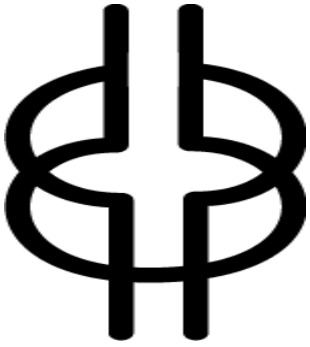
Ghana Journal of Education: Issues and Practice (*GJE*)



NYANSAPO – "Wisdom Knot"

Symbol of wisdom, ingenuity, intelligence and patience

Ghana Journal of Education: Issues and Practices (GJE)
Vol. 3, December 2017, pp. 138 – 158

The Internal Consistency Reliability of Scores in Diploma in Basic Education Examination conducted by the Institute of Education, UCC, Ghana

Jonathan Osae Kwapong
College of Education Studies, University of Cape Coast, Cape Coast, Ghana
kwapjoges@yahoo.com

Abstract
The purpose of the study was to determine the internal consistency reliability of the scores that students of Colleges of Education in Ghana obtain for the Diploma in Basic Education examination. The stratified random sampling technique was employed to select the scripts of 600 students for each examination paper from 12 Colleges of Education. The courses selected for the study were English (FDC121), Mathematics (FDC122) and Integrated Science (FDC124) whose examination was conducted in the second semester of the 2015/2016 academic year. Cronbach's alpha was computed for the internal consistency reliability. The results showed a reasonably strong internal consistency indicating that candidates' performance is reasonably consistent across items on each test paper and the items constituting a paper, to some extent, are homogeneous. However, it was observed that there was the need to improve upon the internal consistency of the scores. Consequently, it was recommended that the Institute of Education intensifies the orientation on test construction for item writers and conditions in the testing environment should be improved for efficient administration of the examinations.

Key words: internal consistency, cronbach alpha, standard error of measurement, item homogeneity, reliability, replications, true scores, error scores

Introduction

Whenever a test is administered, the test user would like some assurance that the results could be replicated if the same individuals were tested again under similar circumstances (Crocker & Algina, 1986). It is this consistency (reproducibility) of test scores that is called reliability. In practical terms, reliability is the degree to which individuals' deviation scores, or z-scores, remain relatively consistent

over repeated administration of the same test or alternate forms (Crocker & Algina, 1986). Subsequently, Haertel (2006) opined that the concern of reliability is to quantify the precision of test scores and other measurements. Haertel, further explained that reliability is concerned solely with how the scores resulting from a measurement procedure would be expected to vary across replications of that procedure. This suggests that test scores from a single administration may not be wholesome. In view of this, Spearman (1913) cited in Crocker and Algina (1986) described test scores as fallible measures.

Spearman (1913) cited in Crocker and Algina (1986) went on to explain that any observed score could be envisioned as a composite of two hypothetical components- a true score and an error score which is expressed mathematically as X=T+E where X represents observed or raw score, T represents the true score and E the error score. From the equation, it can be deduced that the greater the error (E) the wider the difference between the observed score and the true score and the smaller the error the less the difference between the observed score and the true score. The latter is the wish of every test developer and user for the greater the uncertainty associated with the result of measurement, the less confidence should be placed on the measurement (Haertel, 2006). Since both the test developer and user expect the confidence people place on the decisions that arise out of the use of the test to be high, they would like the error associated with the test result to be relatively low. This corroborates Miller, Mclntire and Loveler's (2011) definition that a reliable test is one that can be trusted to measure each person approximately the same way every time it is used.

According to AERA, APA and NCME (2014), a true score is a hypothetical error-free value that characterises the variable being assessed. It is conceptualised as the hypothetical average score over an infinite set of replications of the testing procedure. In other words, the true score is the mean or expected value, of an examinee's observed scores obtained from a large number of repeated testings (Crocker & Algina, 1986). This means that the scores obtained in the different replications are not the same and that there may be difference between the true score and the score obtained by an individual on a single administration. This difference between the true score and the observed score constitutes the error score. That is X-T=E. It is on this basis that Crocker and Algina defined the error of measurement as the

discrepancy between an examinee's observed test score and his or her true score.

Diploma in Basic Education (DBE) is a programme run by the Colleges of Education in Ghana. The programme leads to the award of DBE certificate which qualifies one to teach in Basic schools in Ghana (KG1 to JHS3). A DBE score is a composite of two scores. These are the internal score which is conducted and scored by the college (continuous assessment) and the external score (end of semester), which is conducted and scored by the Institute of Education of the University of Cape Coast (UCC).

The Institute of Education has put in place a structured process of marking the scripts of the candidates. The examiners for the marking are tutors from the Colleges of Education. The Principal of each college selects representatives for each course offered in the college for appointment by the IOE. The marking is conducted in conference and the examiners are put in groups of three or four under a team leader selected among the examiners based on his/her experience. The chief examiners who are university lecturers prepare marking schemes for their respective course papers.

The marking begins with coordination of the examiners of the marking scheme. During the coordination, the chief examiner of each paper leads the team of examiners to thoroughly discuss the marking scheme. Where there are disagreements with the scheme, the examiners deliberate and arrive at a consensus. The outcome of the scheme at the end of the coordination becomes the accepted scheme for the marking. When the assistant examiners mark, the marked scripts are vetted by the team leaders who record the marks obtained by each candidate on broadsheets.

The marks on the broadsheets are crosschecked by checkers with the marked scripts. Errors detected are corrected before the scores are keyed into the computer programme. The scores are scaled down to 60% and added to the internal component of 40% to obtain the composite of the DBE scores. These are printed out for another checking to ensure that scores from the corrected broadsheets have accurately been imputed. From the eventual scores, grades are assigned for each candidate and based on the grades obtained for all the prescribed courses a student's final performance for the programme is

determined. That is, whether the student qualifies to be certificated as a teacher or not.

In spite of the structures put in place by the Institute of Education to ensure error-free scores, measurement error cannot be avoided totally. For example, candidates might have cheated but succeeded without been noticed by invigilators. Or candidates might have guessed correct answers. Such situations lead to random errors andis may reduce the usefulness of the test scores. Literature shows that the error component of an observed score arises from a number of factors. These include content sampling, inattention on the part of the student, guesses, misreading of items, variations in testing conditions, administration errors, fluctuations in the level of the examinee's motivation, levels in distractions and variations in scoring due to scorer subjectivity (AERA, APA & NCME, 2014; Crocker & Algina, 1986; Haertel, 2006; Fraenkel, Wallen & Hyun, 2012). In view of this, the DBE obtained scores may also be contaminated.

These sources of error are categorized into systematic and random errors. Systematic measurement errors are those which consistently affect an individual's score because of some particular characteristic of the person or the test that has nothing to do with the construct being measured (Crocker & Algina, 1986). For example, if a candidate at the JHS level gets a question on integration, which is not included in the JHS syllabus, in a mathematics test wrong and provided no revision takes place afterwards, anytime the test is conducted again the candidate would have that item wrong. This item would not affect the candidate's performance over replications of the test. In this case the candidate's scores over replications will remain the same, hence any variations in scores attributable to this error is systematic. On the other hand, random errors of measurement affect an individual's score on the basis of chance. Random errors may be caused by guessing, distractions in the testing room, administration errors etc. Random errors may affect an examinee either in the positive or negative direction.

Goforth (2015) noted that reliable measure is one that contains no or very little random measurement error. This implies that anything that might introduce arbitrary or haphazard distortion into the measurement process, results in inconsistent measurements. However, Goforth observed that reliable measure needs not be free of systematic error in order to be reliable; it only needs to be consistent.

Consequently, between the two types of test errors, psychometrics are more concerned with the random errors.

Although systematic errors do not result in inconsistent measurement they may cause test scores to be inaccurate and thus reduce their practical utility. Random errors, on the other hand, reduce both consistency and practical utility of the test scores (Crocker & Algina, 1986). If it is found that test scores are not consistent, their usefulness would be in doubt and prospective users would lose confidence in them. It is, therefore, the expectation of test developers and users that the error component of the observed score of a test is reduced in order to make the observed score closer to the true score. This expectation is realized when reliability is high. This is because reliability is high if the scores of each person is consistent over replications of the testing procedure and is low if the scores are not consistent over replications (AERA, APA & NCME, 2014). Consequently, Crocker and Alginer opined that test developers have a responsibility to demonstrate the reliability of scores obtained from their tests.

Measurement experts have identified a number of procedures for estimating reliability. Miller, Mclntire and Lovler (2011), Haertel (2006) and Allen and Yen (1979) identified four methods of checking reliability. These are test-retest, parallel forms, internal consistency and scorer reliability or agreement. However, Crocker and Algina (1986) categorized reliability estimates into two depending on the number of administrations. The first one consists of procedures requiring two test administrations which include alternate forms, test-retest and test-retest with alternate forms. The second category involves procedures requiring a single test administration. The latter consists of split-half method and methods based on item covariances. Crocker and Algina observed that both methods yield an index of internal consistency.

In spite of the varied methods, the most appropriate procedure to adopt when determining the reliability of a test depends on the intended use of the test scores (Crocker & Algina, 1986) and the population being tested (AERA, APA & NCME, 2014). Consequently, Crocker and Algina suggested that the test developer should identify the sources of measurement error that would be most detrimental to useful score interpretation and design a reliability study that permits

such errors so that their effects can be assessed. This suggests that not all the reliability methods are suitable for a particular study.

Parallel/alternate forms reliability estimates may be ideal but are often difficult to obtain. Even if parallel/alternate forms are available, there may be resistance to the burden of repeated testing, especially in school settings (Haertel, 2006). Thus, there has been an abiding interest in methods for estimating reliability from a single administration of a single test form.

The internal consistency is explained as coefficients based on relationships/interactions among scores derived from individual items or subsets of the items within a test, with all scores accruing from a single administration. It is a measure of how related the items are to each other (Miller, Mclntire & Lovler, 2011). In other words, if a test is internally consistent, then the items in that test are really measuring the same thing (Banyard & Grayson, 2000). In view of this Miller, Mclntire and Lovler argued that if a test is internally consistent then knowledge of how a person answered one item on the test would provide information that would help correctly predict how the person answered another item on the test. In another sense, internal consistency estimates are designed to determine how consistently examinees' perform across items or subsets of items on a test form. In this way, the test user can estimate how consistently examinees' performance on the test can be generalized to the domain of items that constituted the test form. Crocker and Algina (1986) observed that if examinees' performance is consistent across subsets of items within a test, the examiner can have some confidence that the performance would be generalised to other possible items in the content domain.

All internal consistency estimation procedures yield values that are functions of the correlation between separately scored parts of a test (Crocker and Algina, 1986). Crocker and Algina further posited that when examinees perform consistently across items within a test, the test is said to have item homogeneity and such items measure the same type of performance or represent the same content domain. In addition, such items must also be well written and be free of technical flaws that may cause examinees to respond on some basis unrelated to the content.

Crocker and Algina (1986) advocated for the internal consistency to be always examined. This is because its coefficient is an index of both item homogeneity and quality. They, therefore, cautioned

item writers and users to be wary of conditions that will cause examinees not to perform consistently across items on a test and subsequently reduce the internal consistency. Crocker and Algina identified conditions that make examinees not to perform consistently across items to include:

1. When items on a single administration are drawn from diverse areas;
2. When items are drawn from single area but some items test major concepts and some others are based on minor points;
3. If some of the items are poorly written to the extent that examinees may misinterpret the questions or answer to the degree of their testwiseness rather than their knowledge.

The most widely known method using the internal consistency yields a split-half reliability estimate (Allen & Yen, 1979). With the split-half method the test is divided into two parts which are alternate forms of each other (Allen and Yen, 1979; Miller, Mclntire & Lovler, 2011). The individual scores on the two halves are then compared. Allen and Yen suggested that attempts should be made to choose these parts so that they are parallel or essentially tau-equivalent. This means that the two halves must be equivalent in length and content for this method to yield an accurate estimate of reliability (Miller, Mclntire & Lovler, 2011; Haertel, 2006). Consequently, Allen and Yen posited that if the halves are parallel, the reliability of the whole test is estimated using the Spearman-Brown formula. However, if the halves are essentially tau-equivalent, coefficient alpha (α) can be used to calculate the reliability of the entire test.

According to Miller, Mclntire and Lovler (2011) the best way to split the test is to use random assignment to place each question in one half or the other. Miller, Mclntire and Lovler, explained that the random assignment is likely to balance errors in the score that can result from order effect, difficulty and content. Another way to measure internal consistency is to compare individual scores on all possible ways of splitting the test into halves. This method compensates for any error introduced by lack of equivalence in the two halves (Miller, Mclntire & Lovler, 2011). Consequently, KR-20 formula was proposed for computing the internal consistency of tests whose questions are

dichotomously scored (Kuder and Richardson,1939) and Cronbach (1951) also proposed coefficient alpha that calculates internal consistency for items that have more than two possible responses.

Computation of alpha is based on the reliability of a test relative to other tests with same number of items, and measuring the same construct of interest (Hatcher, 1994). According to Goforth (2015) Cronbach's alpha is a measure used to assess the reliability, or internal consistency, of a set of scale or test items. Goforth further explained that the reliability of any given measurement refers to the extent to which it is a consistent measure of a concept, and Santos (1999) observed that alpha is an index of reliability associated with the variation accounted for by the true score of the "underlying construct". In fact, Cronbach's alpha determines the internal consistency or average correlation of items in a survey instrument to gauge its reliability and it is one way of measuring the strength of that consistency.

Cronbach's alpha is computed by correlating the score for each scale item with the total score for each observation (usually individual survey respondents or test takers), and then comparing that to the variance for all individual item scores. The resulting α coefficient of reliability ranges from 0 to 1 in providing this overall assessment of a measure's reliability. If all of the scale items are entirely independent from one another (i.e., are not correlated or share no covariance), then $\alpha = 0$; and, if all of the items have high covariances, then α will approach 1 as the number of items in the scale approaches infinity. In other words, the higher the α coefficient, the more the items have shared covariance and probably measure the same underlying concept.

Although the Standards for what makes a "good" α coefficient are entirely arbitrary and depend on one's theoretical knowledge of the scale in question, many methodologists recommend a minimum α coefficient between 0.65 and 0.8 (or higher in many cases); α coefficients that are less than 0.5 are usually unacceptable, especially for scales purporting to be unidimensional (Goforth, 2015). Other literature had suggested that a coefficient alpha of 0.70 is adequate for reliability of tests (Nunnally, 1978; Cascio, 1991; Schmidt, 1996). However, Cascio (1991) suggested that reliability should be greater than 0.90 and (Green, Salkind & Akey, 2000) in the Statistcal Procedure for Social Sciences (SPSS) corroborating with this view, noted that the coefficient alpha of 0.89 is an indication that the scale scores are

reasonably reliable. However, in determining the adequacy of the internal consistency one must consider the standard error of measurement (SEM) as it gives a realistic estimate of how much error exists in an individual's obtained score (Miller, Mclntire & Lovler, 2011).

Considering the role teachers play in the education of the child, parents, stakeholders and of course, the general public look for evidence to boost their confidence in the teachers who teach their wards. There have been indicators that give concern for stakeholders to be inquisitive about the reliability of the scores that qualify teachers to teach children at the basic schools.

Anamuah-Mensah, Mereku and Ghartey-Ampiah (2008) reporting on the 2007 edition of Ghana's participation in Trends in International Mathematics and Science Study (TIMSS) observed that Ghana's performance was poor. The test was conducted for Grade 8 (JHS 2 in Ghana) students of forty-four countries in Mathematics and Science with Ghana ranking 43rd in Mathematics and last in Science. The students who represented Ghana were a sample of students taught by teachers who had completed the DBE programme and certificated by UCC. If this performance at the International level is anything to go by, then one will wonder the kind of marks that qualified those teachers to obtain the certificates to teach. It was not surprising that, making reference to Ghana's performance at TIMMS during his inaugural lecture, Ghartey-Ampiah (2016) wondered at the type of the content knowledge possessed by these teachers.

Added to this are studies that have questioned the credibility of Senior School Certificates awarded by West Africa Examination Council (WAEC) in Nigeria. Achigbe and Bassey (2012) reported that the Nigerian educational scene had been riddled with a lot of controversies with the approval of a new and indigenous examining body, the National Examination Council (NECO), in 1999 to conduct the Senior School Certificate Examination (SSCE) alongside the more experienced WAEC. They observed that such action had raised the consciousness of stakeholders and agitations of the general public on the credibility of the SSCE being conducted by WAEC. In support of this view is the study of Ajuonuma and Mkpa (2009) which indicated that the credibility of public examinations conducted by WAEC in Nigeria was being queried. They observed that WAEC's certificates

were being subjected to public scrutiny locally and many foreign countries. Ajuonuma and Mkpa, therefore, wondered if the universities have been admitting the right students.

The Institute of Education, as an examination body, should not wait for the public to strike before it puts its house in order. It will be useful for it to learn and avoid the WAEC's experience in Nigeria. It is for these reasons that this study would want to examine the internal consistency of the DBE examination scores on which decisions are taken about the certification of students of the Colleges of Education in Ghana. Hence, the problem of the study is the internal consistency reliability of DBE external scores obtained by the Institute of Education which are used to determine the qualification of students of the Colleges of Education as teachers.

Purpose of the study

The purpose of the study is to examine the internal consistency reliability of the DBE external examination scores obtained by the Institute of Education, UCC. Specifically, the study will examine;

1. The internal consistency of the external examination scores of the DBE.
2. The relationship of the items constituting the papers in measuring common constructs of the DBE examination papers.

Research question

The study was guided by the following research question.

1. What is the internal consistency reliability of the DBE external examination scores?

Methodology

Research Design

The study is mainly a descriptive survey design. Borg and Gall (1983) described descriptive studies as those aimed at finding out the state of objects. Descriptive survey is an attempt to obtain data from members of a population or a sample to determine the current status of that population with respect to one or more variables (Burnham, Gilland, Grant, & Layton-Henry, 2004; Fraenkel, Wallen & Hyun, 2012). A survey is often conducted to obtain description of a particular

group of individuals (Gravetter & Forzano, 2006). This design is suitable for the study because data were collected from the current natural setting of colleges of education to obtain the desired information. The study was conducted using a sample from the population of colleges of education in Ghana. Gravetter and Forzano observed some advantages of a survey to include its flexibility and efficiency in collecting a wide variety of information about different variables. However, one disadvantage has been noted to be its low response rate and non-response bias. In order to address such weaknesses the researcher made a number of follow-ups to the colleges for the collection of the data.

Population

The population of the study consisted of all students who offered first year second semester core courses in English (FDC121), Core Mathematics (FDC122) and Integrated science (FDC124) in all public and private Colleges of Education in Ghana for the 2015/2016 academic year. As at the 2015/2016 academic year, there were thirty eight (38) public and eight (8) private Colleges of Education in Ghana. English, Mathematics and Integrated Science were selected because they were core courses taken by all students offering the General Programme which is offered in all the 46 Colleges of Education. The total number of students was 13,352 (Report on the 2015/2016 first year end-of-second semester examination results).

Sample and Sampling Techniques.

The stratified random and simple random sampling techniques were adopted in selecting the sample. The study was conducted in 12 Colleges of Education constituting 26.1% of the colleges. Using the stratified random sampling technique, two colleges were randomly sampled from each of the five zones of public Colleges of Education in Ghana. These zones were Northern, Ashanti/BA, Eastern/Greater Accra, Volta and Central/Western Zones. In addition to these, two private Colleges of Education were randomly selected.

For each College of Education Zone, the names of all the colleges were written on pieces of paper, folded and placed in a bowl. The researcher shook the bowl vigorously and asked a twelve year old girl to pick two at random with replacement. This was done to ensure

equal chance of selection. The two selected colleges from each zone constituted the sample for the zone. The same process was used to select the sample for the private colleges. In each college, a sample of 50 students' marked scripts for each course was randomly selected for the study. Fifty scripts were packed in each envelope. Any of the fully packed envelopes for each of the selected courses from each of the sampled colleges was randomly selected. This means that 600 scripts (4.5%) were sampled for each course.

Research Instrument

The main instrument used in the study was document analysis guide. A document is an instrument in language which has, as its origin and for its deliberate and express purpose to become the basis of, or to assist, the activities of an individual, an organisation or a community (Webb & Webb, 1932 cited in Burnham, Gilland, Grant & Layton-Henry). Webb and Webb cited in Burnham, Gilland, Grant and Layton-Henry opined that the social investigator must insist on the original document or an exact verbatim copy and that the aim of the investigator must be to consult the original source. The instrument sought to examine documents/records of students' external examination scores of English (FDC 121), Mathematics (FDC 122) and Integrated Science (FDC 124). The Integrated Science and Mathematics papers consisted of objective (dichotomously scored) and essay items and the English paper consisted of five (5) sections (A, B, C, D and E) with seven questions. Candidates were to answer one question out of two from sections A and E and answer all the questions in Sections B, C and D. Consequently, for the English, candidates answered five items in all. The Mathematics (FDC 122) paper was composed of two sections (A and B). Section A consisted of 15 compulsory objective items and Section B was made of five items out of which candidates were to answer three. In this section candidates were required to show working. The integrated Science paper, on the other hand, consisted of four sections (A, B, C and D). Section 'A' part had a 40 dichotomously scored items. Each of Sections B, C and D consisted of two subjective items from which candidates were to answer one from each section.

One advantage of examination of records is that it is relatively quick and complete since all the relevant information is usually stored in one location (Borg & Gall, 1983). Borg and Gall cautioned that the

use of the technique involves invasion of subjects' privacy. In view of this clearance was sought from the appropriate authorities of the Colleges of Education, Institute of Education and Institutional Review Board (IRB) of the University of Cape Coast.

Data Analysis

The data were analysed by adopting the Cronbach's alpha. The Statistical Programme for Social Science (SPSS) was employed to compute the statistics. The internal consistency of the external scores of the three papers was computed using the Cronbach's alpha. Cronbach's alpha was found suitable to determine the internal consistency of the scores of the target courses due to the structure of the test papers which consisted of both objective and essay items. Crocker and Algina (1986) observed that Cronbach.'s alpha can be used to estimate the internal consistency of items which are dichotomously scored or items which have a wide range of scoring weights including essay items.

Results and Discussions

Research question: What is the internal consistency reliability of the DBE external examination scores?

To answer the research question, the internal consistency reliabilities of the external papers of the selected courses (English, FDC121; Mathematics, FDC122; and Integrated Science, FDC 124) were computed by the researcher using the Cronbach alpha. For the FDC 121 each of the five sections was considered as a subtest of the test paper with each section consisting of a maximum of 20 marks. This means that FDC 121 was made up of five subtests.

Section A of FDC 122 had a total of 40 marks and each of the three questions of Section was B allotted 20 marks. In order to have uniform scores for all the items, the researcher divided section A scores by two. In effect, for the FDC 122 four questions or subtests with 20 marks each were used to compute the internal consistency. The distribution of scores in Integrated Science was the same as that of Mathematics and so FDC 124 was also composed of four subtests. The results of the Cronbach's alpha are presented in Table 1.

Table 1: Results of Cronbach's alpha for the three papers

Paper title	Paper code	Valid cases	No. of subtests	Coefficient alpha (α)	SEM of the single adm
English	FDC 121	497	5	0.66	6.51
Mathematics	FDC 122	395	4	0.69	7.05
Int. Science	FDC 124	591	4	0.66	5.95

From Table 1 coefficient alpha of the three papers range between 0.66 and 0.69. These approximated to one decimal place gives 0.7 with SEM ranging between 6 and 7. It can, therefore, be deduced that the internal consistency of the external DBE examination scores may be considered satisfactory. This is because the results are in conformity with other literature that a coefficient alpha of 0.70 is adequate for reliability of tests (Nunnally, 1978; Cascio, 1991; Schmitt, 1996). To buttress this, Goforth (2015), observing that standards for what makes a good coefficient alpha are controversial, noted that many methodologies recommend a minimum alpha coefficient of 0.65 and that alpha coefficient of less than 0.5 is unacceptable. The results further conform with the American National Election Study scale in 2008 cited in Goforth (2015), whose coefficient alpha of 0.67 was considered as reasonably strong. The results therefore suggest a reasonably strong internal consistency reliability of the DBE test scores.

The fact that the range of internal consistency coefficients for the three papers is very small (0.66 – 0.69) suggests that the internal consistency of the scores of the other DBE papers might hover around the same range of values. If this is the case, then, the results depict that candidates' performance is consistent across items on the DBE papers. This implies that the items in the respective papers, to a very large extent, measure common or related constructs (concepts). This is an indication that one can predict a candidate's performance on an item from the candidate's performance on another item. Miller, Mclntire and Lovler (2011) noted that if a test is internally consistent then knowledge of how a person answered one item on the test would provide information that would help correctly predict how the person answered another item on the test. It, therefore, suggests that the test item writers follow a common standard in developing the items and the scoring

procedure is uniform across examination papers. Consequently, one can conclude that the methods the Institute of Education has been adopting in developing items and the scoring procedures for the DBE examinations are, to some extent, effective.

However, this observation may not be all that accurate because the alpha coefficient of 0.66, SEM=6.51 for English, 0.66, SEM 5.95 for Integrated Science and 0.69, SEM=7.01 for Mathematics are just at the fringes of the minimum alpha coefficient considered as satisfactory. Given the view of the minimum acceptable alpha coefficient of 0.9 (Cascio, 1991) and 0.89 (Green, Salkind and Akey, 2000), the internal consistency reliability coefficients could be considered as low. This is buttressed by the high standard error of measurement (between 6 and 7) (Table 1). It means that the error margin is high with regard to the difference between the true score and the observed score. For example, if a candidate scored 70 (raw score) in Mathematics, and given a 95 confidence interval or 0.05 level of significance, the true score is likely to be located between 56.28 and 83.72 ($70 \pm 1.96 \times 0.7$). This illustrates that there is 95% chance that the true score in Mathematics of the candidate whose raw score is 70 lies between 56.28 and 83.72, an interval of 26.5. The interval is too wide and illustrates a wide variation between the true score and the observed score. This is in line with Santos' (1999) observation that alpha is an index of reliability associated with the variation accounted for by the true score of the "underlying construct." In fact the results indicate that the variation of the observed score from the location of the true score is too wide with regard to the SEM.

Studies show that the error component of an observed score arises from a number of factors. These include content sampling, inattention on the part of the student, guesses, misreading of items, variations in testing conditions, administration errors, fluctuations in the level of the examinee's motivation, levels in distractions and variations in scoring due to scorer subjectivity (AERA, APA and NCME, 2014; Crocker & Algina, 1986; Haertel, 2006; Fraenkel, Wallen & Hyun, 2012). Crocker and Algina also noted that low internal consistency is likely to result from the following.

1. That the items on each paper might have been drawn from diverse (unrelated) areas.

2. The items might be drawn from single area or while some items test major concepts others are based on minor points.
3. Some of the items were poorly written to the extent that examinees might have misinterpreted the questions or answer to the degree of their testwiseness rather than their knowledge.

It then follows that if the internal consistency of the DBE papers is low, then they might have been caused by some, if not all of these factors and the scores may contain some amount of errors.

Conclusion and Recommendations

The study has shown that the internal consistency reliability of the DBE external examination scores is high (coefficient alpha for the three papers is about 0.7) and that the performance of candidates on the examination papers is reasonably consistent across items of the DBE papers. This implies that the items in the respective papers, to some extent measure the same construct. However, the coefficient alpha range of 0.66-0.69 with high SEM (6-7) indicate errors in the observed scores. Based on the identified factors that generate errors in observed scores by measurement experts (AERA, APA & NCME, 2014; Crocker & Algina, 1986; Haertel, 2006; Fraenkel, Wallen & Hyun, 2012), the following are recommended in order to improve upon the internal consistency reliability of the DBE external examination scores.

1. The item writers must pay particular attention to the construction and use of test specification table during the process of test construction by writing items which relate content with objective effectively.
2. Each item should reflect major content or topic. To a very large extent, items drawn from trivial content must be avoided.
3. The information carried out in each item must be as clear as possible such that candidates will depend on knowledge acquired rather than guesses and testwiseness in answering questions. Items should be devoid of any ambiguity.
4. Each examination paper must be moderated by a panel comprising, at least, a subject expert, language specialist and an assessment expert. This will ensure that issues bordering on content, clarity, language and principles of test construction are effectively addressed.

5. In administering the test the testing environment should be given utmost consideration. Adequate attention must be paid to illumination of the examination room to ensure that there is enough light that will enable each candidate see clearly to read and understand each question. Also, the room must be devoid of any distractive sounds that may distract candidates' attention in the course of answering the questions. Furthermore, the seating arrangements should be spacious enough to avoid candidates from obtaining any form of assistance from each other. These and other measures will ensure that candidates answer the questions with no or little interference thereby reducing the error margin associated with their observed scores. Consequently, with the items being homogeneous, internal consistency will be high.
6. Finally, further studies could be conducted involving other courses beside Mathematics (FDC 122), English (FDC 121) and Integrated Science (FDC 124) in any of the College of Education zones or the same courses in other zones to confirm or nullify the results of this study.

References

Achigbe, M. O., & Bassey, E. O. (2012). The effect of senior school certificate examination conducted by WAEC and NECO on public perception of the examinations in Nigeria. *Journal of Educational Assessment in Africa,7*,77-87.

AERA, APA & NCME. (2014). *Standards for Educational and psychological testing*. Washington, D.C: American Educational Research Association.

Ajuonuma, J. O., & Mkpa, N. D, (2009). The predictive validity of West African Senior Secondary Certificate Examination for academic performance in the university. *Journal of Research in National Development, 7*(1), 1-8.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Illinois; Waveland Press Inc.

Ampiah, J. G. (2016, March). *The pre-tertiary science education in Ghana: curriculum, teaching resources and students' performance.* Inaugural lecture University of Cape Coast.

Anamuah-Mensah, J., Mereku, D. K. & Ampiah, J. G. (2008). *Trends in International Mathematics and Science Study*. Accra; Adwinsa Publications.

Banyard, P. & Grayson, A. (2000). *Introducing psychological research*. (2nd ed.). New York, Palgrave Macmillan.

Borg, W. R. & Gall, M. D. (1983). *Educational Research*. (4th ed). New York; Longman Inc.

Burnham, P. Gilland, K. Grant, W. & Layton-Henry, Z. (2004). *Research Methods in Politics*. Hampshire; Palgrave Macmillan.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ:Prentice Hall.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Florida; Holt Rinehart and Winston Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 297-334.

Fraenkel, J. R., Wallen, N. E. & Hyun, H. H. (2012). *How to design and evaluate research in education.* (8th ed.). New York; McGraw-Hill Company.

Goforth, C. (2015). *Using and interpreting Cronbach's alpha*. University of Virginia Library Research Data Services + Sciences.

Gravetter, F. J., & Forzano, L. B. (2006). *Research methods for the behavioural sciences* (2nd ed.). Belmont; Wadsworth.

Gravetter, F. J., & Wallnau, L. B. (2005). *Essentials of Statistics for behavioural Sciences*. Belmont; Wadsworth.

Green, S. B., Salkind, N. J. & Akey, T. M. (2000). *Using SPSS for windows, analysing and understanding data*. (2nd ed.). New Jersey; Prentice Hall.

Haertel, E. H. (2006). *Educational measurement*. (4th ed.) K. L. Brennan (Ed.). West Port: Praeger Publishers.

Hatcher, L. (1994). *A step-by-step approach to using the SAS(R) system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute.

Kuder, G. E., & Richardson, M. W. (1939). The calculation of test reliability coefficients based on the method of rational equivalence. *Journal of Educational Psychology, 30*, 681-687.

Miller, L. A., Mclntire, S. A. & Lovler, R. L. (2011). *Foundations of psychological testing*. (3rd ed.). California; Sage publication Ltd.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of Extension, 2*(37).

Schmidt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.