominant religion in the region is Christianity (77.5%) followed by Islam (13.2%). The proportion with no religion is relatively high (7.3%). Agriculture, trading, manufacturing, and community, social and personal services are the four major economic activities in the region (Ghana Statistical Service). The diverse nature of the characteristics of the region's population made Ashanti Region a preferred area for test fairness investigations. Again,

my familiarity with the region's landscape facilitated the collection of the data within the limited stipulated time for the submission of the final work.

Population

Amedahe (as cited in Oduro-Okyireh, 2008) defined population as the target group about which a researcher is interested in gaining information and drawing conclusions. In this study, the target population consisted of test developers and teachers who are involved in large-scale district-mandated testing programmes in the Ashanti Region. This was made up of 2 testing agencies (private testing organisations) and 8,424 teachers in the region. According to the JCTP (2004), test developers are people and organizations that construct tests, as well as those that set policies for testing programmes.

However, for the purposes of the study, the accessible population consisted of key informants at CePME, and 1,296 teachers, representing 162 public JHSs in three Districts/Municipalities of the Ashanti Region that administer mandated assessment materials, prepared by CePME. The three Districts/Municipalities are Atwima Nwabiagya, Asante-Akim Central, and Mampong.

Sampling Procedure

Teddlie and Yu (2007) defined sampling as the process of selecting subgroups from a population of elements such as people, objects or events. This study adopted the multilevel mixed methods sampling technique. Multilevel mixed methods sampling is a general sampling strategy in which probability and non-probability sampling techniques are used at different levels of the study (Teddlie &Tashakkori, 2009).

Two separate samples were selected for test developer, and teachers in the study. Creswell (2007) defined a sample as a small proportion of a population selected for observation and analysis. Due to the research's sub-purpose of exploring the roles of the test developer by gathering in-depth data, 3 key informants were purposively sampled for the test developer level. In addition, 9 test questions were also purposively sampled in order to explore the roles of the test developer. On the other hand, a sample size of 300 teachers was used in order to analyse quantitative data.

A purposive sampling technique, critical case sampling, was utilized at the test developer level. Critical case sampling was utilized in sampling 3 key informants and 9 test questions. The 3 informants were sampled based on their knowledge of activities at CePME and expertise in educational measurement. Moreover, the 9 test questions were sampled because of its religious and ethnic content that poses fairness concerns, and also by the fact that the 9 test questions are administered in all participating schools.

The two-stage cluster sampling technique was adopted in selecting teachers in a random manner. In accordance with the proportional number of schools in the three districts, the cluster sampling technique was first utilized to randomly sample 50 clusters of public JHSs from the three participating districts in the region (See Appendix A), with each cluster consisting of approximately equal number of teachers. Then using a stratified random sampling technique, I selected 6 teachers from each of the 50 clusters of JHSs sampled, which added up to a sample size of 300 teachers, representing 23% of the accessible population. Table 1 shows a distribution of the number of schools and teachers sampled from each of the three districts.

Table 1- *Distribution of Sampled Schools and Teachers in Each District/Municipality*

Name of District/Municipality	Total Number of JHSs	Number of JHSs Sampled	Number of Teachers Sampled
Atwima Nwabiagya	74	23 (46%)	138
Mampong	54	17 (33%)	102
Asante-Akim Central	34	10 (21%)	60
Total	162	50	300

Source: Field survey, Boakye (2016)

On arrival in the respective cluster of schools sampled, I identified all teachers in the school with the help of the head teacher, and using sampling with replacement method, the required number of teachers (i.e., 6) was selected. The names of the teachers were written on pieces of paper and placed in an urn. The slips of paper were then picked one after the other without looking into the urn. Once a name of a teacher was picked, it was recorded as a sample and put back into the urn. The urn is reshuffled and the process was repeated till the required number of teachers for each school was obtained. This was to maintain the same probability for teachers in each school.

The large size of the JHSs sampled (i.e., 50), and the small nature of the cluster size (i.e., 6 teachers from each school) would compensate for the problem of decreased reliability in cluster sampling technique, due to the likelihood that people living in the same cluster tend to be homogenous, and to have similar characteristics.

Data Collection Instruments

The instruments used for this study included document and interview guide for the test developer's strand, and questionnaire for the teachers'

strand. The unobtrusive measure comprised official documents such as test questions developed by CePME. According to Merriam (1988, p. 118), "documents of all types can help the researcher uncover meaning, developing [*sic*] understanding, and discover insights relevant to the research problem." Documents review provided first-hand information on the kind of test questions given to students and the nature of the tasks they perform. Merriam (2001) contend that document analysis has the potential to reveal information that the interviewee is not ready to share. Moreover, document review was relatively inexpensive and serves as good source of background information. However, it was time consuming to collect, review, and analyse several test questions. Additionally, information gathered on test questions could be biased because of selective survival of information (Finn & Jacobson, 2008).

Data was also solicited from the test developer through the use of semi-structured interviews. "In semi-structured interviews, a researcher is able to refocus the questions, or prompt for more information, if something interesting or novel emerges" (Baškarada, 2014, p. 16). An interview protocol (See Appendix B) was devised to ensure that major topics relating to the research problem were covered. I based the content of the interview protocol, primarily, on aspects of the Code of Fair Testing Practices in Education (JCTP, 2004) that are relevant to my study. Interviews included mainly open-ended questions that yielded narrative data. According to Merriam (2001), interviewing is the best technique to use when conducting intensive case studies of a few selected individuals. Additionally, interviews were useful for gaining insight into fair testing practices, and also allowed respondents to describe what was important to them. Nevertheless, interviews are susceptible

to interview bias, time consuming and expensive, and may seem intrusive to the respondents (Finn & Jacobson, 2008).

Validity of the qualitative data was discussed in terms of trustworthiness. According to Lincoln and Guba (1985, p. 290), trustworthiness refers to findings that are "worth paying attention to". Trustworthiness was established by using the following strategies (Teddlie & Tashakkori, 2009):

1. Using data from documents and interviews to best represent the realities of the test developer (triangulation technique).
2. Asking interviewees to verify the researcher's interpretation and representation of their reality (member checks).

The use of official documents (i.e., test questions) and purposive sampling of key informants for guided interviews also ensured authoritative and credible data.

Nkpa (as cited in Gichuru, 2014) defined questionnaire as a carefully structured instrument for data collection in accordance with specifications of the research questions or hypotheses. Questionnaires were used to illicit responses from teachers in order to answer related research questions. A five-section questionnaire (See Appendix C), made up of mainly closed-ended items, was developed.

Section A of the questionnaire dealt with items involving test preparation practices in the district-mandated testing programme. Item 12 in this section sought for information on content coverage, i.e., students' opportunity to learn the contents and topics covered by the test. Section B covered the administration of the district-mandated test. Information mainly

sought for included testing environment (i.e., items 13, 26 and 27), cheating (i.e., items 14, 15, 17, 23, 24, and 25), test directions (i.e., items 18, 19 and 20), and test anxiety (i.e., items 16, 21 and 22). Items 28 to 33 of this section elicit information on test security.

Section C was concerned with grading of students' test performance, and interpretation of students' performance on the district-mandated test. Items 34 to 42 elicited responses on factors that influence students' grades whereas items 43 to 52 were concerned with factors considered in the interpretation of students' test performance. Section D dealt with practices concerning reporting and uses of district test results. Items 53 to 59 drew responses on teachers' reporting practices while items 60 to 71 extracted information on uses made of students' results in the external examination. Finally, Section E sought for information on respondents' background, specifically, teachers' level of training in educational measurement (i.e., item 72).

I developed the questionnaire described above after reviewing the related literature on fair testing practices. Items on the questionnaire were categorically scored, and multiple-scored on a four-point Likert type scale. My thesis supervisors at the University of Cape Coast, who are experts in educational measurement and research methods helped establish the validity (i.e., content and construct validity) of the questionnaire. A pilot test was also carried out in Public Basic Schools within the Cape Coast Municipality that administer assessment materials prepared by CePME. Following the pilot test, errors identified on the instrument were corrected and the final instrument made.

The reliability (i.e., internal consistency) of the questionnaire was estimated using Cronbach's co-efficient alpha. According to Nitko (2001, p. 69), "because co-efficient alpha is a more general version of KR20, it can be used with either dichotomously or polytomously scored items." The questionnaire had a Cronbach's alpha of 0.82 as an estimate of its reliability coefficient (See Appendix D). According to Pavet, Diener, Colvin and Sandvick (as cited in Anhwere, 2009), any co-efficient alpha above 0.70 is considered appropriate.

The questionnaire was preferred at this level due to the large size of respondents that were sampled in order to make generalization statements. This made it unfeasible, in terms of time and funds to interview every respondent (Osuola, 2001). It also reduced chance of researcher bias because the same questions were asked of all respondents. Tabulation of closed-ended responses on the questionnaire were easy and straightforward process. Nevertheless, the items on the questionnaire might not have the same meaning to all respondents. Also, I was unable to probe for additional details from the respondents, using a questionnaire (Gay, Mills, & Airasian, 2009; McMillan & Schumacher, 2001).

Pre-testing of questionnaire

The questionnaire was pre-tested in four JHSs in the Cape Coast Metropolis of the Central Region of Ghana, namely, Apewosika M/A, OLA Presbyterian, Imam Khomeini Islamic, and Kwaprow M/A. The sample for the pre-testing was 40 teachers, comprising 10 teachers from each of the four schools sampled for pre-testing.

Aside the fact that assessment instruments prepared by CePME are

administered in the four schools sampled, these schools also represent the diversity of JHSs in the Ashanti Region that administer assessment instruments prepared by CePME. Apewosika M/A, OLA Presbyterian and Kwaprow M/A JHSs were chosen to represent mainstream schools while Imam Khomeini JHS was chosen, specifically, to represent Islamic schools. Besides, the teachers of these schools had similar qualifications and characteristics with teachers in the accessible population of the study.

The questionnaire was personally administered to the 40 teachers in the four JHSs. A space was provided on the questionnaire for respondents to express their views in writing on the clarity and ambiguity of the items on the questionnaire. The analysis of the teachers' views helped to improve the construct validity of the instrument. For example, 7 items were reworded and 3 items clarified. Again, the pre-testing of the questionnaire provided the opportunity in assessing the instrument's suitability, and also the expediency of the data collection procedure. Specifically, the items on the instrument were reduced from 103 items on the pilot test instrument to 72 items on the final instrument.

Data Collection Procedures

Both primary and secondary data were used for the study. Secondary data were collected through sources such as books, journals, theses and dissertations, and any other relevant literature. Primary data were collected from questionnaires, interview protocols, and documents. Data collection for both qualitative and quantitative methods occurred concurrently. As part of the researcher's ethical considerations, and also by the fact that data collection, especially interviews, requires the researcher to spend an amount of

time with participants, I obtained formal permission from all participants. Again, all participants for the teachers' level were provided with information about the study's aims and purposes (See Appendix C). The respondents for the test developer's level were requested to sign an informed consent form (See Appendix E). Moreover, participants were informed of their right to withdraw, at any stage, from the study.

The interviews were mainly face-to-face, and conducted in the English language. I recorded the majority of the interviews, with permission of the interviewees. The choice of tape recorders ensured dependability of the data. Data on test questions were collected on characteristics such as contrast, type size, leading, spacing, typeface, justification, line length, and blank space. These, according to Thomson et al. (2002), increase the legibility of a test, and its eventual fairness. Moreover, data on rules for writing test items were also collected on the test questions.

The questionnaire was personally administered to the randomly selected teachers in the 50 sampled public JHSs for the study. Permission was sought from head teachers of the respective schools with an introductory letter, specifying the study's topic and purpose. In each school, I further explained the purpose of the study to the sampled teachers, and assured them of anonymity and confidentiality of their participation in the study. For instance, participants were instructed not to identify themselves or their respective schools and districts on the questionnaire. Finally, I provided opportunities for respondents to seek clarification on issues that were not clear to them before responding to the questionnaire. According to Trochim (as cited in Anhwere, 2009), this will help erase respondents' biases and

prejudices.

The data collection process started on May 11, and ended on June 4, 2016, thus, spanning a period of 24 days. Out of the 300 questionnaires administered, 251 representing 84% response rate, were retrieved. It should be noted that I made follow-ups to many of the schools to collect the completed questionnaires.

Data Processing and Analysis

This study employed the multilevel mixed data analysis (Teddlie & Tashakkori, 2009), and involved the utilization of quantitative and qualitative data analysis techniques. Qualitative content analysis was used to subjectively interpret qualitative data collected on the interview protocol and test questions. Content analysis was used to analyse data deductively. In deductive analysis, codes and categories are derived from previous literature (Kavanagh, 2013). Evidence of fair testing practices was then displayed by showing coding categories with exemplars and offering descriptive evidence.

Quantitative data obtained through questionnaires were edited, coded, and entered into the Statistical Product for Service Solution (SPSS) computer software for analysis. Both descriptive and inferential statistics were employed to analyse data, where appropriate. The items on the questionnaire comprised of Likert-scale items, categorical (i.e., dichotomous and multiple-choice) type items, and multiple response items. Items: 1 – 11, 13 – 27, and 53 – 70 were measured on four point Likert scales indicating, "Strongly agree" (scored 4) to "Strongly disagree" (scored 1). Items 28 to 33 were also measured on four point Likert scales indicating, "always" (scored 4) to "never" (scored 1) whereas items 34 to 42 were measured and coded as follows: "very large

extent" (scored 4) to "not at all" (scored 1). Item 12 was scored categorically as "above 90%" (scored 3), "between 60% - 90%" (scored 2), and "below 60%" (scored 1). Items: 43 – 52, and 71 were also measured and coded as "yes" (scored 2) and "no" (scored 1). Finally, item 72 was coded as follows: "trained" (scored 2) and "never trained" (scored 1). It must be emphasized that negatively worded items had coding weights reversed directly for them.

Section F of the teachers' questionnaire was on background information of the respondents. These responses were analysed using frequency and percentage tables.

Research question one

In what ways does a test developer follow the standard approved practices of test fairness?

The contents of test questions, and interviewees' responses to the interview protocol, which represented test developer's practices concerning the development, administration, scoring, interpretation, and reporting of district test results were analysed with qualitative content analysis. Firstly, audio recording of interviews were transcribed verbatim. All transcribed scripts of interview protocol, and test questions were immediately coded using pre-determined coding categories. Themes were used as the unit of analysis (i.e., coding). Thus, codes were assigned to chunk of message of any size as far as they together represent a pre-determined category. Moreover, "fuzzy boundaries between categories" (Zhang, 2004, p. 3) were employed, and therefore, a theme could belong simultaneously to more than one category.

Research question two

In what ways do teachers adhere to fair test preparation practices?

The scores of responses on items 1 to 11 of the questionnaire, which represented actual test preparation practices of teachers, were used in answering this research question. The data on this research question were analysed using one-sample t test. The one-sample t test examined whether the means obtained from teachers' responses to the individual questionnaire items were significantly different from a known mean or test value of 2.5. The test value (i.e., 2.5) is the midpoint of the scores assigned to teachers' responses to the questionnaire items (i.e., "Strongly agree" [scored 4] to "Strongly disagree" [scored 1]). Items with obtained means that are statistically significant and greater than 2.5 were discussed as practices adopted by teachers in preparing students for the test, and vice versa. The level of significance was .05.

Research question three

What significant difference in fair test preparation practices exists between teachers who have received training in educational measurement, and those who have received no training in educational measurement?

The scores of the responses to items 1 to 11 of the questionnaire were used to answer this research question. An independent-samples t test was computed between teachers who have received training in educational measurement and those who have receive no training in educational measurement. The t-test examined whether there was any statistically significant difference in the test preparation practices of trained and untrained teachers. The level of significance was .05.

Research question four

In what ways do teachers adhere to fair test administration practices?

The scores of responses to items 13 to 27 and 28 to 33 of the questionnaire, which represented 15 actual test administration practices of teachers and 6 test security practices in the participating schools respectively, were used in answering this research question. A one-sample t test was used to analyse data on this research question. The test examined whether the means obtained from teachers' responses to the questionnaire items were significantly different from a known mean or test value of 2.5. The test value (i.e., 2.5) is the midpoint of the scores assigned to teachers' responses to the questionnaire items (i.e., "Strongly agree" [scored 4] to "Strongly disagree" [scored 1]). Items with obtained means that are statistically significant and greater than 2.5 were discussed as practices adopted in administering the district test, and vice versa. The level of significance was .05.

Research question five

What significant difference in fair test administration practices exists between teachers who have received training in educational measurement, and those who have received no training in educational measurement?

The scores of the responses to items 13 to 27 of the questionnaire were used to answer this research question. An independent-samples t test for equality of means was computed between teachers who have received training in educational measurement and those who have receive no training in educational measurement. The t-test examined whether there was any statistically significant difference in the test administration practices of trained and untrained teachers. The level of significance was .05.

Research question six

What factors influenced students' grades on the district-mandated test?

One-sample t-test was used to analyse the scores of respondents' responses to items 34 to 42 of the questionnaire, which represented 9 factors that influenced students' grades on the district-mandated test. A one-sample t test was used to test whether or not the observed means for each of the factors influencing students' grades were significantly different from a hypothesized mean of 2.5. The hypothesized mean (i.e., 2.5) is the average of the scores assigned to teachers' responses to the questionnaire items (i.e., "Strongly agree" [scored 4] to "Strongly disagree" [scored 1]). For each item (i.e., factor) that the observed mean was found to be significantly different from and greater than the hypothesized mean of 2.5, a conclusion was drawn that, that factor influenced students' results on the test, and vice versa. The level of significance was .05.

Research question seven

What significant difference exists between teachers who have received training in educational measurement, and those who have received no training in educational measurement in terms of factors influencing students' grades?

This question was answered by analysing the data on items 34 to 42, using an independent-samples t-test for equality of means. The t test compared the group mean scores for teachers who have received training in educational measurement, and those who have receive no training in educational measurement, with respect to factors that influenced students' results on the district test. The test examined whether there was any statistically significant difference in the grading practices of trained and untrained teachers. The level of significance was .05.

Research question eight

What factors are considered by teachers in interpreting students' performance on the district-mandated test?

Using a frequency and percentage distribution table, this research question was answered by analysing the scores of respondents' responses to items 43 to 52 of the questionnaire.

Research question nine

What are teachers' fair reporting practices of district-mandated test results?

One-sample t-test was used to analyse the scores of respondents' responses to items 53 to 59 of the questionnaire, concerning how teachers fairly report students' results on the district-mandated test. The t test was used to test whether or not the obtained means for each of the reporting practices were significantly different from an assumed mean of 2.5. The assumed mean (i.e., 2.5) is the midpoint of the scores assigned to teachers' responses to the questionnaire items (i.e., "Strongly agree" [scored 4] to "Strongly disagree" [scored 1]). For each practice that the obtained mean was found to be significantly different from and greater than the assumed mean of 2.5, a conclusion was drawn that, that practice is utilized in teachers' reporting practices, and vice versa. The level of significance was .05.

Research question ten

What are teachers' fair uses of district-mandated test results?

This question was answered by analysing the data on items 60 to 70, using one-sample t test. The test examined whether the means obtained from teachers' responses to the questionnaire items were significantly different

from a known mean of 2.5. The known mean (i.e., 2.5) is the midpoint of the scores assigned to teachers' responses to the questionnaire items (i.e., "Strongly agree" [scored 4] to "Strongly disagree" [scored 1]). Items with obtained means that were statistically significant and greater than 2.5 were discussed as uses made of students' results, and vice versa. The level of significance was .05. Also, frequency and percentage table was used to analyse respondents' responses to item 71, which enquired whether uses made of students' test results were based only on students' performance on the district-mandated test or other student information were considered.

Chapter Summary

The methodology that was necessary to address the research questions was presented in Chapter Three. I adopted a mixed method approach to the study. This involved the use of document review and semi-structured interview guide, and questionnaire to collect data from test questions and 3 staffs of CePME, and 300 teachers randomly selected from the participating schools. The data collected were analysed using qualitative content analysis, one-sample t test, independent-samples t test, and frequency and percentages.

The sampling procedure for the test developer's level of this study makes it difficult to generalize the results to the whole population of test developers in Ashanti Region. Instead, analytical generalization (i.e., comparing the results to previously developed concepts on test fairness) was made. I cannot however judge the honesty and truthfulness of such responses made by respondents in the interviews or on the questionnaires.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

The study sought to investigate the fair testing practices of a test developer, and teachers in a district-mandated testing programme for Public JHSs in the Ashanti Region of Ghana. A mixed-methods approach to research was adopted in conducting this study. Document review, semi-structured interview guide, and questionnaire were used to collect data in order to answer the research questions.

In this chapter, qualitative data collected at the test developer's strand were analysed using qualitative content analysis while quantitative data obtained from responses on the questionnaire items were analysed using one-sample t-test, independent-samples t-test, and frequency and percentage tables. The results of both qualitative and quantitative analysis are presented with discussions.

Results

Background information

I carried out the study in 50 public JHSs in the Ashanti Region of Ghana that administer assessment materials prepared by CePME, with a sample size of 251 teachers. The 50 JHSs were located in three districts (See Table 2 for the distribution of JHSs according to districts). The number of teacher respondents from each school ranged from three to six (See Appendix A).

Moreover, 3 people, representing CePME, were interviewed in addition to an analysis of data on 9 test questions, including Integrated Science for JHS 1, 2 and 3; Religious and Moral Education for JHS 1, 2 and 3; and Social Studies for JHS 1, 2 and 3. These subjects are assessed in all JHSs across the country. Also, the 3 people selected for interviews had extensive knowledge of the activities at CePME. They have also received training in educational measurement.

Pre-service training in educational measurement

Item 72 of the questionnaire sought to find out whether respondents had received training in educational measurement. However, the responses for in-service training and course/school training were added and taken as one. The result is presented in Table 2.

Table 2- *Frequency Distribution of Respondents who Received Training or Those who did not Receive Training in Educational Measurement*

Status	Frequency	Percentage (%)
Received training in educational measurement.	162	64.5
Did not receive training in educational measurement.	89	35.5
Total	251	100

Source: Field survey, Boakye (2016)

From Table 2, 162 of the total respondents, representing 64.5%, specified that they have received training in educational measurement while the remaining 89, representing 35.5% of the respondents, indicated that they have never received training in educational measurement.

Research question one

In what ways does a test developer follow the standard approved practices of test fairness?

This question sought to find out the test developer's practices that ensured equal opportunity for all students taking the district-mandated test. Three personalities, representing CePME, were selected for interviews. In addition, nine test questions were also analysed. Data collected were analysed using qualitative content analysis. Based on relevant literature, I identified four themes as coding categories, including (1) vigilant efforts not to discriminate against students or groups of students, (2) efforts to establish uniform procedures for using the assessment, (3) efforts to align the test to curriculum and instruction, and (4) efforts to eliminate deficiencies in the test instruments. These themes were then used to code all transcribed interviews' responses, and test questions.

Vigilant efforts not to discriminate against students or groups of students

This theme presents the practices of CePME that ensured that all diversity of students taking the test is considered in the development of the district-mandated test. Interviewee's responses indicated that students' gender, age, geographical locations, ability and opportunity to learn the content of the test were considered to a greater extent. This is what the interviewees had to say:

"Gender is considered through the examples that we give. We make sure they are not biased. So granted am editing a script and someone used only male names for certain items, I will change it. We also make sure they do not stereotype certain gender for certain jobs or tasks."

"We set the test questions to cut across the country. Students in urban and rural areas take the same test so we think about the diverse nature of the locations of the schools."

"We have told our printers that these are questions that children are going to engage in and therefore the printout should be very clear and legible."

Probing further on students' ability and opportunity to learn, one of the interviewee's mentioned that:

"Generally, that's why in developing the test we ensure that the items are varied in terms of difficulty, and covers almost all the topics to be taught in that term. Specifically, we even add a few items of the previous class as the SBA suggest."

"We also give them the scheme of work which in actual fact is the same as the syllabus so that teachers can try and complete that scheme of work before students take the exam."

However, other student characteristics, such as religion, socio-economic status and special needs were less considered as interviewees felt that the nature and contents of the syllabus sometimes makes it difficult to consider such characteristics. This was the concern expressed by one of the interviewees:

"Basically, we use the syllabus, and with the syllabus, some of these stereotypes, for example religion and socio-economic activities like fishing or farming, we can't do anything about them."

Moreover, when asked about the background characteristics that are considered in selecting item writers, one of the interviewees affirms that:

"We do not consider the background characteristics of our item writers. For

good item writing, it does not depend on one's gender or other background characteristics such as religion or region of residence."

The interviewees further shared their views on how they ensured that the content or language used on the test does not offend any student or groups of students taking the test. In the view of the interviewees:

"In addition, when they have finished writing the items, we normally employ item reviewers who look at the language content. They are not experts in writing items but they will look at the language to check for such things as ambiguity and offensive content."

Klinger and Luce-Kapler (2007) stated that test items that respect the diversity of the assessment population are sensitive to test taker characteristics and experiences such as gender and age. According to Redfield (2001), coverage of the full range of student knowledge and skills in the content areas tested will help ensure that students at all levels of learning have an opportunity to demonstrate their knowledge and skills.

Efforts to establish uniform procedures for using the assessment

Interviewees in this study also made mention of practices that ensure that test takers have comparable context in which to demonstrate the knowledge and skills assessed on the district-mandated test. The interviewees disclosed that:

"The first thing that we do is that we actually ensure that all the schools will take the exams on the same day, and at the same time. So we develop our own timetable and insist that the participating schools strictly adhere to it."

"We give them marking scheme for each class. The marking scheme is prepared by the item writers, and is prepared at the same time that they write

the test items."

The interviewees further shared their views on the security of the testing process. All the interviewees expressed that employees of CePME were briefed about the purposes and confidentiality of the activities undertaken by the company, and therefore, there is a consensus effort to maintain the integrity of the company's activities. Similarly, the interviewees shared their views on specific measures that ensure security of test items during item writing and transportation of test instruments. The interviewees expressed that:

"But in actual fact most of them would not be aware when the items written will be used because they have written more than one test. In addition, we have a secure bank of test items and we blend the items that are currently developed with those that we already have in the item bank. This makes it difficult for any item writer to predict the content of any test."

"What we do is that we package the materials school-by-school and subject-by-subject for each district. We package them very well that people cannot open and take materials out. Also, we have a security post that ensures that any package which is taken out is taken out by the person who is responsible to take the materials out so that in case of any leakage we can trace the source."

However, one of the interviewees expressed concerns about test security. In the view of the interviewee:

"When the materials leave our premises, how it is handled in such a way that its security is not compromised is our biggest challenge."

For this reason, the test materials for all subjects were delivered to the

participating schools on the last working day (mostly Fridays) preceding the examination week. However, for districts that the schools are clustered, they were able to curtail security issues by housing the test materials at the district office or in a particular selected school. The materials were then released to the various participating schools as at when (i.e., the day) they are supposed to be administered.

A review of the instructions on the test questions showed that the test developer also specified adequate instructions for taking the test. Test directions were specific as to the regulations governing the test. The instruction read:

*"This booklet consists of two papers. Answer Paper **2**, which comes first in your answer booklet, and Paper **1** on your Objective Test answer sheet. Paper 2 will last **1** hour, **15** minutes after which the answer booklet will be collected. Do **not** start Paper **1** until you are told to do so. Paper **1** will last **45** minutes.*

Instructions also included weights (marks) assigned to parts of the test, and all constructed-response items. Instructions for the Objective Test encouraged students to complete all questions, without any penalty for guessing. According to Downing (2004), students should be encouraged to complete all written exercises and questions and not leave anything blank.

Efforts to align the test to curriculum and instruction

The interviewees mentioned practices that ensured that students of participating schools have had exposure to instruction, knowledge and information that afford them the opportunity to learn contents covered by the test. These efforts were evident mainly in the development of a scheme of work, which is based on the GES approved syllabus. The scheme of work is

shared with all participating schools at the beginning of an academic term, and then based on the scheme of work; a test specification is developed for item writing. In the views of the interviewees concerning content coverage:

"The item writers write items on all of the topics to be covered in a term. Moreover, the items on the test are distributed for all topics that are treated in a particular term. Generally, we make sure every topic is considered in assembling the test items"

Interviewees further expressed that the scheme of work also enables teachers to inform students about the contents covered by the test since the scheme of work puts skills and knowledge into academic terms. Also, the purpose of the study, which is to motivate teachers to cover a higher percentage of the syllabus, ensures that students are assessed on contents that have been taught. In the view of one of the interviewees:

"If we could get our teachers to cover a higher percentage of the syllabus, then they also need an external assessment. Because external assessment questions are based on the GES curriculum, if pupils are assessed externally for terminal exams, then the tendency for the teachers to cover as much of the syllabus as possible will be high."

Efforts to eliminate deficiencies in test items

The interviewees made mention of practices that help to eliminate deficiencies in the test items. One of the interviewees expressed that:

"The items are written by experienced teachers from Colleges of Education and Basic Schools. At the beginning of the conference item writing, an expert in measurement and evaluation walk them through the principles of item writing."

A critical review of the 9 test questions also showed that the test developer adheres to majority of the guidelines or rules for developing test items. On multiple-choice items, rules such as the following were completely observed by the test developer:

1. *Base each item on a specific problem stated clearly in the stem.*
2. *Include as much of the item as possible in the stem.*
3. *Avoid changing pages in the middle of an item.*
4. *Avoid using unnecessarily difficult vocabulary.*
5. *State the stem in positive form (in general). However, whenever negative wording is used in the stem, emphasize it.*
6. *Keep the alternatives homogeneous in content.*
7. *Keep the alternatives parallel in form.*
8. *Keep the alternatives similar in length.*
9. *Avoid the use of specific determiners.*
10. *Use plausible distracters.*
11. *Avoid the alternatives "all of the above" and "none of the above" (in general).*
12. *Be consistent in the number of options used.*

There was not a single violation of the above listed rules. However, rules such as (1) use proper grammar, punctuation, and spelling, (2) use a vertical format for presenting alternatives, (3) the expected response should not be put at the beginning of the stem, and (4) the alternatives must be in alphabetical/sequential order, were to some extent violated. For example, Items 18 and 39 on the objective test for Integrated Science for JHS 3 and Social Studies for JHS 1 read:

18. Which of the following structures develops in to[*sic*] *the fruit after fertilization?*

A. Ovary *B. Ovale*
C. Stigma *D. Stamen*

39. ……. was a great powerful Ga King?

A. Okai Kwei
B. Ayitey Quaye
C. Borketey Okai
D. Anyetei Okran

Violations of grammar, punctuation and spelling rule is particularly unexpected as one of the interviewees mentioned that proofreading is done on the printed hard copies of test questions. This is because typographical errors, which are missed at all other quality control stages can easily be identified when proofreading is done on printed hardcopies (Downing, 2004).

On the other hand, rules for writing essay items such as allowing adequate time to answer the question, providing helpful instructions and guidance, focusing questions on a single issue, and using several relatively short essay questions were observed to a greater extent. On average, 15-20 minutes' test time per each constructed-response item is deemed adequate (Nitko, 2004). However, the frequent use of optional questions and illustrative verbs such as 'state' and 'list' violates two important rules for writing essay items. A typical example of such violations is item 3(b) on the essay test for RME, which read:

3. b. Complete this proverb of Solomon. "The wise store up knowledge" ……

Moreover, recommendations for increasing the legibility of test instruments were observed by the test developer to a greater extent. These included (1) black type on off-white paper, (2) use of serif fonts typeface, (3)

use of 12-point font size, (4) left justified margins, (5) line spacing of more than 1.0, (6) text printed completely in normal mixed-case text, (7) lines of text roughly less than 12 words per line, and (8) 50/50 ratio of test and graphics compared with blank space on the test questions.

Research question two

In what ways do teachers adhere to fair test preparation practices?

Research question 2 sought to find out the kind of practices that teachers utilize in the preparation of students for the district-mandated test. Items 1 to 11 of the questionnaire were used in answering this question. The items were scored, using a four point Likert scale, as "strongly agree" (scored 4) to "strongly disagree" (scored 1).

A one-sample t-test was used to examine whether the means obtained from teachers' responses to the questionnaire items were significantly different from a known mean of 2.5. The known mean of 2.5 is the midpoint of the scores assigned to teachers' responses to the questionnaire items. Therefore, items with obtained means that are statistically significant and greater than 2.5 were discussed as practices adopted by teachers in preparing students, and vice versa. The results are shown in Table 3.

Table 3- *One-sample t-test Statistics on Test Preparation Practices*

Test Preparation Practice	N	Mean	SD	T	df	Sig. (2-tailed)
I motivate students to do their best	251	3.56	.599	28.068	250	.000
I notify students in advance of testing	251	3.50	.695	22.842	250	.000
I inform students about actions that constitute misconduct, and consequences of such acts	251	3.37	.750	18.462	250	.000
I inform students about the purpose, and uses of test results	251	3.31	.686	18.715	250	.000
I inform students about the coverage of the test	251	3.23	.689	16.816	250	.000
I inform students about how their performance will be graded	251	3.23	.825	13.967	250	.000
I inform students about the directions for taking the test	251	3.10	.715	13.203	250	.000
I inform students about the format of the test questions	251	3.08	.823	11.239	250	.000
I inform students about appropriate test taking strategies/skills	251	3.03	.812	10.301	250	.000
I modify previous test questions, and practice it with students	251	2.99	.772	10.017	250	.000
I take previous district test and give them as practice in class	251	2.91	.827	7.825	250	.000

Test value = 2.5

Source: Field survey, Boakye (2016)

From Table 3, the result of the one-sample t test indicated that all the obtained means on teachers' test preparation practices are statistically and significantly different from a known mean of 2.5 (i.e., all p-values are less than .05). Teachers therefore agreed to the following practices in preparing students for the district test: "I motivate students to do their best" (*M* = 3.56, *SD* = .599); "I notify students in advance of testing" (*M* = 3.50, *SD* = .695); "I inform students about actions that constitute misconduct, and consequences of such acts" (*M* = 3.37, *SD* = .750); "I inform students about the purpose, and uses of test results" (*M* = 3.31, *SD* = .686); "I inform students about the coverage of the test" (*M* = 3.23, *SD* = .689); "I inform students about how their performance will be graded" (*M* = 3.23, *SD* = .825); "I inform students about the directions for taking the test" (*M* = 3.10, *SD* = .715); "I inform students about the format of the test questions" (*M* = 3.08, *SD* = .823); and "I inform students about appropriate test taking strategies/skills" (*M* = 3.03, *SD* = .812).

Other test preparation practices including "I modify previous test questions, and practice it with students" (*M* = 2.99, *SD* = .772) and "I take previous district test and give them as practice in class" (*M* = 2.91, *SD* = .827) were also agreed to as practices adopted by teachers in this study.

Research question three

What significant differences in fair test preparation practices exist between teachers who have received training in educational measurement, and those who have received no training in educational measurement?

This research question sought to find out whether there was any significant difference in test preparation practices between teachers who have

received training in educational measurement and those who have receive no training in educational measurement. To answer this question, the responses to items 1 to 11 on test preparation practices were used. Using teachers' level of training as independent variable and test preparation practices as dependent variable, an independent-samples t-test of equality of means was conducted to determine whether teachers' level of training had any statistically significant influence on their test preparation practices. The result is shown in Table 4.

Table 4- *Independent-samples t-test Statistics on Test Preparation Practices of Teachers Trained, and Those Not Trained in Educational Measurement*

Group	N	Mean	SD	t	Df	p-value
Trained	162	35.17	3.869			
				-.688	249	.492
Not Trained	89	35.57	5.412			

Source: Field survey, Boakye (2016)

The result indicated that based on the responses of the teachers, the independent-samples t-test of equality of means is not statistically significant (t (249) = -.688, p > 0.05). This implies that there was no statistically significant difference between teachers who have received training in educational measurement and those who have received no training in educational measurement in terms of test preparation practices.

Research question four

In what ways do teachers adhere to fair test administration practices?

This question sought to investigate two major issues involved in the administration of district-mandated test. These are actual test administration practices, and practices that ensure the security of test materials, before,

during and after test administration. Items 13 to 27 of the questionnaire were used to determine the actual practices adopted by teachers in administering the test. The responses were on a four-point Likert scale with categories from "strongly agree" (scored 4) to "strongly disagree" (scored 1).

A one-sample t test was used to analyse data on this research question. The test examined whether the means obtained from teachers' responses to the individual questionnaire items were significantly different from a known mean of 2.5. The known mean of 2.5 is the midpoint of the scores assigned to teachers' responses to the questionnaire items. Items with obtained means that were statistically significant and greater than 2.5 were discussed as practices adopted by teachers in administering the test, and vice versa. The results are presented in Table 5.

Table 5- *One-sample T Test Statistics on Test Administration Practices*

Test Administration Practices	N	Mean	SD	t	df	Sig. (2-tailed)
I ensure that the testing environment is free from distractions	251	3.58	.584	29.241	250	.000
I assign sitting/seating arrangement for students	251	3.41	.745	19.434	250	.000
I make announcements about the test time at regular intervals	251	3.41	.786	18.263	250	.000
Students start the test promptly, and stop on time	251	3.37	.647	21.327	250	.000
I take note of any problem or irregularities encountered	251	3.21	.725	15.464	250	.000
The sitting arrangement prevents students from copying each other's work	251	3.16	.772	13.452	250	.000
I, single-handed, supervise more than 25 students in one room	251	3.15	.833	12.311	250	.000
I ensure that students are working in the appropriate sections of the test	251	3.06	.730	12.062	250	.000
Ambiguous test questions are clarified only to students who ask about them	251	2.93	.912	7.511	250	.000
I make special adjustments or changes for students with special needs/injuries	251	2.88	1.011	6.023	250	.000
Students are threatened dire consequences if they fail	251	2.84	1.063	5.019	250	.000
I perform other unrelated professional duties while administering the test	251	2.79	.998	4.646	250	.000
I make remarks about the quality or quantity of students' work during testing	251	2.44	.984	-.930	250	.353
I inform students about how past students of specific gender performed on the test	251	2.43	.958	-1.219	250	.224
Students are told to work faster in order to finish on time	251	1.88	.830	-11.819	250	.000

Test value = 2.5

Source: Field survey, Boakye (2016)

From Table 5, the result of the one-sample t test, based on the responses of the teachers, indicated that 13 out of 15 obtained means on teachers' test administration practices are statistically and significantly different from a known mean of 2.5 (i.e., p-values are less than .05). The teachers therefore agreed to practices that; eliminate distractions or barriers to students' maximum performance, ensured uniform directions for all students taking the test, and prevent cheating among students. These practices included "I ensure that the testing environment is free from distractions" (*M* = 3.58, *SD* = .584); "I assign sitting/seating arrangement for students" (*M* = 3.41, *SD* = .745); "I make announcements about the test time at regular intervals" (*M* = 3.41, *SD* = .786); "Students start the test promptly, and stop on time" (*M* = 3.37, *SD* = .647); "I take note of any problem or irregularities encountered" (*M* = 3.21, *SD* = .725); "The sitting arrangement prevents students from copying each other's work" (*M* = 3.16, *SD* = .772); "I ensure that students are working in the appropriate sections of the test" (*M* = 3.06, *SD* = .730); and "I make special adjustments or changes for students with special needs/injuries" (*M* = 2.88, *SD* = 1.011).

The other test administration practices that were agreed to by the teachers included "I, single-handed, supervise more than 25 students in one room" (*M* = 3.15, *SD* = .833); "Ambiguous test questions are clarified only to students who ask about them" (*M* = 2.93, *SD* = .912); "Students are threatened dire consequences if they fail" (*M* = 2.84, *SD* = 1.063); and "I perform other unrelated professional duties while administering the test" (*M* = 2.79, *SD* = .998). The only test administration practice that was disagreed to by the teachers is "Students are

told to work faster in order to finish on time" (*M* = 1.88, *SD* = .830).

Items 28 to 33 of the questionnaire sought to find out the frequency of test security practices in the various participating schools. The items contained statements to be responded to by using a four-point Likert scale with categories from "always" (scored 4) to "never" (scored 1). A one-sample t test was used to examine whether the obtained means for the six individual test security items on the questionnaire were significantly different from a known mean of 2.5. Items with obtained means that were statistically significant and greater than 2.5, as shown in Table 6, were discussed as security practices adhered to in the schools sampled for this study.

Table 6- *One-sample t-test Statistics on Test Security Practices*

Test Security Practice	N	Mean	SD	T	Df	Sig. (2-tailed)
Prior to testing, the test materials are kept in a secure place where students cannot have access to it.	251	3.76	.570	35.179	250	.000
Prior to testing, the test materials are kept in a secure place where unauthorized teachers cannot have access to it.	251	3.63	.712	25.038	250	.000
The tests are administered during the official scheduled testing time.	251	3.60	.676	25.827	250	.000
Duplications of test or answer booklets, including photographing, photocopying, or copying by hand is allowed.	251	1.77	.952	-12.166	250	.000
Teachers discuss the test questions prior to testing.	251	1.52	.797	-19.530	250	.000
Teachers leave testing materials unattended to, before, during or after test administration.	251	1.41	.797	-21.665	250	.000

Test value = 2.5

Source: Field survey, Boakye (2016)

From Table 6, based on the teachers' responses to the items, the one-sample t test indicated that all the test security items are statistically significant (i.e., all p-values are less than .05). Therefore, the following test security measures were implemented in the participating schools: "Prior to testing, the test materials are kept in a secure place where students cannot have access to it" (M = 3.76, SD = .570); "Prior to testing, the test materials are kept in a secure place where unauthorized teachers cannot have access to it" (M = 3.63, SD = .712); and "The tests are administered during the official scheduled testing time" (M = 3.60, SD = .676).

Other practices that violate test security protocols including "Duplications of test or answer booklets, including photographing, photocopying, or copying by hand is allowed" (M = 1.77, SD = .952); "Teachers discuss the test questions prior to testing" (M = 1.52, SD = .797); and "Teachers leave testing materials unattended to, before, during or after test administration" (M = 1.41, SD = .797), were never practiced in the participating schools.

Research question five

What significant differences in fair test administration practices exist between teachers who have received training in educational measurement, and those who have received no training in educational measurement?

This research question sought to find out whether there was any significant difference in actual test administration practices between teachers who have received training in educational measurement and those who have received no training in educational measurement. To answer this question, the responses to

items 13 to 27 on actual test administration practices were used.

Using teachers' level of training as independent variable and actual test administration practices as dependent variable, an independent-samples t-test of equality of means was conducted to determine whether there were any significant differences in the means of trained and untrained teachers in terms of actual test administration practices. The result is shown in Table 7.

Table 7- *Independent-samples t-test Statistics on Test Administration Practices of Teachers Trained, and Those not Trained in Educational Measurement*

Group	N	Mean	SD	t	df	p-value
Trained	162	43.73	4.815			
				-.719	249	.473
Not Trained	89	44.24	6.207			

Source: Field survey, Boakye (2016)

The result indicated that based on the responses of the teachers, the independent-samples t-test for equality of means is not statistically significant (t (249) = -.719, p > .473). This indicates that there was no statistically significant difference between teachers who have received training in educational measurement and those who have received no training in educational measurement in terms of actual test administration practices.

Research question six

What factors influenced students' grades on the district-mandated test?

Items 34 to 42 of the questionnaire asked teachers to indicate the extent to which students' grades on the district test are influenced by factors that are irrelevant to classroom achievement testing. The responses were assessed using a

four-point Likert scale with categories from “very large extent” (scored 4) to “not at al” (scored 1).

A one-sample t-test was used to test whether or not the observed means for each of the factors influencing students’ grades were significantly different from a hypothesized mean of 2.5. The hypothesized mean of 2.5 is the average of the scores assigned to teachers’ responses to the questionnaire items. For each item that the observed mean was found to be significantly different from and greater than the hypothesized mean of 2.5, a conclusion was drawn that, that factor influenced students’ grades on the district-mandated test, and vice versa. The results are presented in Table 8.

Table 8- *One-sample t-test Statistics on Factors that Influenced Students' Grades on the District-mandated Test*

Factors That Affect Students' Scores	N	Mean	SD	t	Df	Sig. (2-tailed)
Students' efforts to learn	251	3.15	.916	11.203	250	.000
Student's class attendance	251	3.06	.951	9.325	250	.000
The language student speaks in class/school	251	3.03	.901	9.286	250	.000
Student's participation in class discussion	251	2.96	.914	8.049	250	.000
Student's relationship with the teacher	251	2.53	1.059	.507	250	.613
Student involvement in religious activities in class/school	251	2.51	1.089	.087	250	.931
Student's moral virtues such as indecent language and dress code	251	2.39	1.020	-1.640	250	.102
Student gender	251	2.25	1.038	-3.802	250	.000
Bonus marks for extra work or responses on the district test	251	2.11	.890	-6.984	250	.000

Test value = 2.5

Source: Field survey, Boakye (2016)

From Table 8 above, the one-sample t test indicated that six out of nine items were statistically significant (i.e., p-values are less than .05). Therefore, factors that were considered as influencing students' grades were as follows: "Student's efforts to learn" (*M* = 3.15, *SD* = .916); "Student's class attendance" (*M* = 3.06, *SD* = .951); "The language student speaks in class/school" (*M* = 3.03, *SD* = .901); and "Student's participation in class discussion" (*M* = 2.96, *SD* = .914). Factors that had no influence on students' grades were "Student gender" (*M* = 2.25, *SD* = 1.038), and "Bonus marks for extra work or responses on the district test" (*M* = 2.11, *SD* = .890).

Research question seven

What significant differences exist between teachers who have received training in educational measurement, and those who have received no training in educational measurement in terms of factors influencing students' grades?

This research question also sought to find out whether there was any significant difference in the means of teachers who have received training and those who have received no training in educational measurement in relation to factors that influenced students' grades on the district-mandated test. The result is shown in Table 9.

Table 9- *Independent-samples t-test Statistics on Factors that Affect Students' Grades on the District-mandated Test*

Group	N	Mean	SD	t	df	p-value
Trained	162	23.93	5.040	-.304	202.13	.762
Not Trained	89	24.11	4.422			

Source: Field survey, Boakye (2016)

Degrees of freedom reduced because Levene's test shows violation of homogeneity of variances assumption.

The independent-samples t-test for equality of means shows no statistically significant difference, (t (202.13) = -.304, p > 0.05). This implies that there was no statistically significant difference between teachers who have received training in educational measurement and those who have received no training in educational measurement in terms of factors that influenced students' grades on the district-mandated test.

Research question eight

What factors are considered by teachers in interpreting students' performance on the district-mandated test?

Three groups of factors can impact the valid and useful interpretations of students' performance on a district-mandated test, namely; psychometric factors, test taker factors, and contextual factors (JCTP, 1993). Items 43 to 52 of the questionnaire sought to find out which of these factors are considered by teachers in the interpretation of students' performance on the district-mandated test. The responses were categorized as "yes" (scored 2) and "no" (scored 1). The

distribution of the teachers' responses is shown in Table 10.

Table 10- *Frequency Distribution of Factors Considered in Interpreting Students' Performance on the District-mandated Test*

Factors that Introduce CIV	Yes (%)	No (%)	Total (%)
The level of vocabulary, and sentence structure of the test.	73.7	26.3	100
Irregularities or problems encountered during testing.	69.7	30.3	100
The difficulty level of questions.	69.3	30.7	100
Student's ability.	68.9	31.1	100
Sufficiency or adequacy of test time.	67.7	32.3	100
The alignment of the test content to curriculum and instruction.	63.7	36.3	100
Student's opportunity to learn the content of the test.	59.4	40.6	100
Student's motivation to perform.	55.8	44.2	100
Student's test-taking skills or strategies.	52.2	47.8	100
Student's socio-economic background.	49.8	50.2	100

Source: Field survey, Boakye (2016)

Overall, the responses of the teachers indicated that psychometric factors are considered by most of the teachers in the interpretation of students' performance. These included "The level of vocabulary, and sentence structure of the test" (73.7%); "The difficulty level of the questions on the test" (69.3%); and "Sufficiency or adequacy of test time" (67.7%). In addition, contextual factors such as "Irregularities or problems encountered during testing" (69.7%); "The alignment of the test content to curriculum and instruction" (63.7%); and "Student's opportunity to learn the content of the test" (59.4%) were also

considered by most of the teachers in the interpretation of students' performance on the test.

From Table 10, the least factors considered by teachers were that of test taker characteristics, which included "Student's ability" (68.9%); "Student's motivation to perform on the test" (55.8%); "Student's test-taking skills or strategies" (52.2%); and "Student's socio-economic background" (49.8%). Therefore, it could be said that teachers in this study did not consider, to a very large extent, test taker characteristics in the interpretation of students' results.

Research question nine

What are teachers' fair reporting practices of district-mandated test results?

Items 53 to 59 of the questionnaire asked teachers to indicate the extent to which they agreed to practices used in reporting students' results on the district-mandated test. The responses were assessed using a four-point Likert scale with categories from "strongly agree" (scored 4) to "strongly disagree" (scored 1). One-sample t-test was used to analyse the scores of teachers' responses to these items. The t-test was used to test whether or not the obtained means for each of the reporting practices were significantly different from an assumed mean of 2.5. The assumed mean of 2.5 is the midpoint of the scores assigned to teachers' responses to the questionnaire items. For each practice that the obtained mean was found to be significantly different from and greater than the assumed mean of 2.5, a conclusion was drawn that, that practice was utilized in teachers' reporting practices, and vice versa. The results are presented in Table 11.

Table 11- *One-sample t-test Statistics on Fair Reporting Practices*

Reporting practices	N	Mean	SD	t	Df	Sig. (2-tailed)
I use clear and simple language on report cards.	251	3.70	.477	39.727	250	.000
I discuss with students ways of improving their achievement.	251	3.40	.705	20.197	250	.000
I report district test results in a timely fashion.	251	3.26	.639	18.809	250	.000
I report separately, on report cards, student's achievement, effort and attitude.	251	3.12	.818	12.080	250	.000
I share specific examples of students' strength and weaknesses, on report cards.	251	3.04	.809	10.573	250	.000
I modify reporting procedures for students with special needs and/or injuries.	251	2.76	.920	4.493	250	.000
I publicly display students' test results with visible grades.	251	2.33	.925	-2.832	250	.005

Test value = 2.5

Source: Field survey, Boakye (2016)

From Table 11, based on teachers' responses to the seven items on reporting practices, the one-sample t-test indicated that all the items were statistically significant (i.e., p-values are less than .05). Hence, teachers' responses indicated that the following reporting practices were adopted by teachers in the participating schools: "I use clear and simple language on report cards" (*M* = 3.70, *SD* = .477); "I discuss with students ways of improving their achievement" (*M* = 3.40, *SD* = .705); "I report district test results in a timely fashion" (*M* = 3.26, *SD* = .639); "I report separately, on report cards, student's achievement, effort and attitude" (*M* = 3.12, *SD* = .818); "I share specific examples of students' strength and weaknesses, on report cards" (*M* = 3.04, *SD* = .809); and "I modify reporting procedures for students with special needs and/or injuries" (*M* = 2.76, *SD* = .920).

The practice, "I publicly display students' test results with visible grades" (*M* = 2.33, *SD* = .925), according to teachers' responses, was disagreed to as a practice adopted by teachers in reporting students' test results.

Research question ten

What are teachers' fair uses of district-mandated test results?

This question sought to investigate two major issues involved in the uses made of district-mandated test results. Items 60 to 70 of the questionnaire were used to determine the actual uses made of district-mandated test results by teachers. These included both high-stakes and low-stakes uses of students' results. The responses were on a four-point Likert scale with categories from "strongly agree" (scored 4) to "strongly disagree" (scored 1).

One-sample t-test was used in analysing the data on these items. The test examined whether the means obtained from teachers' responses to the individual questionnaire items are significantly different from a known mean of 2.5. The known mean of 2.5 is the midpoint of the scores assigned to teachers' responses to the questionnaire items. Items with obtained means that are statistically significant and greater than 2.5 were considered as uses made of students' results, and vice versa. The results are shown in Table 12.

Table 12- *One-sample t-test Statistics on Uses Made of District-mandated Test Results*

Uses Made of Test Results	N	Mean	SD	t	df	Sig. (2-tailed)
Evaluate students' progress	251	3.41	.729	19.867	250	.000
Give feedback to students	251	3.33	.732	18.069	250	.000
Rank students by achievement level	251	3.27	.753	16.230	250	.000
Give feedback to parents	251	3.21	.726	15.517	250	.000
Identify students in need of extra support	251	3.17	.787	13.428	250	.000
Determine students grades	251	3.14	.822	12.404	250	.000
Assess my teaching effectiveness	251	3.09	.830	11.222	250	.000
Diagnose students learning needs	251	3.06	.825	10.751	250	.000
Make classroom instructional decisions	251	3.03	.867	9.724	250	.000
Promote/retain students in class	251	2.93	.899	7.622	250	.000
Group students by achievement level	251	2.90	.866	7.256	250	.000

Test value = 2.5

Source: Field survey, Boakye (2016)

From Table 12, the results of the one-sample t-test indicated that all 11 items on uses made of students' test results were statistically significant (i.e., all p-values are less than .05). Hence, the teachers responses to the questionnaire items indicated that the following uses were made of district-mandated test results: "Evaluate students' progress" (*M* = 3.41, *SD* = .729); "Give feedback to students" (*M* = 3.33, *SD* = .732); "Rank students by achievement level" (*M* = 3.27, *SD* = .753); "Give feedback to parents" (*M* = 3.21, *SD* = .726); "Identify students in need of extra support" (*M* = 3.17, *SD* = .787); "Determine students grades" (*M* = 3.14, *SD* = .822); "Assess my teaching effectiveness" (*M* = 3.09, *SD* = .830); "Diagnose students learning needs" (*M* = 3.06, *SD* = .825); and "Make classroom instructional decisions" (*M* = 3.03, *SD* = .867).

Other uses made of students' test results included "Promote/retain students in class" (*M* = 2.93, *SD* = .899) and "Group students by achievement level" (*M* = 2.90, *SD* = .866).

Item 71 on the questionnaire asked respondents to indicate whether uses made of students' district-mandated test results were based only on students' performance on the test or other student information were considered. The responses were categorized as "yes" (scored 2) and "no" (scored 1). Table 13 shows the data on this item.

Table 13- *Frequency Distribution of Procedure Used in Making Decisions about Students*

Decisions About Students are Based Only on District Test Results	Frequency	Percentage (%)
Yes	100	39.8
No	151	60.2
Total	251	100

Source: Field survey, Boakye (2016)

From Table 13, it could be observed that 100 (39.8%) of the respondents made decisions about students based on only district-mandated test results. The rest, 151 (60.2%) considered other students' information in making uses of district-mandated test results.

Other results

Item 12 of the questionnaire sought to find out the proportion of the syllabus for an academic term that teachers were able to complete. The responses to the item were scored categorically as "more than 90%" (scored 3); "between 60 and 90 percent" (scored 2); and "Less than 60%" (scored 1). Frequency distributions have been used in analysing item 12, as shown in Table 14.

Table 14- *Frequency Distribution of the Percentage of the Syllabus Completed in an Academic Term*

Percentage of the Syllabus	Frequency	Percentage (%)
More than 90%	171	64.5
Between 60% and 90%	68	27.1
Less than 60%	21	8.4
Total	251	100

Source: Field survey, Boakye (2016)

The results from Table 14 indicated that, out of the 251 teachers for the study, 162 (64.5%) reported that they were able to complete more than 90% of the syllabus for an academic term while 68 (27.1%) and 21 (8.4%) of the teachers indicated that they were able to complete between 60 and 90 percent, and less than 60% of the syllabus respectively.

Discussion of Results

This section presents the discussion of the results. It involves the discussion of the findings of the study in relation to published literature and empirical findings on test fairness. The diverse nature of students in a district in terms of age, gender, religion, geographic location, and socio-economic background necessitates that the practices of test developers and test users, particularly teachers, are fair to all students taking the mandated test. A prime threat to test fairness comes from aspects of the test or testing process that may produce CIV, and therefore, incorrectly and systematically increase or decrease test scores for some students (Haladyna & Downing, 2004). Carefully following standard practices of testing however help to reduce CIV in students' scores. Based on the objectives of the study, the discussion is outlined as follows:

1. Test developer's fair testing practices.
2. Teachers' fair testing practices in preparing students for the district test.
3. Teachers' fair testing practices in administering the district test.
4. Teachers' fair testing practices in grading students' performance on the test.
5. Fair testing practices of teachers in interpreting students' performance on the test.
6. Fair testing practices of teachers in reporting district-mandated test results.
7. Teachers' fair uses of district-mandated test results.

Test developer's fair testing practices

Factors that cause differences in students' scores, and are not attributable to the construct that the district-mandated test is designed to measure should be considered in achievement testing process. This is because these factors are irrelevant to the construct measured in achievement testing (Bachman, 1990). Test developer's (CePME) consideration of students' characteristics would ensure that the district test is appropriate for all intended test takers irrespective of student's characteristics such as gender, age, and ability, which are irrelevant to the constructs assessed on the district-mandated test. A test that is developed with due consideration of the characteristics of the intended examinees reflect the test takers' actual ability (Klinger & Luce-Kapler, 2007).

Moreover, the distribution of a scheme of work to the participating schools would ensure that students are tested on contents that they have had the opportunity to learn. Students' opportunity to learn the content of the tests was further affirmed by the responses of the teachers, which indicated that a

significant majority of the teachers (64.5%) were able to complete more than 90% of the syllabus for an academic term. In contrary to the findings of CRDD (2005) which reported that the majority of teachers in that study completed only 60% of the contents of the syllabus for basic schools in Ghana, the higher percentage of the syllabus covered by significant majority of the teachers in this study could be explained by the assertion of the test developer that the purpose of the district-mandated test is to motivate teachers to cover a higher percentage of the academic syllabus.

The test developer's effort in ensuring an extensive coverage of the content areas or topics to be tested, through the use of a test specification, would ensure that students at all levels of learning have an opportunity to demonstrate their knowledge and skills on the test. Test specification would also ensure a representative sample of content and cognitive objectives (Suskie, 2000). Construct underrepresentation resulting from under-sampling or biased sampling of the content domain by the assessment instrument (Schouwstra, 2000), which in turn benefits students who have mastered that aspect of the curriculum would be avoided. Construct underrepresentation leads to underperformance on the part of some test takers (Messick, 1989).

An analysis of the test questions revealed that majority of the rules for writing test items (i.e., multiple-choice and essay items) were adhered to by the test developer. In addition, test instructions were sufficient and comprehensive. This would help to diminish construct-irrelevant test variance and yield more accurate test scores. However, a few violations of such rules as witnessed on the

test questions is still problematic as they undermine the quality of individual items and the test as a whole (Downing, 2004; Tarrant & Ware, 2008). For example, violations of grammar and punctuation rules could make reading more difficult and time consuming for some students (Bachman, 1995; Lang & Wilkerson, 2008). Such violations increase unnecessary searching and reading on the part of the test taker (Etsey, 2012). In addition, the use of optional test items would make it difficult for teachers to analyse students' performance on the constructed-response items (Etsey, 2012). Also, frequent use of illustrative verbs (i.e., 'state' and 'list') assess lower-order outcomes such as students' ability to recall knowledge, which should not be evaluated by essay questions (Reiner et al., 2002). Flawed test questions would tend to present more of a passing challenge for students taking the test (Downing, 2002).

Effective item writers are trained, not born (Haladyna, 2004). The training of item writers during item writing workshops would help prevent problems of poor-quality, flawed test questions that test trivial content (Haladyna). Thus, one of the more important validity issues, and the eventual fairness of the district-mandated test concerns the selection and training of item writers (Haladyna, 1999). However, the finding that item writers received training on item writing is inconsistent with the finding that some of the test items violated standard item writing rules. This contradiction, probably, supports the statement that training in assessment methods does not necessarily lead to quality test items (Amedahe, 1989).

The analysis of the interview data further indicated that the test developer employed experts to review the test items in order to detect language and other representations that might be interpreted differently by different groups of students, and for content that might be offensive to some test takers. Such sensitive review panel, probably, helped to guard against construct-irrelevant language that may offend some test takers, and against construct-irrelevant context that may be more familiar to some students than others (AERA et al., 2014). For instance, analysis of the test questions indicated that context or topics that have been identified as likely sources of CIV, such as regionalisms, sports, accidents, illnesses, natural disasters, death and dying, rape, and advocacy (Cole & Zieky, 2001) were completely avoided in the development of the test items. Also, test questions were free of gender, political, and religious bias. There was not a single citation of a political personality, party, organization, or slogan in the test items. Moreover, the three major religions in Ghana were fairly represented by items on the RME tests. Careful editing of test content by experts yield more accurate test scores that reflect students' true performance (ETS, 2009).

The district test can be said to be much more accessible to a wide range of students since the test developer adhered to the principles of universal design. Principles of universal design make tests better for every student (Johnstone et al., 2006), as they enable all students to decipher test items more easily (Thompson et al., 2002). For instance, the test developer's use of black type on off-white paper would reduce glare (Thompson et al. 2002) while the use of serif fonts on educational textbooks would help test takers to better recall on the test items as a

result of students' familiarity with serif fonts. Students would also recall text better as it is left justified and unjustified on the right. Moreover, the markings of serif fonts would make the row of lines on the test questions to be separated more easily; consequently, reading becomes easier for the test takers (Gasser, Boeke, & Haffernan, 2005). Type size has significant effect on reading speed, thus 12-point size would be read faster than others (Chandler, 2001). Universally designed assessments help to advance assessment participation and performance for all test takers, including students with special needs.

A fixed timetable for the administration of the district-mandated test in all participating schools, as indicated by the interviewees, would ensure that all students taking the test are given equitable treatment in the administration of the test. Such standardization practice would curtail problems such as test leakages that give some students an unfair advantage over others. The provision of scoring keys and guides by the test developer, as indicated by the interview data, are also important features of standardization of the district-mandated test. Scoring guides would ensure consistency in the scoring process, and also help to reduce errors resulting from hand scoring of student responses. Scoring rubrics provide feedback to students concerning how to improve their performances (Rudner & Shafer, 2002).

Ideally, item writers for external mandated examinations are given an assignment to produce a small number of items. This ensures that the security of the testing programme is maintained (Tierney, 2013). Nonetheless, for the fact that item writers for this examination write two or three sets of test questions for

specific subjects somewhat ensured that test questions were not leaked to participating schools before they were administered. This is because it would be difficult for any item writer to recall specific test items that would appear on a particular test. Moreover, the security of the test content would be maximized with the use of the secure bank of test items, stored in a library with highly restricted access, as indicated by the interviewees. To the extent that item writers were unaware which specific items would appear on the district-mandated test, the security of the test would be assured. In addition to the above, the packaging of the test materials school-by-school and subject-by-subject, as indicated by the interviewees, would reduce problems of test insecurity. Teachers and school authorities, prior to testing, would not need to do sorting of test materials, which threatens test security.

The development of a scheme of work by the test developer would ensure that there is a consensus between the test developer and teachers as to what students should be able to do at the end of an academic term (i.e., learning objectives), what should be taught in that academic term, and the content that would be assessed at the end of the academic term. Thus, it can be said that the district test is aligned to the curriculum (Briggs, 2009), and therefore, assesses what should be taught in the participating schools. Items on the district-mandated test would reflect a pre-established set of content standards that specify the knowledge and skills students are expected to acquire as a function of schooling. According to Popham (2002), if a criterion-referenced examination has two main

characteristics, i.e., explicit test specifications and congruent test items, then a more accurate interpretation of an examinee's performance can be determined.

Teachers' fair testing practices in preparing students for the district test

Items 1 to 11 of the questionnaire sought to investigate the test preparation practices of teachers in the participating schools. Analysis of the survey data indicated that teachers agreed to all the test preparation activities stipulated on the questionnaire. However, nine out of the 11 test preparation practices were considered ethical since they give students an opportunity to learn and be informed about the content of the district-mandated test. These included practices such as motivating students to do their best; notifying students in advance of testing; and informing students about how their performance will be graded. The other ethical preparation practices adopted by the teachers included informing students about, actions that constitute misconduct and its consequences; purpose and uses of test results; coverage of the test; directions for taking the test; format of test questions; and appropriate test taking strategies. This result supports previous findings that teachers in the Ashanti Region of Ghana adopt ethical test preparation practices (Oduro-Okyireh, 2008).

Informing test takers about appropriate test-taking strategies would reduce test anxiety and improve students' attitudes toward the test, and thereby, improves students' chances of showing their actual knowledge on the content of the district test (Dodeen, 2015; Maxwell et al., 2012). McCabe et al. (2006) advocated communicating clear expectations for honest behaviour, and communicating clear consequences for dishonest behaviour in order to ensure test security and

equitable treatment of all test takers. Making students aware of the purposes of the mandated testing programme would improve both their attitude and motivation toward the testing programme (JCSEE, 2003). Informing students, prior to the use of an assessment method, about the scoring procedures to be followed would also help to ensure that both students and their teachers hold similar expectations (Camara, 2007). Research findings showed that being well grounded in the test content and form can improve test performance (ISBE, 2014). Ethical test preparation would ensure that students are tested for their knowledge and ability, not their test-taking skills (ISBE, 2014).

If some students have received test preparation, and others have not, differences in the performance of these groups of students might be attributable to the fact that some students did not receive test preparation. However, the results of the independent-samples t test indicated that there was no significant difference in the test preparation practices of trained and untrained teachers in educational measurement. This implies that students of both trained and untrained teachers received similar test preparation. Gipps (1995) noted that there should be some evidence that all students have received uniform and ethical test preparation.

However, the other two preparation practices adopted by the teachers are considered unethical as it amounts to teaching the test. Providing test preparation on actual previous test questions and/or modified previous test questions implies teaching the students nearly identical questions that will appear on a district test, and therefore, are unethical, and creates a biased score. Such practices constitute cheating, and give some students an unfair advantage over others. It also confines

instruction to a mere sample of the knowledge and skill domain represented by the district-mandated test (Redfield, 2001). Teaching the test would request students to cram for the examination rather than prepare for a broad curriculum (Cheng, 1998). This defeats the purpose of the district-mandated testing programme, which is to motivate teachers and students to cover a higher percentage of the curriculum, as indicated by the test developer.

Teachers' fair testing practices in administering the district-mandated test

According to Plake and Jones (2002), the pre-administration procedures for paper-and-pencil tests include verifying the integrity of materials prior to the test date and maintaining their security and confidentiality. An analysis of the questionnaire data indicated that practices that ensured test security such as securing test materials from unauthorized teachers and students, and adhering to the official scheduled testing time, were always done in the participating schools. Other practices that violate test security protocols such as discussing test questions prior to testing, duplicating test materials, and leaving testing materials unattended to, were never practiced in the participating schools. The results support the statement that states and school districts should adopt policies, which clearly forbid school personnel to look at "high stakes" test questions except as needed during administration (Cizek, 1999). Testing should be conducted in a fair and ethical manner, which includes security (Cizek, 2012). Security of the district-mandated test helps to maintain the meaning and integrity of students' test scores (Xiaomei, 2014).

In addition to pre-administration security procedures, practices that reduce cheating during actual test administration improve the security of mandated examination. Practices that eliminate cheating, and were practiced (agreed to) by the teachers include (1) I assign seating/sitting arrangement for students, (2) students start the test promptly, and stop on time, and (3) the sitting arrangement prevents students from copying each other's work. Henning (2012) stated that students should be seated in an arrangement that prevents them from seeing the work of other students. Henning further noted that assigning seating would prevent students from choosing cheating partners.

However, other practices agreed to by the teachers, including 'I, single-handed, supervise more than 25 students in one room' and 'I perform other unrelated professional duties while administering the test', would encourage cheating behaviours among students during test administration. Such test administration practices do not communicate to students that teachers expect honest academic behaviour (Cizek, 1999). To maintain test security, the teachers should grant their full attention to the testing site at all times (Gordon & Fay, 2010).

The district-mandated test is a potentially stressful event due to its high-stakes nature such as grading and promotions that are made of students' test results. Henceforth, practices that further make test takers anxious should be avoided. However, the analysis of the questionnaire data showed mixed results. Responses of the teachers indicated that students are *not* told to work faster in order to finish on time. This would help to eliminate students' test anxiety. But

threatening dire consequences if students fail, as indicated by the teachers' responses, could increase anxiety level of test takers and prevent them from demonstrating their true performance on the test (Gordon & Fay, 2010). Teachers' threatening behaviour could be explained as a result of the additional pressure to improve students' results due to the accountability purpose of external tests (Popham, 2001; Redfield, 2001). Nonetheless, when instructions accompanying a test create threat, it results in additional pressure that disrupts the performance of students because they have fewer cognitive resources to devote to the tasks on the test (Alter et al., 2009; Croizet et al., 2001; Steele, 1997; Wright, 2008).

To ensure that assessment results adequately reflect what test takers know and can do, it is important to remove any problems in the testing process that introduce sources of bias. Analysis of the data indicated that teachers agreed to; ensuring that the testing environment is free from distractions, taking note of irregularities encountered during testing, and making special adjustments in the testing process for students with special needs. Such practices help to maintain comfortable testing environment that permits all test takers, including students with special needs, to demonstrate their true level of attainment on the constructs measured. Providing testing accommodations ensures equitable treatment of all test takers and reduces test bias because obstacles to accommodations can result in inappropriate interpretations of test scores for test takers from different groups (AERA et al., 2014).

Lack of understanding of an assessment task may prevent students' maximum performance. However, teachers' practices such as ensuring that

students are working in the appropriate sections of the test and announcing test time at regular intervals increase students' understanding of the assessment task. According to Popham (2001), teachers should describe the time limits, explain how students might distribute their time among parts for those assessment instruments with parts, and describe how students should record their responses. Nevertheless, clarifying ambiguous test questions only to students who ask about them is unethical and unfair to students who did not have access to such information. This is because students would have been able to supply the correct answers if they have had access to that information (Popham).

Overall, implementation of practices such as specifying uniform directions and comfortable environment, and consistent security procedures ensures that differences in test administration conditions do not inadvertently influence the performance of some test takers relative to others on the district-mandated test. Moreover, the results of the independent-samples t-test indicated that there was no significant difference in the actual test administration practices of trained and untrained teachers in educational measurement. This implies that students of both trained and untrained teachers are taken through similar administration conditions, and therefore, differences in their scores cannot be attributable to administration conditions.

Teachers' fair testing practices in grading students' performance on the test

Care should be taken to ensure that students' grades are not influenced by factors that are not relevant to the purpose of an assessment (JCTP, 1993). If grades are intended to measure student achievement, then they likely should not

take into account school behaviours such as participation in class and comportment (McMillan, 2011). However, analysis of the quantitative data showed mixed results. On one hand, teachers' responses to the questionnaire items indicated that grades were not influenced by students' gender, and bonus marks for extra work. Such practices ensure absence of bias in the district-mandated testing programme. For instance, students who simply followed test instructions or valued succinctness in their responses would be provided equal opportunity to demonstrate their knowledge since bonus marks were not awarded to extra responses (AERA et al., 2014).

On the other hand, the quantitative data showed that students' grades on the district-mandated test were influenced by factors such as students' efforts to learn, class attendance, language, and participation in class discussions. This result supports the findings of other studies (Green et al., 2007; Tierney et al., 2011; Zoeckler, 2005), as teachers in these studies also continue to weigh student effort in grading. Nonetheless, such practices award credit for response characteristics that are irrelevant to the constructs being measured on the test (AERA et al., 2014). Students' scores on the district-mandated test are supposed to provide an accurate undiluted indicator of students' mastery of instructional objectives (Wormeli, 2006).

Also, the results of the independent-samples t-test indicated that there was no significant difference in the test performance grading practices of trained and untrained teachers in educational measurement. This implied that scores of students of both trained and untrained teachers were influenced by the perceptions

and predispositions of the teachers, and therefore, were not deemed fair. This finding further indicated that teachers in this study do not strictly adhered to the instructions of the scoring guides provided by the test developer. Thus, the findings support the statement that it is not sufficient to assume consistency in scoring because a scoring scheme has been provided for grading (Anamuah-Mensah & Quagrain, 1998).

Fair testing practices of teachers in interpreting students' performance on the test.

Items 43 to 52 were to find out factors considered by teachers in the interpretation of students' performance on the district-mandated test. The result indicated that teachers considered, to a large extent, both psychometric and contextual factors in the interpretation of students' performance on the district-mandated test. Sizeable majority of the teachers (i.e., above 60%) indicated that they considered factors such as the level of vocabulary and sentence structure of the test, irregularities encountered during testing, the difficulty level of the questions, student's ability, adequacy of test time, and the alignment of test content to curriculum. Interpretation of scores on any test should take place along with a thorough knowledge of the technical aspects of the test, student's personal and social context, and limitations in the assessment methods used (AERA et al., 2014). Considering limitations in the assessment methods used is particularly important in this testing programme as the interview' data indicated that the test developer do not strictly adhered to rules for writing test items such as the use of optional questions for essay tests.

However, students' factors were *not* considered by a sizeable majority of the teachers. Factors, such as students' opportunity to learn, students' motivation to perform, student's test-taking skills, and student's socio-economic background, were considered by less than 60% of the teachers. To the extent possible, characteristics of all students, including gender, age, socio-economic status, and ability, must be considered throughout all stages of classroom testing, including interpretation of test results, so that barriers to fair assessment can be reduced. Not taking into account student's prior opportunity to learn could lead to misdiagnosis and inappropriate placement, which could have significant consequences for the test taker (AERA et al., 2014).

Regarding how to increase test fairness, the most frequently reported comment is whether or not the content of tests covers what have been taught in class (Jaturapitakkul, 2013). The finding that approximately 41% of the teachers in the participating schools do not consider students' prior opportunity to learn is particularly worrying as teachers' responses to Item 12 of the questionnaire indicated that approximately 36% of the teachers were unable to complete more than 90% of the curriculum for an academic term. Students of such teachers cannot and should not be held accountable for knowledge and skills they have had no opportunity to acquire (Herman & Choi, 2012). An achievement test that assumes a particular syllabus would not be a fair test for students who did not follow that curriculum (Willingham & Cole, 1997). Students should be exposed to the knowledge, skills, and dispositions that are measured on mandated test.

Without this type of exposure, it is not fair to expect students to have mastered the contents of the test (Lang & Wilkerson, 2008).

Fair testing practices of teachers in reporting district-mandated test results

Another main finding of the study was an answer to research question nine which sought to find out the kind of practices adopted by teachers in reporting students' results on the district-mandated test. The results showed that teachers agreed to six out of seven reporting practices outlined on the questionnaire, including the use of clear and simple language, timely reporting of test results, discussing with students' ways of improving their achievement, reporting separately students' achievement, effort and attitude, modifying reporting procedures for students with special needs, and sharing specific examples of students' strengths and weaknesses. This result supports previous study by Webber et al. (2009), which reported majority of educators indicating agreement or strong agreement to the item: 'Teachers regularly discuss with students' ways of improving their grades'.

Examinees have a right to a precise, timely, meaningful, and useful report of their performance on a district-mandated test. Also, performance reports that are written in simple language facilitate recipients' understanding (Cheng, 1998). The district-mandated test served as a learning experience as students were provided with prompt feedback about which of their answers were correct and which were incorrect (Cizek, 2012). Prompt feedback serves as a guidance system that keeps students on track of how to master the subject matter as it assists them in developing future plans for continued learning (JCSEE, 2003).

Moreover, it must be emphasized that teachers in this study disagreed to the item, "I publicly display students' test results with visible grades". This practice is considered ethical and fair because students' results on a district-mandated test are supposed to be reported in a confidential manner that respects the integrity of individual students (Cheng, 1998).

Teachers' fair uses of district-mandated test results

The last research question sought to find out uses made of district-mandated test results by teachers of the participating schools. The findings indicated that, in general, teachers in the study put students results on the district test to all uses outlined on the questionnaire. These uses included both high-stakes and low-stakes decisions. The high-stakes decisions mainly comprised of determining student grades, assessing teaching effectiveness, making classroom instructional decisions, and promoting and retaining students in class. The other decisions are considered as low-stakes, and it includes diagnosing students' learning needs, giving feedback to parents and students, evaluating students' progress, ranking and grouping students by achievement level, and identifying students in need of extra support.

The result supports the findings of NBETPP (2003), which reported the following as top uses of mandated test results by teachers in their study:

1. Assess my teaching effectiveness.
2. Give feedback to parents.
3. Give feedback to students.
4. Evaluate student progress.

The result is further supported by the findings of a survey conducted by Herman and Golan (1991), which reported that external examinations substantially affect teachers' instructional planning.

Mandated assessments are necessary to gauge learning, to monitor student progress, and to identify students who may need extra support (Johnson, 2008). Summative assessment at the district level also serves as an accountability measure that is generally used as part of the grading process (SQA, 2014). Moreover, assessing teaching effectiveness might be considered as a fair use of students' results since analysis of the interviews' data indicated that the test developer make effort to align the test to national standards through the development of a scheme of work and test specification (Johnson, 2008).

However, analysis of teachers' responses to Item 71 of the questionnaire indicated that approximately 40% of teachers in the study made decisions (i.e., uses of test results) based on only students' results on the district-mandated test. Such practice is unfair because measurement experts agree that it is inappropriate to use performance on a single test for making high-stakes decisions for students, teachers and schools (Popham, 2001; Rudner & Schafer, 2002). Multiple indicators of students' attainment or performances are essential so that students who are disadvantaged on one assessment have an opportunity to offer alternative evidence of their mastery of the objectives of the curriculum (Linn, 2003). To obtain a more complete picture or profile of a student's knowledge and skills, more than one assessment method should be considered. This helps to minimize

inconsistency brought about by different sources of measurement error (Henning, 2012).

Moreover, diagnosing students learning needs based on achievement test results is inappropriate and unfair. This is because, according to the test developer, the district test, as an achievement test, was based on the broad curriculum for an academic term, and therefore, sampled test items that cover majority of the topics to be treated in an academic term. This makes it difficult for the test to cover any specific topic in detail, and therefore, unlike diagnostic test results, students' results on an achievement test cannot be used for diagnostic purposes.

Summary of Key Findings

Analysis of the qualitative data showed that the test developer eliminates CIV through a number of practices, including considering students' characteristics in the development of test, adhering to standard item-writing rules and principles of UDA, and the use of sensitivity review panel. Test security was ensured through the use of item bank and test materials' packaging procedures. Also, standardization of the test was maintained through the provision of a fixed timetable and scoring guides. Lastly, the development of the scheme of work ensured that students are tested on contents that have been taught.

On the other hand, the quantitative data showed that majority of the test preparation practices adopted by both trained and untrained teachers in educational measurement were considered fair to all test takers. It was also found that practices that ensure test security were adhered to in the participating schools.

In addition, majority of the test administration practices adopted by both trained and untrained teachers in educational measurement ensured comfortable testing environment, reduced students' test anxiety, and increased students' understanding of the assessment tasks. In terms of grading of students' test performance, teachers' responses indicated that majority of the factors considered introduce CIV in students' results.

The quantitative data further indicated that teachers considered, to a large extent, both psychometric and contextual factors in the interpretation of students' performance on the district-mandated test. Students were also given precise, timely, meaningful, useful, and confidential report of their performance on the district-mandated test. Lastly, both high-stakes and low-stakes decisions in the classroom were based only on students' test results.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Overview of Research Problem and Methodology

The study sought to investigate fair testing practices of a test developer and teachers in a district-mandated testing programme in the Ashanti Region of Ghana. The study was primarily aimed at finding out whether practices adopted by the test developer and teachers, with regard to developing the district test, preparing students for the test, administering and scoring of the test, interpretation and reporting of test results, and uses made of students' test results, provides all test takers an equal opportunity to demonstrate their knowledge and skills on the test. 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.

A multilevel mixed methods study was adopted. Three personalities, representing CePME, were purposively selected in order to explore the practices of the test developer. In addition, nine test questions were selected to further gather evidence of the test developer's practices. In order to gather data on teachers' practices, a sample of 50 public JHSs were randomly selected from three districts in the Ashanti Region that administers assessment materials, prepared by CePME. The two-stage cluster sampling technique was further employed to sample six teachers from the 50 clusters of JHSs sampled. The

sample size for the study comprised 300 teachers.

The main instruments for data collection included an interview protocol and test questions for the test developer's strand of the study, and a 72-item questionnaire for the teachers' strand of the study. The data collected were analysed mainly by qualitative content analysis, frequency and percentage tables, one-sample t test, and the independent-samples t test.

Summary of Results

Analysis of the test developer's practices indicated that test taker's characteristics, such as gender, age, ability, and opportunity to learn, were considered in the development of the district-mandated test. Also, experts were employed to look at the sensitivity nature of items on the test in terms of language and other representations that might be offensive to some test takers. Instructions on the test were specific as to the regulations governing the test. Test instructions encouraged students to complete all items on the test and also included weights assigned to parts of the test. To a large extent, item writing rules for multiple-choice and essay items, and recommendations for increasing the legibility of test items were also incorporated into the development of the test. However, a number of flawed items, resulting from violations of very few item-writing rules, were identified.

In terms of test security, the use of a secure test bank and the packaging of test materials into school-by-school and subject-by-subject protected the integrity of the test items. The test developer also indicated that the district test was administered during an official scheduled time, which was strictly adhered to in

all participating schools. The test developer, through the provision of a scoring scheme, also ensured consistent scoring procedures.

The test developer further made mention of effort to include test items on all content areas and topics taught in a term. This effort was evident in the development and used of test specifications during item writing. Moreover, a scheme of work was provided to all participating schools in order to align the test to both curriculum and instruction.

Under test preparation, responses of the teachers indicated that they agreed to all of the test preparation practices outlined on the questionnaire. However, two of such practices, including modifying previous test questions as practice test in class, and preparing students on actual previous district test questions, amounts to teaching the test, and thus, were considered unfair and unethical. Also, there was no statistically significant difference in test preparation practices between teachers who have received training in educational measurement and those who have received no training in educational measurement.

Under test administration, analysis of the questionnaire data indicated that teachers' practices ensured security of the district test, maintained comfortable testing environment, and specified uniform directions for all test takers. However, few of the practices agreed to by the teachers, including I, single-handed, supervise more than 25 students; I perform other unrelated professional duties during testing; students are threatened dire consequences if they fail; and ambiguous test questions are clarified only to students who ask about them, do not ensure test security, comfortable environment, and uniform directions.

Furthermore, there was no statistically significant difference in test administration practices between teachers who have received training in educational measurement and those who have received no training in educational measurement.

Concerning grading of students' test performance, the results indicated that students' grades on the test were influenced by students' comportment and other behaviours, such as students' efforts to learn and participation in class discussions, which are irrelevant to the constructs measured on the district-mandated test. The results further revealed that there was no statistically significant difference, between teachers who have received training in educational measurement and those who have received no training in educational measurement, in terms of factors influencing students' grades on the test.

Under interpretation of students' test performance, more than 60% of the teachers in this study indicated that they considered psychometric and/or contextual factors in the interpretation of students' performance. Majority of test takers' factors (i.e., 4 out of 5 factors) were considered by less than 60% of the teachers in this study in interpreting students' performance on the test.

Concerning how teachers report students' results on the district-mandated test, the results showed that teachers practiced all seven ethical and fair reporting practices sought for by items on the questionnaire.

Concerning uses made of students' test results on the district-mandated test, the findings indicated that, in general, teachers in this study put students results to all uses outlined on the questionnaire, including both high-stakes and

low-stakes uses. More importantly, the results further indicated that 40% of the teachers made high-stakes and low-stakes decisions based on only district-mandated test results.

Conclusions

Generally, statements about the fairness of the test developer and teachers' testing practices can be made in terms of three broad views, including practices that ensure (1) absence of bias against any student or group of students, (2) equitable treatment of all students taking the test, and (3) students' opportunity to learn the content of the mandated test. It is evident from the study that, on the whole, the test developer ensured that the district test was not biased against any student or group of students taking the test. This was evident in practices such as consideration of test taker characteristics and sensitivity review of test content by experts. Such practices ensured that the test do not discriminate against groups of students. Additionally, comprehensive test instructions, training of item writers, and compliance to item writing rules and legibility recommendations ensured that characteristics of the test itself do not impede on the performance of any student taking the test.

Also, standardization procedures such as fixed timetable for the administration of the test in all participating schools, and the provision of scoring scheme ensured that all test takers are given equitable treatment in the administration of the district-mandated test. The packaging of test materials into schools-by-schools and class-by-class, and the use of item bank prevent test insecurity such as leakages of test questions, which gives some students an unfair

treatment over others. The provision of the scheme of work ensured that students are given an opportunity to learn the contents covered by the test while the development of the test specification ensured that students at all levels of learning have an opportunity to demonstrate their knowledge and skills on the district test. It could be concluded, therefore, that the practices of the test developer, to a large extent, were fair to all students taking the test.

The results of the study further indicated that on test preparation, administration of the test, interpretation of students' performance, reporting of test results, and uses made of test results, teachers generally reported that they engaged in practices that provided all students an equal opportunity to demonstrate their knowledge and skills on the district-mandated test. Ethical test preparation and reporting practices, such as the ones agreed to by the teachers, provided students with information that afford them an opportunity to learn the content of the test, and also ensured that the test does not discriminate against certain groups of students. On test administration and interpretation of students' test performance, teachers reported they engaged in a sizeable number (i.e., majority) of practices that were considered to be fair to all students. For example, interpreting students' performance in the light of psychometric, contextual and test taker factors, and ensuring comfortable testing environment eliminated test bias while specifying uniform directions and test security procedures promoted equitable treatment of all examinees. Furthermore, majority (i.e., 60%) of the teachers indicated that they made unbiased and ethical uses of test results by *not* basing decisions in the classroom on only students' performance on the district-

mandated test. Thus, it could be concluded that in terms of test preparation, administration, interpretation, reporting and uses made of test results, the practices of teachers, to a large extent, were fair to all students taking the test.

However, on grading of students' test performance, the results of this study indicated that majority of the teachers' practices introduce CIV in students' results on the test, and therefore, create a biased result that, potentially, could discriminate against individual students, as well as certain groups of students. It could, therefore, be concluded that in terms of grading students' test performance, the practices of teachers, to a large extent, were not fair to any student or groups of students taking the district-mandated test.

Finally, the results of the study indicated that under test preparation, administration, and grading of students' test performance, teachers who have received training in educational measurement did not report practices that were statistically and significantly different from that of their colleagues who did not receive training in educational measurement. It could, therefore, be concluded that training in educational measurement had little or no impact on teachers' fair testing practices in terms of preparing students, administering the test, and grading students' test performance. This gives an indication that students received similar treatment on the district-mandated test, irrespective of whether the teacher is trained or untrained in educational measurement.

Recommendations

In view of the study's findings and the conclusions arrived at, the following recommendations are made to improve upon the degree of fairness of

the district-mandated testing program:

1. As regards the number of flawed items found on the test questions, resulting from very few violations of item writing rules, the test developer should institute a formal content review panel to appraise the technical quality of the items, looking for items that are free from such flaws as ambiguity and factual inaccuracy (ODE, 2011). Unlike the sensitivity review panel utilized by the test developer, content review panel could examine the extent to which items conform to widely accepted item-writing rules (Haladyna et al., 2002), and also the alignment of test content to specific standards (ODE). According to Haladyna (1999), all items must go under review for content, item writing violations, and grammatical errors.
2. On test preparation, the results indicated that teachers provide test preparation on previous district test questions, which is considered unethical and unfair among measurement experts as it leads to teaching the district-mandated test. I, therefore, recommend that practice questions be provided to teachers along with the scheme of work. Practice test questions will provide clear instructions and, support students' understanding of what will be required during the assessment, which eventually reduces construct-irrelevant test variance. Also, it will explain the overall design of the test and describe the specific content that appears on the test, conveying to teachers what their students can expect on the district-mandated test (ODE, 2011).
3. A comprehensive test manual is essential to properly administering the mandated test. Procedures for the standardized administration of a mandated

test should be carefully documented by the test developer and followed carefully by the test administrator. Therefore, on the finding that teachers testing practices increased students' anxiety level, encouraged cheating among students, and introduced CIV in students' scores, it is recommended that a comprehensive test manual should be developed by the test developer, and shared with all participating schools. The test manual will improve upon the degree of standardization of the assessment system in order to further ensure consistent test administration procedures. The test developer will promote fair testing practices by establishing standardized testing procedures, and communicating these procedures to test administrators and test takers (Plake & Jones, 2002).

4. In addition to the above, I also recommend that the test developer engage the services of highly experienced test administrators as 'chief proctors' who will serve as monitoring teams, and assume full responsibility for all aspects of the district-mandated test, including supervision of test administration and grading procedures (Cheng, 1998). These professional proctors could comprise of heads of the various participating schools and measurement experts at the Districts' Offices of Education. Chief proctors are key to successful, secure, well-organized and well-administered large-scale examinations (Downing, 2004). Use of monitoring teams would also improve test security.

Suggestions for Further Research

The following are suggested for future research:

1. In order to accept or refute the findings of the study, and generalize them for the whole of the country, it is suggested that the study is replicated in other regions of the country at the JHS level. Also, the practices of other private testing companies may be considered in future research.
2. It is also suggested that further research on the problem of test fairness of district-mandated test could include the calculation of DIF statistics, which considers students' actual scores on the test. DIF is a statistical procedure for judging whether test items are functioning in the same manner for different groups of test takers. It will also help to determine whether students with equal ability but representing different groups do not have the same probability of responding correctly to items on a district-mandated test (McNamara & Rover, 2006).

REFERENCES

AERA, APA, NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

AERA, APA, NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Airasian, P. W. (1999). *Assessment in the classroom: A concise approach* (2nd ed.) New York, NY: McGraw-Hill.

Alberta Education. (2009). *The Alberta student assessment study: Final report.* Retrieved from http://education.alberta.ca/department/ipr.aspx

Alberta et al. (2006). *Rethinking classroom assessment with purpose in mind: Assessment for learning, assessment as learning, assessment of learning*. Retrieved from www.wncp.ca

Allen, J. D. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House, 78(5),* 218-223

Alter, A. L., Aronson, J., Darley, J. M., Rodriguez, C., & Ruble, D. N. (2009). Rising to the threat: Reducing stereotype threat by reframing the threat as a challenge. *Journal of Experimental Psychology, 46(2010)*, 166-171

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

Anamuah-Mensah, J., & Quagrain, K. A. (1998). Teacher competence in the use of essay-tests: A study of secondary schools in the Western Region of Ghana. *The Oguaa Educator, 12(1)*, 31-43

Ananda, S. (2003). *Rethinking issues of alignment under No Child Left Behind*. San Francisco: WestEd.

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

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

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford, UK: Oxford University Press.

Bachman, L. F. (1995). *Fundamental consideration in language testing.* Oxford: OUP.

Baharloo, A. (2013). Test fairness in traditional and dynamic assessment. *Theory and Practice in Language Studies, 3(10),* 1930-1938

Baškarada, S. (2014). Qualitative case studies guidelines. *The Qualitative Report, 19(24),* 1-18

Baxter, G., & Mislevy, R. J. (2005). *The case for an integrated design framework for assessing science injury (PADI Technical Report 5).* Menlo Park, CA: SRI International.

Bouville, M. (2008). *The obsession with exam fairness*. Retrieved from http://www.mathieubouville.com/education-ethics/Bouville-exam-fairness

Briggs, D. C. (2009). *Preparation for college admission exams.* National

Association for College Admission Counselling, Colorado, USA.

Bunch, M. B. (2012). Aligning curriculum, instruction, and assessment. *Measurement Incorporated.* Retrieve from www.measurementinc.com

Buren, C., Ziker, C., Brashear, F., & Crosswell, J. (2006). *Response to alignment study findings.* Phoenix, Arizona: Arizona Department of Education.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for University faculty*. Brigham Young University Testing Services, Brigham.

Camara, W. J. (2007). *Standards for educational and psychological testing: Influence in assessment development and use*. Unpublished paper developed for the Joint Committee on Testing Standards.

Camilli, G. (2006). Test Fairness. In R. L. Brennan (Ed.), *Educational Measurement,* 221-256. Washington, DC: American Council on Education/Praeger.

Cannell, J. J. (1989). *The "Lake Wobegon" report: How public educators cheat on standardized achievement tests.* Albuquerque, NM: Friends for Education.

Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences*. (3rd ed. revised). Philadelphia, PA: National Board of Medical Examiners.

Chan, K. K. (2009). *How can a report card facilitate assessment for learning?* Paper presented at the 35th International Association for Educational Assessment (IAEA) Annual Conference, Brisbane, Australia.

Chandler, S. B. (2001). Running head: Legibility and comprehension of onscreen type. Retrieved from http://scholar.lib.vt.edu/theses/available/etd-11172001-152449/unrestricted/chandler.pdf

Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation, 24(3),* 279-301

Cheng, L. (2010). The history of examinations: why, how, what and whom to select? *In L. Cheng, & A. Curtis (Eds.), English language assessment and the Chinese learner* (pp. 13-25). NY: Routledge

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Erlbaum.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17(1),* 31 - 43

Cizek, G. J., Germuth, A. A., & Schmid, L. A. (2011). *A checklist for evaluating K-12 assessment programmes.* Kalamazoo: The Evaluation Center, Western Michigan University. Retrieved from http://www.wmich.edu/evalctr/checklists/

Close, D. (2009). Fair grades. *Teaching Philosophy, 32(4),* 361-398

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38,* 369-382

Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions for medical education activities and self-assessment

modules. *Radio Graphics, 26(2)*, 542-551

CRDD. (2005). *Opportunity to learn English and Mathematics in Ghanaian primary schools.* Accra: Author.

Creswell, J. W. (2007). *Qualitative inquiry and research method: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W., & Plano-Clark, V. (2010). *Designing and conducting mixed methods research* (3rd ed.). Thousand Oaks, CA: Sage.

Croizet, J., Désert, M., Dutrévis, M., & Leyens, J. (2001). Stereotype threat, social class, gender, and academic under-achievement: when our reputation catches up to us and takes over. *Social Psychology of Education*, *4*, 295 - 310

Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013). *Criteria for high-quality assessment.* Stanford, CA: Stanford Center for Opportunity Policy in Education.

Davies, A. (2010). Test fairness: A response. *Language Testing, 27*, 171-176

Dee, T. (2007), Teachers and the gender gaps in student achievement, *Journal of Human Resources, 22(4),* 23-45

DePascale, C. A. (2003, April). *The ideal role of large-scale testing in a comprehensive assessment system*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Dodeen, H. (2015). Teaching test-taking strategies: Importance and techniques. *Psychology Research, 5(2),* 118-113

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions:

Do multiple-choice item writing principles make any difference? *Academic Medicine, 77(10)*, 103–104

Downing, S. M. (2004, April). *The effects of violating standard item-writing principles: The impact of flawed test items on classroom achievement tests and students*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

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

Educational Testing Service (ETS). (2012). *Guidelines for best test development practices to ensure validity and fairness for international English language proficiency assessments*. Princeton, NJ: Author.

Educational Testing Service (ETS). (2014). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Educational Testing Services (ETS). (2009). *ETS guidelines for fairness review of assessments*. Princeton, NJ: Author.

Elhoweris, H., & Alsheikh, N. (2010). UAE teachers' awareness & perceptions of testing modifications. *Exceptionality Education International, 20*, 37-48

Etsey, Y. K. A. (2012). *Assessment in education*. Lecture notes on EPS 311. Unpublished document, University of Cape Coast, Ghana.

Felder, R. M. (2002). Designing tests to maximize understanding. *Journal of Professional Issues in Engineering Education and Practice, 128*, 1–3

Finn, J., & Jacobson, M. (2008). *Just Practice: A social justice approach to social work.* Peosta, IL: Eddie Bowers Publishing.

Gallagher, D. J. (1998). *Classroom assessment for teachers*. Upper Saddle River, NJ: Merrill.

Gasser, B., Boeke, J., Haffernan, M., & Tan, R. (2005). The influence of font type on information recall. *North American Journal of Psychology, 7(2),* 181-188

Gay, L. R., Mills, G. E., & Airasian, P. (2009). *Educational research competencies for analysis and applications.* Columbus: Pearson Merrill Prentice Hall.

Ghana Ministry of Education (GMOE). (1999). *National education forum report.* Accra: Author.

GhanaWebb. (2016). Ashanti Region. Retrieved from http://www.ghanaweb.com/GhanaHomePage/geography/ashanti_region

Gibbons, S., & Chevalier, A. (2007). *Teacher assessments and pupil outcomes*. *Research Papers in Education, 23(2),* 113-123

Gichuru, F. M. (2014). *Classroom assessment practices in Kenyan secondary schools: Teacher perspective*. Unpublished master's thesis, University of Nairobi, Kenya.

Gipps, C. (1995). What do we mean by equity in relation to assessment? *Assessment in Education, 2(3),* 271–282

Gipps, C., & Stobart, G. (2009). *Fairness in assessment*. In C. Wyatt-Smith & J. Cumming (Eds.), Educational assessment in 21st century: Connecting theory and practice (pp. 105-118). Netherlands: Springer Science Business Media.

Gordon, M. E., & Fay, C. H. (2010). The effects of grading and teaching practices on students' perceptions of grading fairness. *College Teaching, 58(3)*, 93-98

Green, K. H., & Emerson, A. (2007). A new framework for grading. *Assessment & Evaluation in Higher Education, 32(4)*, 495-511

Green, S., Johnson, R. L., Kim, D. H., & Pope, N. S. (2007). Ethics in classroom assessment practices: Issues and attitudes. *Teaching and Teacher Education, 2(7)*, 999–1011

Guskey, T. R., & Jung, L. A. (2009). Grading and reporting in a standards-based environment: Implications for students with special needs. *Theory into Practice, 48(1)*, 53–62

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items (3rd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1)*, 17–27

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15(3)*, 309–334

Henning, G. (2012). Twenty common testing mistakes for EFL teachers to avoid. *English Teachers Forum, 20(3)*, 33-36

Herman, J. L., & Choi, K. (2012). *Validation of ELA and Mathematics assessments: A general approach.* Center for Research on Evaluation, Standards, and Student Testing, Los Angeles: University of California.

Herman, J. L., & Golan, S. (1991). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12(25)*, 41-42

Hinnerich, B. T., Hoglin, E., & Johanneson, M. (2011). *Ethnic discrimination in high school grading: Evidence from a field experiment*, pp. 1–36.

Illinois State Board of Education (ISBE). (2014). *Professional testing practices for educators: Illinois Standards Achievement Test (ISAT)*. USA: Author.

International Test Commission (ITC). (2001). International guidelines for test use. *International Journal of Testing, 1(2),* 93-114

Jaturapitakkul, N. (2013). Students' perceptions of traditional English language testing in Thailand. *Academic Journal of Interdisciplinary Studies, 2(3),* 445-452

Johnson, D. (2008). *Stop high-stakes testing: An appeal to America's conscience.* Lanham, MD: Rowman & Littlefield Publishers.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1(2),* 112–133

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Johnstone, C. J., Thompson, S. J., Moen, R. E., Bolt, S., & Kato, K. (2005). *Analyzing results of large-scale assessments to ensure universal design* (Technical Report 41). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Technical41.htm

Joint Advisory Committee on Testing Practices (JACTP). (1993). *Principle for fair assessment practices for education in Canada.* Edmonton, Alberta: Author.

Joint Committee on Standards for Educational Evaluation (JCSEE). (2003). *The student evaluation standards: How to improve evaluations of students.* Washington DC: AERA, APA, NCME.

Joint Committee on Testing Practices (JCTP). (2000). *Rights and responsibilities of test takers: Guidelines and expectations*. Washington, DC: Author.

Joint Committee on Testing Practices (JCTP). (2004). *Code of fair testing practices in education.* Washington, DC: Author.

Kavanagh, L. (2013). *A mixed methods investigation of parental involvement in Irish Immersion Primary Education: Integrating multiple perspectives*. Unpublished doctorate's thesis, University College Dublin, Dublin.

Kellow, J. T., & Jones, B. D. (2008). The effects of stereotypes on the achievement gap: Re-examining the academic performance of African American high school students. *Journal of Black Psychology, 34(1),* 94-120

Khalanyane, T., & Hala-hala, M. (2014). Traditional assessments as a subjectification tool in schools in Lesotho. *Educational Research and Reviews, 9(17),* 587-593

Khoii, R., & Shamsi, N. (2012). A fairness issue: Test method facet and the validity of grammar sub-tests of high-stakes admissions tests. *Literacy Information and Computer Education Journal, 1(1),* 801-809

Klinger, D. A., & Luce-Kapler, R. (2007). Walking in their shoes: Students' perceptions of large-scale high-stakes testing. *The Canadian Journal of Program Evaluation, 22(3),* 29–52

Kunnan, A. J. (2004). Test fairness. In Milanovic, M., & Weir, C., (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27-48). Cambridge, UK: Cambridge University Press.

La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A., & Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment.* Washington, DC: Council of Chief State School Officers.

Lam, T. C. M. (1995). *Fairness in performance assessment*. ERIC digest, EDO-CG-95-25.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 25(1),* 127-135

Lang, W. S., & Wilkerson, J. R. (2008, February). *Accuracy vs. validity, consistency vs. reliability, and fairness vs. absence of bias: A call for quality*. Paper presented at the annual meeting of the American

Association of Colleges of Teacher Education (AACTE), New Orleans, LA.

Lekoko, R. N., & Koloi, S. (2007). Qualms in marking university students' assignment by teaching staff: Does correlation of student's expectations and teacher's feedback matter? *Journal of Business, Management and Training, BIAC, 4*, 34-45

Levy, B. (1996). Improving memory in old age through implicit self-stereotyping. *Journal of Personality and Social Psychology, 71*, 1092–1107

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry.* Beverly Hills, CA: SAGE.

Linn, R. L. (2003). Assessments and accountability. *Educational Researcher, 23(9),* 4–14

Linn, R. L. (2008). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32(3),* 3–13

Lissitz, R. W., & Schafer, W. D. (2002). *Assessment in educational reform.* Boston, MA: Allyn and Bacon.

Maxwell, G. S., Cumming, J. J., Wyatt-Smith, C. M., & Colbert, P. (2012). *Developing students' test-wiseness*. Brisbane: Griffith University.

McCabe, D. L., Trevino, L. K., & Kenneth, D. B. (2001). Cheating in academic institutions: A decade of research. *Ethics & Behaviour, 11*, 219-232

McDonald, M. E. (2008). Developing trustworthy classroom tests. In Penn, B.K. (Ed.), *Mastering the teaching role: A guide for nurse educators* (pp. 275 - 286). Philadelphia: FA Davis.

McMillan, J. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20(1),* 21–44

McMillan, J. H., & Schumacher, S. (2001). *Research in education (5th ed.).* New York; Longman.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Oxford, UK: Blackwell Publishing.

Merriam, S. B. (1988). *Case study research in education.* San Francisco, CA: Jossey-Bass Publishers.

Merriam, S. B. (2001). *Qualitative research and case study applications in education*. San Francisco, California: Jossey-Bass Publishers.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). pp. 13–103. Washington, DC: American Council on Education and Macmillan.

Munson, R., & Parton, C. (2013). *Bias and fairness in state testing*. Retrieved from http://apps.leg.wa.gov/billinfo/summary.aspx?bill=1450&year=2013

National Association of School Psychologists (NASP). (2002). *Large-scale assessments and high stakes decisions: Facts, cautions and guidelines.* Bethesda, MD: Author.

National Board on Educational Testing and Public Policy (NBETPP). (2003). *Perceived effects of state-mandated testing programmes on teaching and learning: Findings from a national survey of teachers.* Boston College: Author.

National Council on Measurement in Education (NCME). (1995). *Code of professional responsibilities in educational measurement (CPR).* Washington DC: Author.

Nicole, F. (2013). *Does teaching women about stereotype threat reduce its effects on math performance?* Senior honours theses. Paper 73. Retrieved from http://digitalcommons.brockport.edu/honors

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.

Nitko, A. J. (2004). *Educational assessments of students.* Englewood Cliffs, NJ: Prentice Hall.

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

Onwuegbuzie, A., & Leech, N. L. (2007). Sampling designs in qualitative research: Making the sampling process more public. *The Qualitative Report, 12(2),* 238-254

Oregon Department of Education (ODE). (2011). *Technical report: Oregon state-wide assessment.* Oregon: Author.

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

Patton, M. Q. (1990). *Qualitative evaluation and research methods. (2nd ed.).* London: Sage.

Penn, B. K. (2009, August). *Test item development and analysis*. Presented at Creighton University School of Nursing Faculty Retreat, Omaha, NE.

Plake, S. B., & Jones, P. (2002, February). *Ensuring fair testing practices: the responsibilities of test sponsors, test developers, test selectors, and test takers in ensuring fair testing practices*. Paper presented at the meeting of the Association of Test Publishers, Carlsbad, CA.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal, 48(4),* 965-995

Popham, W. J. (2001). Teaching to the test. *Educational Leadership, 58(6),* 16-20

Popham, W. J. (2002, April). *High-stakes tests: Harmful, permanent, fixable.* Paper presented at the Annual Conference of the American Research Council, New Orleans, LA.

Popham, W. J., & Lindheim, E. (1980). The practical side of criterion-referenced test development. *NCME Measurement in Education, 10(4),* 1–8. Publications.

Porter, A. C. (1993). *Defining and measuring opportunity to learn.* Madison, WI: University of Wisconsin.

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.

Quality Assurance Agency for Higher Education (QAA). (2012). *Understanding assessment: Its role in safeguarding academic standards and quality in*

higher education. A guide for early career staff (2nd ed.). Retrieved from http://www.qaa.ac.uk/en/Publications/Documents/understanding-assessment.pdf

Rahn, M. L., & Stecher, B. (1997). Making decisions on assessment methods: Weighing the trade-offs. *Preventing School Failure, 41(2),* 1-20

Redfield, D. (2001). *Critical issues in large-scale assessment: A resource guide.* Council of Chief State School Officers. Office of Educational Research and Improvement. Washington, DC.

Reeves, M. (2003, October). *Markets, "missions" and languages of higher education reform in Kyrgyzstan.* Paper Presented to the CESS the Annual Conference. Cambridge, MA.

Reiner, C. M., Bothell, T. W., Sudweeks, R. R., & Wood, B. (2002). *How to Prepare Effective Essay Questions: Guidelines for University Faculty*. Brigham: Brigham Young University.

Rodabaugh, R. C. (1991). Institutional commitment to fairness in college teaching. In L. Fish (Ed.). *Ethical dimensions of college and university teaching* (pp. 37-45). San Francisco: Jossey-Bass.

Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation,* 4(2). Retrieved from http://PAREonline.net/getvn.asp?v=4&n=2

Rudner, L., & Schafer, W. (2002). *What teachers need to know about assessment*. Washington, DC: National Education Association.

Schouwstra, S. J. (2000). *On testing plausible threats to construct validity.* The

Institutional Repository of the University of Amsterdam (UvA). Retrieved from http://dare.uva.nl/document/56520

Scottish Qualifications Authority (SQA). (2014). *Guide to assessment*. Retrieved from www.sqa.org.uk

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20(5),* 8-11

Stainback, W., & Stainback, S. (1996). *Controversial issues in special education. Divergent perspectives* (2nd ed.). Boston: Ally and Bacon Press.

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629

Stobart, G. (2005). Fairness in multicultural assessment. *Assessment in Education: Principle, Policies, and Practices, 12,* 275-87

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education, 30(4),* 331–347

Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice items: Guidelines for University faculty.* Brigham: Brigham Young University Testing Services and the Department of Instructional Science.

Suskie, L. (2000). *Fair assessment practices: Giving students equitable opportunities to demonstrate learning*. Boston, MA: AAHE Bulletin.

Susuwele-Banda, W. J. (2005). *Classroom assessment in Malawi: Teachers' perceptions and practices in Mathematics*. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Virginia.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42,* 198-206

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research. Integrating qualitative and quantitative approaches in the social and behavioural sciences.* Thousand Oaks, CA: Sage.

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal of Mixed Methods Research, 1(77),* 77-100

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments (Technical Report 42)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Technical42.htm

Thurlow, M., Quenemoen, R., Thompson, S., & Lehr, C. (2001). *Principles and characteristics of inclusive assessment and accountability systems (Synthesis Report 40)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from

http://education.umn.edu/NCEO/OnlinePubs/Synthesis40.html

Tierney, R. D. (2013). Fairness in classroom assessment. In J. H. McMillan (Ed.). *SAGE Handbook of Research on Classroom Assessment* (pp. 125-144). Thousand Oaks, CA: SAGE Publications.

Tierney, R. D., Simon, M., & Charland, J. (2011). Being fair: Teachers' interpretations of principles for standards-based grading. *The Educational Forum, 75(3),* 210-227, DOI: 10.1080/00131725.2011.577669

Tunks, J. (2001). The effects of training in test item writing on test performance of junior high students. *Educational Studies, 27(2),* 129-142

Virginia Department of Education (VDE). (2015). *Students with disabilities: Guidelines for special test accommodations.* Retrieved from http://doe.virginia.gov/testing/alternative_assessments/index.shtml.

Wankat, P., & Oreovicz, F. (1993). *Teaching Engineering*. New York: McGraw-Hill.

Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna, (Eds). *Handbook of test development* (p. 155- 180). Mahwah, NJ: Lawrence Erlbaum Associates.

Webber, C. F., Atkin, N., Lupart, J., & Scott, S. (2009). *The Alberta student assessment study: Final report.* Retrieved from http://education.alberta.ca/department/ipr.aspx

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wormeli, R. (2006). Accountability: Teaching through assessment and feedback, not grading. *American Secondary Education, 34(3),* 14-27

Wright, R. J. (2008). *Educational assessment: Tests and measurements in an age of accountability*. Thousand Oaks, CA: Sage.

Wright, S. C., & Taylor, D. M. (2003). The social psychology of cultural diversity: Social stereotyping, prejudice, and discrimination. In M. A. Hogg, & J. Cooper, (Eds.). *Handbook of Social Psychology* (pp. 41-63). London: SAGE Publications.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27(2),* 147-170

Xiaomei, S. (2014). *Test fairness in a large-scale high-stakes language test.* Unpublished doctoral thesis, Queen's University, Ontario, Canada.

Yin, R. K. (2011). *Qualitative research from start to finish*. New York: Guilford Press.

Yip, D. Y., & Cheung, D. (2005). Teachers' concerns on school-based assessment of practical work. *Journal of Biological Education, 39(4),* 156–62

Younger, M., Warrington, M., & Jaquetta, W. (1999). The gender gap and classroom interactions: Reality and rhetoric? *British Journal of Sociology of Education, 20,* 15-45

Zhang, H. Q. (2004). Accession to the World Trade Organization: Challenges for China's travel service industry. *International Journal of Contemporary Hospitality Management, 16(6),* 369-372

Zoeckler, L. G. (2005). *Moral dimensions of grading in high school English.*

Proquest Digital Dissertations UMI (No. AAT3183500).

Zucker, S. (2004). *Administration practices for standardized assessments: Assessment report*. San Antonio, TX: Pearson Education Inc.

APPENDICES

APPENDIX A

JUNIOR HIGH SCHOOLS AND DISTRIBUTION OF TEACHERS SAMPLED

District/ Municipality	JHSs Sampled	Teachers Sampled
Atwima Nwabiagya	1. Abuakwa D/A 'B'	4
	2. Abuakwa D/A 'C'	4
	3. Adankwame Roman Catholic	6
	4. Adibiya Islamic	6
	5. Afari Presbyterian	6
	6. Asenemaso D/A 'B'	4
	7. Asuofia D/A 'A'	6
	8. Bandaogo Islamic	5
	9. Barekese D/A 'B'	6
	10. Barekese Methodist	6
	11. Bokankye D/A	3
	12. Fufuo D/A	6
	13. Jankobaa Roman Catholic	4
	14. Kuffuor D/A	6
	15. Manhyia D/A	6
	16. Mim D/A	6
	17. Nkawie D/A Experimental 'A'	6
	18. Nkawie Islamic	4
	19. Nkawie Panin D/A	6
	20. Nketia D/A	6
	21. Sepaase D/A	5
	22. Toase D/A	6
	23. Toase Saint Peters Roman Catholic	6
Mampong	1. Aframano M/A	3

	2. Apaah M/A	4
	3. Bosofour Roman Catholic	6
	4. Brofoyedru M/A	5
	5. Kofiase Abubakar Islamic	6
	6. Krobo M/A	3
	7. Mampong Seventh-Day Adventist	6
	8. Mensah Saahene	5
	9. Messiah Baptist M/A	6
	10. Mprim M/A	6
	11. Muslim Mission	5
	12. Ninting M/A	6
	13. Nkwanta M/A	5
	14. Nyinapong M/A	6
	15. ST. Monica's Experimental	6
	16. ST. Paul Roman Catholic	5
	17. T. I. Ahmadiyya	5
Asante-Akim Central	1. Konongo Mines M/A 'C'	3
	2. Konongo Islamic	3
	3. Konongo Roman Catholic	6
	4. Odumasi Presbyterian	4
	5. Odumasi ST. Mary's Roman Catholic	4
	6. Kramokrom M/A	4
	7. Kyekyebiase M/A	3
	8. Nyaboo M/A	6
	9. Dwease M/A	4
	10. Konongo Presbyterian 'B'	3

APPENDIX B

UNIVERSITY OF CAPE COAST

INTERVIEW PROTOCOL FOR TEST DEVELOPER

Topic: *Fair Testing Practices of Test Developers and Teachers in District-Mandated Testing Programmes in the Ashanti Region of Ghana.*

Purpose of the Test

1. Why was this mandated testing program instituted in the participating districts?
    a. What benefits or advantages do this testing program has over teacher-made test?

Alignment Analysis

2. How do you determine the knowledge and skills to be tested?
    a. Which materials aid in this process?
    b. Please give a description of such activities?

Item Writing

3. Who are the item writers?
    a. What criteria are used to select item writers?
    b. How many item writers are recruited for each subject?
    c. How many items are assigned to each item writer?
    d. What forms of guidelines or training are given to item writers?
    e. Can you please describe the process of developing the items?
4. What characteristics of test takers are considered in the development of the test?

a. How do these characteristics influence the development of the test?

Review of Test Items

5. How do you ensure that the content or language used on the test does not offend any student or groups of students taking the test?
   a. Who or which group of people forms the panel to review the items written?
   b. What characteristics are considered in forming a panel for the review of the items?
   c. Can you please give a brief description of the review process?

UDA and Testing Accommodations

6. There is a move to include all children of school going age into mainstream education, including students with special needs. What measures are in place to ensure that these students are given equal opportunity to demonstrate their knowledge on your test?
   a. What kinds of opportunities are provided to teachers and/or students to make special request for changes in the testing program?

Assembling of Test Items

7. How do you ensure that the final test questions are a representative sample of the contents taught in schools?
   a. What factors are considered in distributing items/questions to the various contents?
8. How do you determine the difficulty level of the items?
   a. How do you gather information about students' performance on the test?

Informing Test Takers

9. How are students informed of the contents covered by the test?
    a. How are students informed of the format of the test?
    b. What samples of test materials are given to participating schools in order to help students familiarize themselves with the nature of the test?

Test Security

10. How are you convinced that the items written are not leaked to the participating schools even before they are administered?
11. How are test instruments packaged and delivered to the participating districts and schools?
    a. What measures are undertaken to ensure the security of test materials during this transfer period?
    b. What security policies guide the assessment instruments once they have been delivered to the participating schools?
    c. What is the time frame between the transportation of test instruments to the participating schools and the actual test administration?

Test Administration

12. What forms of guidelines or training are given to teachers in the administration of the test?
    a. What forms of assistance are provided to test administrators to help resolve novel situations or difficulties that may pop-up during test administration?

Scoring Process

13. What materials and guidelines are provided for scoring the test?
    a. When are these materials prepared and by whom?
    b. How are performance standards or passing scores determined?
        i. What evidence or rationale supports it?
    c. What measures are in place to monitor the accuracy of teachers' scoring of the test?

Interpretation of Test Results

14. What guidelines are provided to teachers to help them interpret students' test results?
15. In education, students' test results are supposed to be interpreted in the light of the technical or psychometric properties of the test. Which of these properties are communicated to teachers?
    a. How were these properties documented or determined?

Reporting and Uses of Test Results

16. What are the recommended uses of students' test results in this program?
    a. What constitutes a potential misuse of students' results on the test?
17. What policies guide the confidentiality of students' results on these tests?
18. Any additional comments concerning what have been discussed so far?

Thank You Very Much For Your Time, and Contributions.

APPENDIX C

UNIVERSITY OF CAPE COAST

INSTRUMENT FOR TEACHERS' SURVEY

Respondent's Consent:

The purpose of this questionnaire is to elicit information on teachers' testing practices in the end-of-term district-mandated testing program, prepared by the Center for Performance Monitoring and Evaluation (CePME). Your full participation will help make informed decisions about the District testing program. It would therefore be appreciated if you could provide responses to **all** items on the questionnaire, and do it **honestly**.

You are assured of complete **confidentiality** and **anonymity** of all information provided. **Nothing** will ever be published or reported that will associate your name and/or school with your responses to the survey questions. Therefore, you **should not** write your name, and/or school name on any part of the instrument. Your participation in this study is **completely voluntary**.

Again, questions on this survey instrument have gone through a thorough review by professionals at the University of Cape Coast, and have been declared **ethical** for educational research.

You hereby consent to voluntarily participate in this study by providing responses to items of the various sections of this instrument. Thank You.

SECTION A

TEST PREPARATION PRACTICES

Directions: Indicate with a tick [√] your level of practice on the following activities regarding how you prepare students for the test. Where: *SA = Strongly Agree, A = Agree, D = Disagree, and SD = Strongly Disagree.*

	SA	***A***	***D***	***SD***
1. I notify students in advance of testing.				
2. I motivate students to do their best.				
3. I inform students about the format of the test questions.				
4. I inform students about the coverage of the test.				
5. I inform students about the directions for taking the test.				
6. I inform students about how their performance will be graded.				
7. I inform students about actions that constitute misconduct, and consequences of such acts.				
8. I inform students about the purpose, and uses of test results.				
9. I inform students about appropriate test taking strategies/skills.				
10. I take previous district tests and give them as practice in class.				
11. I modify previous test questions, and practice it with students.				

12. What proportion of the syllabus for an academic term are you able to complete?

[ ] More than 90 percent of the syllabus for an academic term.

[ ] Between 60 and 90 percent of the syllabus for an academic term.

[ ] Less than 60 percent of the syllabus for an academic term.

SECTION B

TEST ADMINISTRATION PRACTICES

Directions: Indicate with a tick [√] your level of practice on the following activities regarding how you administer the district test. Where: *SA = Strongly Agree, A = Agree, D = Disagree, and SD = Strongly Disagree.*

	SA	***A***	***D***	***SD***
13. I ensure that the testing environment is free from distractions.				
14. I assign seating/sitting arrangement for students.				
15. Students start the test promptly, and stop on time.				
16. I inform students about how previous/past students of specific gender (either males or females) performed on the test.				
17. The sitting arrangement prevents students from copying each other's work.				
18. I ensure that students are working in the appropriate sections of the test.				
19. Ambiguous test questions are clarified **only** to students who ask about them.				
20. I make announcements about the test time at regular intervals.				
21. Students are threatened dire consequences if they fail.				
22. Students are told to work faster in order to finish on time.				
23. I, single-handed, supervise more than 25 students				

in one room.

24. I perform other unrelated professional duties, such as scripts marking while administering the district test.
25. I make remarks about quality or quantity of students' work during testing.
26. I take note of any problem or irregularities encountered.
27. I make special adjustments or changes in the testing process for students with special needs and/or injuries.

Directions: Indicate with a tick [√] the level of practice on the following activities in your school regarding the security of test materials. Where: *A = Always, VO = Very Often, NO = Not Often, and N = Never.*

	A	***VO***	***NO***	***N***
28. The tests are administered during the official scheduled testing time.				
29. Prior to testing, the test materials are kept in a secure place where **unauthorized** teachers cannot have access to it.				
30. Prior to testing, the test materials are kept in a secure place where students cannot have access to it.				
31. Teachers discuss the test questions prior to testing.				
32. Teachers or school heads duplicate test or answer booklets, including photographing, photocopying, or copying by hand.				

33. Teachers leave testing materials unattended to, before, during or after testing.

SECTION C

GRADING AND INTERPRETATION OF TEST PERFORMANCE

Directions: Please tick [√] the extent to which a student's grade on the district test is influenced by the following factors. Where: *VE = Very Large Extent, FE = Fairly Large Extent, VL = Very Little, and N = Not At Al.*

Student's grade on the district test is influenced by:	***VE***	***FE***	***VL***	***N***
34. The language (i.e., local dialect or English) student speaks in class/school.				
35. Student's involvement in religious activities in class/school.				
36. Student's gender.				
37. Student's efforts to learn or improve academically.				
38. Student's relationship with the teacher.				
39. Student's moral virtues such as indecent language and dress code.				
40. Student's class attendance.				
41. Student's participation in class discussions.				
42. Bonus marks for extra work or responses on the district test.				

Directions: Which of the following factors are considered in interpreting student's performance on the district test?

In interpreting students' performance on the district test, do you consider:	***Yes***	***No***

43. The level of vocabulary, and sentence structure of the test.
44. The difficulty level of the questions on the test.
45. Sufficiency or adequacy of test time.
46. Student's motivation to perform on the test.
47. Student's ability.
48. Student's test-taking skills or strategies.
49. Student's socio-economic background.
50. Student's opportunity to learn the content of the test.
51. The alignment or association of the test content to curriculum and instruction.
52. Irregularities or problems encountered during testing.

SECTION D

REPORTING AND USES OF DISTRICT TEST RESULTS

Directions: Indicate with a tick [√] your level of practice on the following activities regarding how you report students' results. Where: *SA = Strongly Agree, A = Agree, D = Disagree, and SD = Strongly Disagree.*

	SA	***A***	***D***	***SD***
53. I use clear and simple language on report cards.				
54. I report district test results in a timely fashion.				
55. I discuss with students ways of improving their achievement.				
56. I report separately, on report cards, student's achievement, effort, and attitude.				
57. I modify reporting procedures for students with special needs and/or injuries.				

58. I publicly display students' test results with visible grades.
59. I share specific examples of students' strengths and weaknesses, on report cards.

Directions: Indicate with a tick [√] your level of practice on the following statements regarding how you use students' test results. Where: *SA = Strongly Agree, A = Agree, D = Disagree, and SD = Strongly Disagree.*

I use students' results on the district test to:	***SA***	***A***	***D***	***SD***
60. Determine student grades.				
61. Promote/retain students in class.				
62. Diagnose students learning needs.				
63. Make classroom instructional decisions.				
64. Assess my teaching effectiveness.				
65. Give feedback to parents.				
66. Give feedback to students.				
67. Evaluate students' progress.				
68. Rank students by achievement level.				
69. Identify students in need of extra support.				
70. Group students by achievement level.				

71. Decisions about students (i.e., uses made of the test results) are based **only** on students' performance on the district test?

[ ] Yes.

[ ] No.

SECTION F

BACKGROUND INFORMATION

72. Please indicate your level of training in educational measurement. (You may tick more than one option).

[ ] I have had a course or school training in Educational Measurement.

[ ] I have had workshop/in-service training in Educational Measurement.

[ ] None.

Thank You Very Much For Your Time, and Contributions.

APPENDIX D

CRONBACH'S ALPHA RELIABILITY TEST

Reliability Statistics

Case Processing Summary

		N	%
Cases	Valid	251	100.0
	Excluded[a]	0	.0
	Total	251	100.0

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

Reliability Statistics

Cronbach's Alpha	N of Items
.819	72

APPENDIX E

INTERVIEWEE'S CONSENT FORM

This is a research project intended to elicit information on fair testing practices in the end-of-term district-mandated testing programmes, prepared by the Center for Performance Monitoring and Evaluation (CePME), in the Ashanti Region of Ghana. The information collected in this study will help make informed decisions about the fairness of the district-testing program.

I appreciate that this is a busy period for you and crave your indulgence to bear and cooperate with me. If you need to take a break for rest or to take care of something, please inform me and I will stop when you wish.

You are assured of complete **confidentiality** and **anonymity** of all information provided. **Nothing** will ever be published or reported that will associate your name and/or contact with your responses to the questions. Only the researcher has access to information concerning your identity, and has promised to keep your data confidential.

Contact Address: Emmanuel Boakye, University of Cape Coast.
Tel: 054-4623716/027-8535462/020-5789954.
Email: bem2886@yahoo.com

Do you [Name……] agree to participate in the study?

[ ] Yes

[ ] No

Signature/Thumbprint of Respondent: ……