Journal of Educational Development and Practice (JED-P)
Vol. 5, No. 1, December 2021, pp. 23-39

INTERRATER RELIABILITY STUDY OF THE DIPLOMA IN BASIC EDUCATION EXAMINATION IN ENGLISH, MATHEMATICS AND INTEGRATED SCIENCE CONDUCTED BY THE INSTITUTE OF EDUCATION, UCC IN GHANA

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

Abstract
The study was conducted to determine the interrater reliability (ratter agreement) of the Diploma in Basic Education (DBE) examination conducted by the Institute of Education of UCC in Ghana. The population consisted of 13,352 first year students who were admitted for the 2015/2016 academic year to pursue the DBE programme and offered English, Core Mathematics and Integrated Science. Using the stratified random sampling technique, 600 scripts of each course were sampled from twelve Colleges of Education for the study. The Pearson Product Moment Correlation Coefficient and paired samples t-test were used for the analyses. The results showed a high interrater reliability in the three courses English (r=0.819 at 0.05α), Mathematics (r=0.878 at 0.05α) and Integrated Science (r=0.867at 0.05α) courses. In addition, the hypothesis testing revealed that the differences between raters in Mathematics and Integrated Science were not significant at 0.05α, p>0.05 indicating that the differences had no impact on total scores. However, in English the differences were found to be significant at 0.05α. It was recommended that the Institute of Education intensifies its coordination sessions with examiners with special emphasis on the English examiners. It was further suggested that team leaders should be more vigilant with vetting of scripts.

Keywords: interrater reliability, scorer agreement, total score, error scores

Introduction

Diploma in Basic Education (DBE) was a three year teacher training programme run by Colleges of Education (COE) in Ghana (Institute of Education, UCC, 2005). The products after successful completion of the programme were awarded Diploma Certificates by the University of Cape Coast (UCC). This certificate qualified one to teach in Basic schools in Ghana (KG1 to JHS3). The University was, therefore, responsible for the conduct of the assessments leading to the award of the certificates. The assessment process was managed by the Institute of Education of UCC on behalf of the University.

A DBE examination score is a composite of two scores. These were the internal score which is obtained from internal assessment conducted and scored by the college (continuous assessment) and the external score (end of semester), is obtained from external assessment conducted and scored by the Institute of Education of the University of Cape Coast (UCC).

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

The marking began with coordination of the examiners of the marking scheme prepared by the chief examiner. During the coordination, the chief examiner of each paper led the team of examiners to thoroughly discuss the marking scheme. Where there were disagreements with the marking scheme, the examiners discussed and arrived at a consensus. The outcome of the scheme at the end of the coordination became the accepted scheme for the marking. When the assistant examiners mark, the marked scripts were vetted by the team leaders who recorded the marks obtained by each candidate on broadsheets.

In spite of the lofty arrangements put in place for the marking, one cannot fathom the credibility or otherwise of the scores obtained by the examiners. But Crocker and Algina (1986) observed that 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 and Algina termed this 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) indicated 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 (1904 to 1913) cited in Crocker and Algina (1986) described test scores as fallible measures. Spearman 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, the greater the error (E) the wider the difference between the observed score and the true score. Similarly, 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 many 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.

Measurement error reduces the usefulness of test scores. It limits the extent to which test results can be generalized beyond the particulars of a given replication of the testing procedure. It also reduces the confidence that can be placed on the results from any single measurement. Literature shows that the error component of an observed score arises from several factors including variations in scoring due to scorer subjectivity (AERA, APA & NCME, 2014; Crocker & Algina, 1986; Haertel, 2006; Fraenkel, Wallen & Hyun, 2012).

Assessment and psychology experts contend with two types of errors-random and systematic. Between systematic and random errors, assessment experts 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 reduce both consistency and practical utility of the test scores (Crocker & Algina, 1986). If it is found that test scores are not consistent, its usefulness would be in doubt and prospective users would lose confidence in it. 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 when 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.

Task-to-task variances in the quality of an examinee's performance and ratter-to-ratter inconsistencies in scoring represent independent sources of measurement error. When such a situation occurs in a reliability study it is necessary to indicate which of these sources are reflected in the data. Consequently, the AERA, APA and NCME (2014) reported that when subjective judgment is incorporated in test scoring, evidence should be provided on interrater consistency in scoring and within-examinee consistency over repeated measurements (AERA, APA & NCME, 2014). This idea is corroborated by Anastasi and Urbina (2007) in their observation that one source of error score

variance is scorer variance. This means that some of the errors inherent in observed scores are caused by differences that result from differences in the ratings of different raters for the same test.

Such errors generate unfairness to affected testees, especially where decisions have to be taken on these scores. In the first place, it is likely that some of those who have to fail may pass and some of those who have to pass may fail. In addition, where such results are required for selection, for example, for admission, employment, or award it is likely that some of those who better qualify may be rejected in favour of some of the less qualified candidates.

In a discussion with the coordinator of examination at the Institute of Education of the University of Cape Coast (UCC) in 2014, it was noted that eight candidates applied for remarking of their scripts in Mathematics, Science and English Language courses. After the remarking, the coordinator noted that seven of them had scores less than what they originally obtained. For the eighth candidate, he went and checked the records and observed that the examiner failed to add the score the candidate obtained for section A which was 24. After adding the 24 the candidate obtained grade C+ instead of grade E which is a failure grade. As a result, that paper was not remarked. If this candidate had not applied for remarking, he/she would have failed amidst social and mental torture he/she would have suffered.

What is more worrying is that among the results that were released there might be some who had suffered from error scores but because they were not bold enough or did not have adequate financial resource did not apply for remarking. This means that some of those who passed might have failed and some of those who failed must have passed. Such situation defeats the assessment principle of fairness. Consequently, Anastasi and Urbina (2007) suggested that such a situation requires a measure of scorer reliability or interrater reliability. Interrater reliability, consistency or agreement is the level of consistency with which two or more judges rate the work or performance of test takers (AERA, APA & NCME, 2014). Miller, Mclntire and Lovler (2011) named it as scorer reliability and described it as the amount of consistency among scorers' judgments. It is concerned with how consistent the judgements of scorers are. According to Crocker and Algina (1986) the most flexible and useful approach for estimating interrater reliability is through the application of generalizability and other indices of agreement such as percentage of

agreement among scorers on codes assigned to specific items or sets of items. Although the indices of these methods are informative, Crocker and Algina contended that they are conceptually different from reliability estimates and consequently, advised that they should not be considered as substitutes for reliability estimates.

Scorer reliability can be found by having a sample of test papers independently scored by two examiners (AERA, APA & NCME, 2014). However, depending on the method to be used more than two examiners can do the scoring. The two scores obtained by each test taker are correlated using the Pearson product moment correlation coefficient and the correlation coefficient is a measure of scorer reliability (Anastasi & Urbina, 2007). Fraenkel, Wallen and Hyun (2012), making reference to scorer reliability, observed that "what is desired is a correlation coefficient of, at least, 0.90" p159. This means that for an interrater reliability estimate to be desirable the scorers must agree on, at least, 90% of the scoring. Consistent with this assertion are the studies on Wisconsin Card Sorting Test (WCST) by Axelrod, Goldman and Woodard (1992). In one of the studies three assessors rated the WCST data and using intraclass correlation the researchers obtained correlation coefficients of 0.92, 0.93 and 0.88. The researchers described the interrater reliability as very high agreement.

Again, 'studying the reliability of open ended mathematics items according to the classical test theory and generalizability theory', Neşe and Selahattin (2010) analysed the ratings of four raters using Kendall's Concordance Coefficient obtained coefficients of 0.90 and 0.97 for the paired raters. The researchers consequently, observed that these values support the fact that there is consistency between the raters' scores. In other words, there was high agreement in rating between scorers. Relating the consistency of the scores of the raters, Neşe and Selahattin examined the difference in the mean scores. Using the test of dependent samples analysis of variance, they found that there was a statistically significant difference between mean scores ($F=13.801$, $p<.05$) of the raters. Banyard and Grayson (2000) opined that if independent observers cannot agree on a high percentage of the observations that they make, then the usefulness of the observations is called into question.

Pearson (2009) conducted an interrater reliability study for the New York State involving grades 3 – 8. The study consisted of scoring of Mathematics items by local raters and audit raters. The results, which

were described as very high degree of agreement, had reliability coefficients of 0.99 for all grades. The mean score differences between the two sets of raters also ranged between 0.93 and 0.99 which was observed as close scoring agreement between the local and audit raters. Consequently, Pearson observed that the statistics provided valuable evidence of the reliability and consistency in students' total scores across local and audit scoring methods. Examination of the differences between local scoring and audit scoring also showed a high degree of consistency. The largest mean difference between local and audit scoring was 0.2, occurred in only three of the items in three different grades. Considering that two of the three items are 3-point items, the difference of only 0.2 represents 7% of the maximum points for those items. All other items had an absolute mean difference of 0.1 or less.

Considering the number of examiners who score the DBE examination scripts, it is not quite certain the score obtained by one examiner will be replicated when a different examiner scores the same script. In addition, the fact that not all testees who feel dissatisfied with their results have the zeal and the resources to call for remarking means that there is the likelihood that some of the DBE result published by UCC could be erratic. The problem of the study is, therefore, to investigate the level of interrater reliability or scorer consistency in the ratings of the DBE examination scripts to ensure that if there is any error associated with the DBE examination scores it will not be attributable to the differences in scorers' ratings.

Research questions

To guide the study the following research question was developed:

What degree of difference in rating exists among the different raters in rating the DBE end-of-semester examination scripts?

Hypothesis

The following hypothesis was tested to support the study:

H_o: The ratings of scripts of the DBE end-of-semester examinations by a particular ratter are not significantly different from ratings of other raters.

Methodology

The study is mainly a descriptive survey design. 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, and Layton-Henry, 2004; Fraenkel, Wallen and Hyun, 2012). A survey is often conducted to obtain description of a particular group of individuals (Gravetter and Forzano, 2006). This design is suitable for the study because data was collected from the current natural setting of colleges of education to obtain the desired information. Gravetter and Forzano (2006) observed some advantages of a survey to include its flexibility and efficiency in collecting a wide variety of information about different variables. One disadvantage has been noted to be its low response rate and non-response bias. In order to address such weakness the researcher sought official permission from the Director of the Institute of Education for the collection of data.

The population for the study consisted of all first year students who were admitted to pursue the Diploma in Basic Education programme in the Colleges of Education in Ghana for the 2015/2016 academic year and offered English, Core Mathematics and Integrated Science. In that year there were 38 public and eight private Colleges of Education in Ghana. The total number of first year students for that academic year was 13,352 (Awards Committee of the Institute of Education, UCC's Report on the 2015/2016 first year end-of-second semester examination results).

The stratified random and simple random sampling techniques were adopted in selecting the sample. The study was conducted in twelve Colleges of Education constituting 26.1% of the population. Using the stratified random sampling technique, two colleges were randomly sampled from each of the five public Colleges of Education or PRINCOF Zones in Ghana. In addition to these, two private Colleges of Education were randomly selected. For each PRINCOF 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 ten-year-old girl to pick two of the folded papers at random. The first one chosen was returned before the second was picked. This was done to ensure equal chance of selection. The two selected colleges 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 fifty (50) students' 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 were randomly selected. This means that six hundred (600) scripts (4.5%) were sampled for each course. This means that 1800 scripts were selected for the three courses. Studies had indicated a minimum of 200 sample size would provide acceptable statistics. For example, Schmidt, Hunter and Uzry (1976) observed that sample size of 200 or more may be needed to reflect validity levels of population data accurately at level 90% of the time.

Instruments

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 cited in Burnham, Gilland, Grant & Layton-Henry, 2004). Webb and Webb 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. In corroborating with the Webb and Webb, Frankel, Wallen and Hyun (2012) explained documents to mean any kind of information that exists in some type of written or printed form that may be original works or copies. By implication, the Webb and Webb suggested that documents are the most dependable sources of information.

One advantage of examination of records is that it is relatively easy to use, quick and complete since all the relevant information is usually stored in one location (Borg & Gall, 1983; Fraenkel, Wallen & Hyun, 2012). Gravetter and Forzano (2006) noted that surveys relatively provide easy and efficient means of gathering a large amount of information and found the response rate to be high since in most cases, they are administered in person.

Despite these strengths of the documentary analysis guide, some disadvantages have been identified with this instrument. Borg and Gall (1983) cautioned that the use of the technique involves invasion of subjects' privacy. In view of this, the researcher sought clearance from the appropriate authorities of the Colleges of Education, Institute of

Education, and Institutional Review Board (IRB) of the University of Cape Coast.

Data collection

The researcher sought permission from the Director of the Institute of Education, UCC to access the examination scripts of the sampled colleges before they were scored. The scripts sampled included Mathematics (FDC122), English (FDC121) and Integrated Science (FDC124). All these courses were offered in first year second semester. In all, six hundred scripts were selected for each course. The researcher photocopied the sampled scripts and returned the original script for official marking.

After the official marking, the researcher selected some of the examiners involved with the official marking using the agreed marking schemes they used to do the official marking. The researcher ensured that no examiner marked scripts of his/her college and those they marked during the official marking. The examiners were given adequate time to complete marking in such a way that they would avoid marking under any pressure. After the marking, the scores obtained from the sampled scripts during the official marking were compiled differently from the scores obtained by the selected examiners who marked the photocopied scripts. By this, the researcher obtained two sets of scores emerging from the same scripts. The two sets of scores were used for the computation of the correlation coefficient.

Data analysis

The study was meant to inquire about the differences in rating that exist among the different raters of the DBE end-of-semester examination scripts. Interrater reliability was computed for the analysis. This was meant to determine the stability of the test scores across raters. In other words, it was used to determine the extent to which the ratings of the different raters showed consistency. The scores of the photocopied scripts of the sampled students which were marked by selected examiners were correlated with the scores obtained by official examiners the Institute of Education employed for the marking exercise. The two sets of scores were correlated using the Pearson's Product Moment Correlation Coefficient (r) to determine the correlation between the two sets of scores (Anastasi & Urbina, 2007). The correlation coefficient obtained indicated the interrater reliability.

The paired samples t-test was also computed to determine the level of significance of any difference that may emerge.

Fraenkel, Wallen and Hyun (2012) have suggested that a correlation coefficient of at least 0.90 is desirable. This means that for an interrater reliability estimate to be desirable the scorers must agree on at least 90% of the ratings. In this case a correlation coefficient of 0.9 or greater would indicate that the scores are stable across raters. That is, irrespective of the examiners who marked any paper at least, 90% of the ratings would, generally, be acceptable to other examiners. A positive index of correlation implies that if a ratter rates a script with a high score another ratter is likely to rate the same script with a high score and the reverse holds. On the other hand, a negative index of correlation would indicate that if a ratter rates a script high, a second ratter is likely to rate the same script low.

The study will also test for the statistical difference of the mean scores of the two sets of raters for each paper. The statistical tool that will be employed to execute this is the paired samples t-test. In this case the statistic of interest will be the sig or *p*-value. A *p*-value greater than 0.05 demonstrates that the difference between the mean scores is not statistically significant. This will mean that the null hypothesis will not be rejected. On the other hand, a p-value of less than 0.05 will mean that there is a significant statistical difference in the mean scores of the two raters. Consequently, the null hypothesis will be rejected at 0.05 level of significance.

Results and Discussion

The research question inquired about the degree of differences in rating that existed among the different raters of the DBE end-of-semester examination scripts. Interrater reliability was computed to answer this research question. This was meant to determine the level of stability of the test scores across raters. The scores of the photocopied scripts of the sampled students which were marked by selected examiners were correlated with the scores obtained by official examiners the Institute of Education employed for the marking exercise. The two sets of scores were correlated using the Pearson's Product Moment Correlation Coefficient (r) to determine the correlation between the two sets of scores (Anastasi & Urbina, 2007). The correlation coefficient obtained indicated the interrater reliability.

The results of the Pearson's Product Moment correlation coefficient are presented in Table 1.

Table 1: Results of the Pearson Product Moment Correlation Coefficient

Raters	Valid N	Mean	Std. Dev	r	r^2	α
English R1	596	45.04	10.71	0.819	0.67	0.05
English R2	596	42.92	10.65			
Maths R1	592	45.66	15.73	0.876	0.77	
Maths R2	592	44.77	15.29			
Science R1	594	49.80	11.36	0.868	0.75	
Science R2	594	49.43	11.86			

The results in Table 1 show that the mean scores of all the ratter 1s of English (English R1) *M*=45, *SD*=10.71, Mathematics (Maths R1) *M*=45.66, *SD*=15.73 and Integrated Science (Science R1) M=49.80, SD=11.36 are greater than the R2s. The mean scores, therefore, show that the scores of the examiners of the official marking were relatively higher than the scores obtained by the examiners selected to mark the photocopied scripts. However, the differences in standard deviations between R1 and R2 in all three subjects are low (not more than 0.5) suggesting that the variations in scores of the two sets of raters are small. But all the correlation coefficients (*r*) are less than 0.90 indicating that the interrater reliabilities are not desirable as suggested by Fraenkel, Wallen and Hyun, (2012). Corroborating 0.90 as the least index for accepting interrater reliability as high, Pearson (2009) in a study conducted for the New York State, described the interrater reliability as very high agreement because the correlation coefficient for all grades was 0.99. The implication for the results of the study is that the ratings of the official marking do not agree substantially with the ratings of the photocopied scripts.

However, approximating the reliability coefficient indices to one decimal place, that of Mathematics (*r*=0.876) and Integrated Science raters (*r*=0.868) could be considered desirable at 0.05 level of significance. This is supported by the results of Axelrod, Goldman, and Woodard (1992) study on the interrater reliability in scoring the Wisconson Card Test whose reliability coefficients included 0.88 (0.92, 0.93 and 0.88) and were considered very high interrater reliability. It, therefore, means that the ratings in Mathematics (FDC122) and

Integrated Science (FDC124) had higher interrater reliability as compared to English (FDC121). This is buttressed by the differences between the mean scores of the two sets of raters in all the three papers. On the average scores of Mathematics and Integrated Science, the differences were less than one (1) but in English the difference was 2.12. The results, therefore, show that the ratings in Mathematics (FDC122) and integrated Science (FDC124) had higher scorer agreement or consistency than in English (FDC121). Banyard & Grayson (2000) opined that if independent observers cannot agree on a high percentage of the observations that they make, then the usefulness of the observations is called into question.

Hypothesis testing

H_o: The ratings of scripts of the DBE end-of-semester examinations by a particular ratter are not significantly different from ratings of other raters.

To test the hypothesis to determine if the difference in ratings of different raters of the same script of the DBE examinations is statistically significant or not, the paired samples t-test was employed. The result is presented in Table 2.

Table 2: Results of the paired samples t-test

Raters	**Valid *N***	**Mean**	***t***	***df***	***P***	***A***
English R1	596	45.04	7.634	595	.00	0.05
English R2		42.92				
Maths R1	592	45.66	1.829	591	0.068	
Maths R2		44.77				
Science R1	594	49.80	0.521	593	0.603	
Science R2		49.43				

Table 2 shows that there are differences between R1s and R2s of the three pairs of groups of raters. In other words, differences exist between the official marking raters and their photocopied counterparts in English (FDC121), Mathematics (FDC122) and Integrated Science (FDC124). But in FDC122 and FDC124 the differences are not significant t (591) = 1.829, p=0.068 for FDC122 and for FDC124, t (593) =0.521, p=0.603. In both cases, p>0.05. In view of this the null hypothesis is accepted for Mathematics (FDC122) and Integrated Science (FDC124) at 0.05α. The implication is that the differences in

the means of FDC122 and FDC124 are not significant and the scorer agreement between the two set of raters is high. It is therefore, safe to conclude that the scorer agreement is high in the ratings of FDC122 and FDC 124 and that the consistency in FDC122 and FDC124 is not by chance or sampling error. The implication is that the differences in ratings that exist in FDC122 and FDC124 did not have any influence on the total scores and that any decision made based on them is valid.

But for the English, the mean difference between the two groups of raters is statistically significant $t(595)=7.634$, $p=00$.This shows that the difference between the two sets of raters in English (FDC121) is real. Consequently, the null hypothesis is rejected for FDC121 at 0. 05α.This means that the ratter agreement observed for English (FDC121) in Table 2 is due to chance or sampling error. This means that differences in scoring between the official raters and photocopied raters in English (FDC121) scores had influence on the total score.

The result relating to the significance of the raters is buttressed by the indices of coefficient of determination (r^2) for FDC121, FDC122 and FDC 124 in Table 1. The coefficient of determination for Mathematics (FDC122) and Integrated Science (FDC124) show that there is a larger proportion of variance shared by both official raters and the photocopied raters. For FDC122, $r^2=0.77$ implying that 77% of the variance is shared by the two sets of raters. This means that there is agreement for 77% of the scores obtained for FDC122. In the case of Integrated Science ($r^2=0.75$) 75% of the variance is shared by the two sets of raters. In other words, 75% of the scores for FDC124 are in agreement by the official raters and the photocopied raters and for that matter, there was high consistency in the ratings. However, for English there was only 67% agreement between the two sets of raters. This means that differences in ratings between the official raters and photocopied raters affected the total score. This suggests that the FDC121 scores are likely to be contaminated with errors as observed by Anastasi and Urbina (2007) that one source of error score variance is scorer variance. In other words, difference in ratter scores introduces errors in test scores.

Supporting the fact that high level coefficient of determination provides a high level of consistency and for that matter differences between raters are not significant is the study of Pearson (2009). The study was to determine the interrater reliability for Grades 3 to 8 Mathematics scores in New York State. The scoring was done by local

and audit raters. Pearson observed that the correlations ranged from 0.96 to 0.99 and the common variance ranged from 0.92 to 0.98. This was described as a high degree of agreement. The common variance, which is the same as coefficient of determination, was high hence, the high degree of scorer agreement. This means that any observed difference between the two sets of raters was not significant.

Conclusion and Recommendations

The study was meant to determine the extent to which different raters agree on the scores obtained by examiners who score the scripts of the DBE examination conducted by the Institute of Education of the University of Cape Coast in Ghana. In other words, it was meant to determine the interrater reliability of the ratings of the DBE examination scripts. The courses involved were English (FDC121), Mathematics (FDC122) and Integrated Science (FDC124). The results showed that the interrater reliability is high in Mathematics (FDC122) and Integrated Science (FDC124). It means that ratings in FDC122 and FDC124 were consistent. In other words, there were agreements of scores obtained by two or more examiners. The coefficient of determination for the two courses suggest that the raters agreed on 75% or more of the scores. However, the interrater reliability of English was different. In addition, the coefficient of determination shows that the raters agreed on less than 70% of the ratings and this was a likely source of errors in the English scores. Anastasi & Urbina (2007) observed that one source of error score variance is scorer variance.

Table1 shows that there were differences between the two groups of raters for the three courses. However, the hypothesis testing (Table 2) established that the differences in the ratings of Mathematics FDC122 and Integrated Science FDC124 were statistically not significant at 0. 05α.This means that the difference in rating in FDC122 and FDC124 did not have any effect on total scores based on which decisions were made. On the other hand, the study established that there is statistically significant difference in English (FDC121) ratings at 0.05α. The results of the hypothesis, therefore, reinforces the fact that the scorer agreement is high in the ratings of FDC122 and FDC 124 and that the consistency in FDC122 and FDC124 is not by chance or sampling error but the contrary holds for FDC121. Hence the conclusion is that the interrater reliability in Mathematics (FDC122)

and Integrated Science (FDC124) is high while that of English (FDC121) is low.

On the basis of the results, it is recommended that the Institute of Education of UCC intensifies the coordination for its examiners, especially, in English (FDC121) so that the examiners understand the scoring rubrics very well. Secondly, team leaders should be more vigilant in vetting the scoring of the assistant examiners. These will help reduce the errors introduced in the total scores. It will eventually make decisions made on the total scores more useful. The researcher further recommends that further studies be conducted in other courses and for different colleges to confirm or nullify this result.

References

AERA, APA & NCME. (2014). Standards for educational and psychological testing. American Educational Research Association.

Anastasi, A. & Urbina, S. (2007). *Psychological testing.* (7th ed.). Prentice- Hall of India.

Axilrod, B. N., Goldman, B. S., & Woodard, I. L. (1992). Interrater reliability in scoring the Wisconsin Card Sorting Test. *The Classical Neuropsychological, 6,* 143-155.

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

Borg, W. R., & Gall, M. D. (1983). *Educational research*. (4th ed). Longman Inc.

Burnham, P., Gilland, K., Grant, W., & Layton-Henry, Z. (2004). *Research methods in politics.* Palgrave Macmillan.

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

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

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

Haertel, E. H. (2006). Reliability. In K. L Brennan (Ed.). *Educational measurement* (4th ed.; pp 65-110). Praeger Publishers.

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

Neşe, G., & Selahattin, G. (2010). Studying reliability of open ended mathematics items according to the classical test theory and generalizability theory. *Educational Sciences: Theory & Practice*, *10*(2), 1011-1019.

Schmidt, F. L., Hunter, J. E., & Uzry, V. W. (1976). Statistical power in criterion-related validity studies. *Journal of Applied Psychology, 61*, 475-485.