INTRODUCTION

Background to the Study

Evaluation is an essential and continuous aspect of teaching and learning. It is the systematic process of identifying the decision options; selecting appropriate information and collecting and analysing information for the purpose of reporting summary data needed by decision makers in selecting among alternatives (Okpala, Onocha & Oyedeji, 1993). Educational evaluation has many purposes both in the education sector and in the society at large. Evaluation induces motivation for learning, which in turn produces good study habits. Evaluation enables the teacher to determine the extent to which behavioural objectives have been achieved. (Okpala, Onocha & Oyedeji, 1993).

Similarly, accurate evaluation information on students' learning, helps teachers arrive at the correct solution to students' learning problem. It is useful in counselling services in schools in the areas of students' education programme, vocational interest, and study habits. It helps principals and teachers take accurate decisions on students in respect of selection and promotion. It is useful for educational policy formulation, parental decisions on children's and wards' education, and also for

decision making in the area of employment of labour (Mamta, 2004; Okpala,Onocha & Oyedeji, 1993).

In the light of the crucial role of accurate evaluation, a considerable effort is being made to ensure objectivity in scoring test. This is most pertinent in the scoring of essay tests. In external examinations, for instance, several steps are taken to create the enabling conditions for objectivity in scoring essay test. Despite the painstaking steps, it is a common fact that unacceptable variation between the original scores and vetted scores still exist. A sample of vetted sheets, for instance, taken by the researcher from the examination department of Edo State Ministry of Education, which had 90 students' scores showed that the variation between original scores and vetted scores ranged between 0 and 15 with a mean of 2.4.

There is an assumption on the part of the public that the marks awarded to candidates in high stakes examinations such as GCSE and GCE are highly reliable and a true reflection of the candidates' abilities with only occasional exceptions. Yet, there are research findings to the contrary. Laming, (1990) and Meadows & Billington, (2005). The level of objectivity of scoring in internal examinations will be worse as there are usually no co-ordination and vetting.

Two major sources of inconsistency in marking are the restricted or the non-detailed marking scheme and some question formats that necessitate markers' opinion during marking thereby giving room to influences from subjective factors such as mood, fatigue, behaviour, handwriting, contrast effect, halo effect, among others.(Hout,1990).

As a result of this problem of inconsistency, some examination bodies resort to various screening exercises to ensure that only experienced markers participate in marking of scripts. This measure often results in the weeding of significant number of potential markers leaving the enormous task to few hands which in turn creates the problem of increasing pressure on the enlisted markers. This is a situation, which may create the problem of fatigue, leading to loss of concentration, which potentially could increase marking error. Fatigue is the physical and mental stress from marking for several hours without rest or break to eat, which begins to take its toll on marker's concentration.

Many other measures for reducing these influences have been suggested and tested without significant result. For instance, Wolf, (1995), Lave and Wenger, (1998) recommended the use of exemplar material. Exemplars

are examples of students' work which are used for scoring examination scripts instead of assessment criteria. However the works of Baird, Greatorex and Bell (2002, 2003) did not support the claim that the use of examplar material alone could increase the marking reliability of essay test.

The use of double marking method has also been recommended by some educationists like Smith, B., Sinclair, H., Simpson, J., van Teijlingen, E., Bond, C., & Tyalor, R. (2002);. Double marking was actually adopted by some examination bodies such as the GCE and CSE examination boards in England in the late 1970s. Though this method yields better reliability than the use of a single marker, the cost implication has made it difficult to use the method. (Meadows & Billington, 2005).

The US and UK examination bodies have embraced e- marking as an alternative to the conventional marking method. However a couple of studies revealed small and inconsistent differences between the reliability of e - marking and the conventional method (Fowles, 2002, Raikes, 2002, Sturman & Kispal, 2003).

Another measure was the adoption of the automated marking, that is using the computer, to assess the mechanical features of candidates' responses (Cohen, Ben-Simon & Hovav, 2003) and in the marking of the short answer tasks format in science (Sukkerieh, Pulman & Raikkes, 2003; Fowles, 2005). Although automated marking is reliable, the validity of this type of marking is potentially threatened by the use of computer keys that will attract undue marks (Ridgeway & Mcmusker, 2004). So it is uncertain whether the use of computer alone will be accepted in the foreseeable future (Meadows & Billington, 2005). Thus Lamprianou (2004) suggested the combination of human marker and a computer, and whenever there is a significant difference in the scores awarded a second human marker would do a second blind marking (having only candidates' numbers on scripts)

Some experts also believe that the holistic and the impression marking methods give more room for subjective interpretation, so they recommended instead the analytic marking in which every facet of an essay question is duly allocated marks in the marking scheme (Huot, 1990, Vanghan, 1991;). Research findings, have confirmed the superiority of the analytic marking to the holistic and impression marking methods (Hout, 1990; Vanghan, 1991;), but studies have shown that in analytic

marking the marker reliability decreases with the increasing complexity of the essay (Delap, 1993a & Ucles, 2000)

Furthermore, some educationists like Okpala, Onocha and Oyedeji (1993) advised against the use of the whole script marking method, in order to enhance the marking reliability of the essay test. Whole script marking is the marking of all the responses in one script before going to another script. This method of marking suffers from the problem of halo effect. Halo effect is the bias, in the marking of a particular item due to a previous impression the marker had of the script owner. This factor tends to make measurement error higher among the scores of upper achievers and that of the lower achievers. Therefore, these experts recommended the use of individual question marking method. Some researchers called this method segmentation (Bakker & Van Lent, 2003; Meadows & Billington, 2005). Segmented marking refers to the marking of one item across the scripts before engaging in the marking of another item.

Segmentation or part making affords the marker the opportunity to compare item response of each candidate to other candidates' responses to the same item there by reducing the level of inconsistency arising from the non - detailed marking scheme, halo effect and the effect of poor

concentration likely to plague the marking of scripts by inexperienced markers.

Nevertheless the objectivity of this method is not without a threat, as non-cognitive elements such as handwriting, and contrast effect may have higher compelling influence on students' scores because of the greater reliance on comparative scoring. Contrast effect is the tendency to underrate or overrate an essay response of an average quality when preceded by a series of responses of excellent quality or responses of poor quality respectively. In addition, if this method is more laborious and time consuming as some examiners claimed (Meadows and Billington, 2005) the effect of fatigue may offset a substantial portion if not all the gains that would have accrued from the method.

The foregoing rationalization can only be resolved by empirical evidence and since this method of marking is being used by some examiners; it is pertinent to have empirical evidence and justification for its use. Bakker and Van Lent (2003) expressed the lack of evidence on the relative effectiveness of segmented marking method in literature. They said that as e-marking becomes common there will be increased opportunities for empirical study of the belief that segmentation can 'add to the objectivity

of the marking. Fowles (2005) discovered the persistence of this gap. This situation was reported by Meadows and Billington (2005), as follows: "Although part versus whole script marking is a topic that might be expected to have received research attention, Fowles (2005) found little reference to this aspect of marking". With respect to the e – marking version of this mode of marking – the digital separation of students' scripts into the different items and each item marked by different markers online, Ofqual (2014) says: "There is currently limited empirical evidence available to enable a robust comparison of the relative merits of whole – script marking and item – level marking".

Despite the progress made so far by scholars in research on marking reliability, conclusive evidence on the relative effectiveness of the segmented marking method, was neither found by the researcher nor previous researchers who did extensive review of literature on this subject (Bakker & Vant Lent, 2003, Fowles, 2005, Meadows & Billington, 2005). Some experts suggested the need for further investigation to assess the relative effectiveness of the segmented marking method (Raikes, 2002; Ucles, 2002; Bakker & Van Lent, 2003; Meadows & Billington, 2005).

In this vein, notable scholars such as Cronbach and Shavelson, (2004) recommended the use of the Standard error of measurement (SEM) as a measure of score reliability. The SEM is the measure of the spread of scores obtained by a single examinee when tested repeatedly without new learning. The SEM is derived statistically because it is not possible to keep examinee from new learning. The SEM unlike the correlation coefficient takes into account the mean score and it is not affected by the spread of scores although it covers only the random component of measurement error in exclusion of the systematic errors such as severity and leniency in marking. In classical test theory the smaller the error component (SEM) relative to the actual score the more reliable the score.

According to Meadow and Billington (2005) there are scholars like Murphy who have also recommended the use of the average mark change (AMC) to estimate the reliability of marks awarded by several markers to an examinee's script. The AMC is the mean of the absolute mark differences awarded by several scorers to an examinee's essay script. The AMC covers both the random errors as well as the systematic errors.

Statement of the Problem

There is substantial evidence in literature of unreliable marking of the essay test in secondary schools and higher education. Many strategies have been tested by researchers yet the desired increase in marking reliability has not been achieved. The search for a solution is still ongoing. One such expected solution is the use of segmented marking method.

In the absence of sufficient empirical evidence one did not know in reality, whether there would be any significant improvement in the objectivity and reliability of essay scores by using segmentation in the place of the usual whole script marking, because of the conflicting probabilities. On one hand the method may reduce halo effect, the effect of non -detailed marking scheme and the effect of poor concentration; while on the other hand, factors like undue influence of handwriting, contrast effect, inexperience and fatigue may out- weigh those advantages of this method.

Therefore the question for this study was, which of the two marking methods, segmentation and whole script would be more effective in reducing the standard error of measurement, marking time, vulnerability

to markers' experience and susceptibility to examinees' handwriting, the general factors in essay marking error?

Purpose of the Study

The main purpose of the study was to experimentally determine the effectiveness of the segmented marking method relative to the whole script method of scoring essay test by comparing both of them using the SEM and AMC as measures of marking reliability. The aim therefore was to use these statistics to ascertain the relative level of objectivity, marking reliability and time efficiency of two groups of markers using the whole script and segmented marking methods. In line with this purpose the specific objectives of this study were to ascertain the following:

1. Standard errors of measurement in the scores awarded by the whole script and segmented marking groups.
2. Average mark changes in the scores awarded by the whole script and segmented marking groups.
3. Relationship between examinees' handwriting and the average mark changes in the marks awarded by the segmented marking group.
4. Relationship between examinees' handwriting and the average mark changes in the marks awarded by whole script marking group.

5. Relationship between marking experience and the standard errors of measurement in the scores awarded by the segmented marking group.
6. Relationship between marking experience and the standard error of measurement in the scores awarded by the whole script marking group.
7. Average times used by the whole script and segmented marking groups.

Significance of the Study

The findings of this study proved in quantitative terms the superior potential effectiveness of the segmented marking method in the marking reliability of essay scores. If examination bodies adopt the recommendations of this study, many teachers will embrace the method and students' results will become more reliable. Thus students, teachers, educational administrators, parents and wards, entrepreneurs and the society will benefit as follows:

Students will gain better understanding as teachers' decisions with respect to instructional objectives, selection of contents and

learning experiences will enhance teaching and learning when the feedback from evaluation becomes more objective.

The school teachers' job will be made easier as students' motivation for learning will be boosted when their grades are a true reflection of their efforts.

Educational administrators will have less worry when there is more reliable marking of essay test. They will have more fulfillment, as evaluation based decisions in areas like selection, promotion, certification, and granting of scholarships, bring more success of the education system.

Examination bodies will have improved public confidence as the results they award becomes more reliable.

With enhanced objectivity of the reports from schools, Parents and Guardians will make better and rewarding decisions on their children's and wards' education.

Finally, the private, public organizations and the society at large will be enhanced as recruitment decisions will bring increase in

the level of effectiveness and success when the evaluation reports upon which such decisions are based become more reliable.

Scope of the Study

This study covers two methods of scoring essay test - whole script marking and the segmented marking methods. The two methods are fundamental to other marking methods.

The essay test used in the study is the Economic theory paper of 1998 NECO examination. This subject was used because it is the specialty of the researcher. The Paper was chosen because there are more essay items than quantitative types in it. Other subjects were not used in the study. The lower levels of education were not covered as external examination was the major focus of the study. NECO examiners were used because of the willingness of NECO Edo state district head to render practical assistance to the researcher.

The study covers the three techniques of estimating marking reliability namely the standard error of measurement (SEM), average mark changes (AMC), and the coefficient of correlation. The standard error of measurement (SEM) (excluding the measurement error of the mean and the relative standard error of measurement) was used for estimating the

marking reliability of the two methods under examination. The average mark change (AMC) was also used as a complementary measure of marking reliability to capture the systematic errors which SEM does not cover. The coefficient of reliability was not used because of its susceptibility to the spread of scores and other limitations of the technique which are discussed in the review of literature.

The two general factors namely the markers experience and examinees' hand writing were used to assess the level of susceptibility of the two methods as confirmatory tests.

Finally the study covers the issue of marking time efficiency. Thus the marking time effectiveness of the two methods was ascertained.

Research Questions

To achieve the specific objectives of this study the following research questions were raised:

1. What are the mean standard errors of measurement in scores awarded by the segmentation and whole script marking groups?
2. What are the average mark changes in the scores awarded by the segmentation and whole script marking groups?

3. What is the relationship between the average mark changes in the scores awarded by the segmentation marking group and the mean ratings of the students' handwriting?
4. What is the relationship between the average mark changes in the scores awarded by the whole script group and the mean ratings of the students' handwriting?
5. What is the relationship between marking experience of the segmentation marking group and the standard errors of measurement in the scores they awarded?
6. What is the relationship between marking experience of the whole script marking group and the standard errors of measurement in the scores they awarded?
7. What are the average marking times used by the segmentation and whole script marking groups?

Hypotheses

Seven hypotheses were tested by the researcher at 0.05 alpha level as operational guide to this study. These are as follows:

1. There is no significant difference between the mean standard errors of measurement in the scores awarded by the segmentation and whole script marking groups.

2. There is no significant difference between the mean average mark changes in the scores awarded by the segmentation and whole script marking groups.
3. There is no significant relationship between the average mark changes in the scores awarded by the segmentation marking group and the students' handwriting.
4. There is no significant relationship between the average mark changes in the scores awarded by the whole script marking group and the students' handwriting.
5. There is no significant relationship between the marking experience of the segmentation marking group and the standard errors of measurement in the scores they awarded.
6. There is no significant relationship between marking experience of the whole script marking group and the standard errors of measurement in the scores they awarded.
7. There is no significant difference between the average marking times used by the segmentation and whole script groups.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

In this chapter, existing literature that are related to this study are presented. The chapter covers conceptual framework on whole script marking, segmented marking, concept of reliability, measures of reliability; theoretical framework on the classical test theory; theoretical studies on Thurston paired comparison of scripts, average mark change, limitation of correlation as a measure of reliability; empirical studies on the evidence of unreliable marking, factors affecting marking reliability, marking methods for improving marking reliability in essay test, and the summary of the literature reviewed.

Conceptual Framework

The conceptual framework is presented as follows:

Whole Script Marking Method. Whole script marking is the marking of all the responses in one script before going to another script (Bakker & Van Lent, 2003; Meadows & Billington, 2005). The reliability of this method of marking is constrained by the problem of halo effect. Halo effect is the bias, in the marking of a particular item due to an earlier impression from the marking of previous items, the marker had of the

script owner. Hence measurement error tends to be higher among the scores of upper achievers and that of the lower achievers. The marking of all the items before going to another script increases marker's familiarity with examinee's handwriting and writing style. This means that the reliability of the marks awarded to items towards the end of the script would be higher than the reliability of marks awarded to items preceding the former. This is most likely to be true with scripts with low hand writing quality.

Whole script marking allows for the general impression marking which tends to increase measurement errors. Holistic ratings may have high interrater reliability mainly because they depend on characteristics in the essays which are easy to identify though irrelevant to 'true' writing ability. Such characteristics included the following: quality of handwriting, word choice, length of essay, and spelling errors. In addition, the academic environment in which holistic scoring usually takes place is not free.

Huot (1990) gave four crucial points against holistic scoring: these are as follows:

a. that holistic ratings correlate with appearance and length;

b. that the product orientation of holistic rating is unsuitable for informed decisions about composition instruction or student writing;

c. that holistic ratings cannot be used beyond the population which generated them, so holistic scoring is useless as an overall indicator of writing quality; and

d. that holistic training procedures alter the process of scoring and reading and distort the raters' ability to make sound choices concerning writing ability.

Furthermore the effect of marker's mood may be heavier on some scripts when the scripts are many.

Segmented Marking Method. Segmented marking refers to the marking of one item across the scripts before engaging in the marking of another item. (Bakker & Van Lent, 2003; Meadows & Billington, 2005). Segmentation or part making allows the marker the opportunity to compare item response of each candidate to other candidates' responses to the same item with the aim of minimising the level of inconsistency arising from non - detailed marking scheme, halo effect and the effect of poor concentration which inexperienced markers are likely to face when marking essay scripts.

There is greater reliance on comparative scoring compared to whole script marking hence non- cognitive elements such as handwriting, and contrast effect may have greater influence on students' scores. Contrast effect is the tendency to underrate or overrate an average quality essay response because of the influence of the preceding series of excellent or poor quality essay responses, respectively. Furthermore, some people feel this method is more tasking (Bakker & Van Lent, 2003) and time consuming hence the effect of fatigue may undo a significant portion if not all the gains that this method is likely to bring.

Concept of Reliability. Rudner and Schafer (2001) maintained that the best way of looking at reliability is to ascertain the degree to which a test reflects the properties of those individuals being measured. Thus, reliability can be defined as "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker" (Berkowitz, Wolkowitz, Fitch & Kopriva, 2000). This definition will be correct if the test scores truly reflect the characteristics of the test takers, otherwise they will vary sharply and unpredictably. Reliability can also be seen as an indicator of zero error when the test is administered.

Ebel and Frisbie (1991) see reliability as how consistent or error - free measurements are. When random error is insignificant it is hoped that scores will be accurate and could be replicated and generalized to other testing occasions and similar test instruments. Theoretically, reliability is the proportion of score variance which is due to systematic variation among the test takers (Meadows & Billington, 2005). This definition is population - specific and portrays reliability as combined features of a test and the examinees.

Group heterogeneity in respect of the trait being measured is a vital factor affecting score reliability coefficient. Ideally coefficient of reliability of measurement is higher for a heterogeneous group in respect of the trait being measured compared to a more homogeneous group. An IQ test, for instance, would be more reliable for a random sample of students compared to a sample of science students.

Classical test theory presumes that only true score variance, varies with group heterogeneity while measurement error variance, does not.. Fan and Yin (2003) believed that this assumption is true on condition. Meadows and Billington (2005) relate this condition as follows:

that when performance levels of the groups are comparable; this assumption appears to be tenable, because the theoretically predicted measurement reliability estimates are largely consistent with the empirically observed measurement reliability estimates.

However, they proved that performance level of a group affects measurement reliability. After offsetting the difference in group variability in the data, measurement error was higher in the scores of the lower performing group , and so their scores had lower measurement reliability. The higher the difference in performance, the more conspicuous the difference in measurement reliability between the high and low ability groups.

Measures of Reliability. It is not possible to calculate a reliability coefficient that conforms to the theoretical definition of reliability for it would require information about the degree to which a population of testees varies in their true scores. Nonetheless there are certain measures often used to estimate the reliability of a group of candidates' scores namely: measure of stability, measure of internal consistency, and measure of equivalence.

Measure of Stability. Measure of stability is one of the ways of estimating the reliability of scores. A test-retest method is used to establish the stability of test scores. The reliability coefficient is calculated by administering the same test twice and correlating the pairs of scores. Wiliam (2000) said that, if a candidate takes a test several times, without new learning , he or she will not have the same score in each occasion. The candidate's level of concentration may vary, the marking standard may vary, the handwriting or the expression might differ and so affect the marking.

A test-retest reliability coefficient theoretically is a functional measure of score consistency since it allows the measurement of consistency from administration to administration directly. This coefficient however is not recommended in practice, because of certain problems and limitations. It needs two administrations of the same test to the same group of candidates and hence more resources and time. If the time interval is short, candidates may remember some of the questions and their responses and If the interval is long, then learning and maturation will affect the results, that is, changes in the candidates themselves.

Where the reliability is only about marking rather than the whole measurement the mark remark reliability test will replace the test retest reliability test. In this case, the scorer/s will have to mark the test script and remark after an interval of time. This method is also constrained by time factor as in the test – retest method.

Measure of Internal Consistency. This is a measure of the degree to which individual items correlate with each other and hence a measure of item homogeneity. It is presumed that items are measuring a common trait if the scores on the items have high correlation. Certain statistics are used for this purpose. The simplest measure of internal consistency is the split-half reliability This coefficient is calculated by splitting a test into two equal halves, correlating the scores on both halves, and then correcting for length using Spearman Brown correction formula because longer tests are more reliable. The split is usually based on odd and even numbered items, or randomly selecting items into two equal sets. The advantage of this approach is that it only requires a single test administration. The limitation of the coefficient however is that it varies with the type of splitting.

> *This is a particular problem when the items are designed to be differentially difficult). Further, it is inappropriate on tests where speed is a factor (that is, where candidates' scores are influenced by how many items they reached in the allotted time (*Meadows & Billington, 2005*).*

The most popular are Cronbach's alpha, the Kuder Richardson Formula 20 (KR-20) and Richardson Formula 21 (KR-21). "Most testing programs that report data from one administration of a test do so using Cronbach's alpha which is functionally equivalent to KR-20. These statistics only require one test administration and they do not require any particular split of items. However they are limited in application to test that measures a single skill area.

> *Where the test aims to measure knowledge, skills and so on across a wide specification, as is the case in GCSE and GCE examinations for example, one would not expect the test to have high internal consistency.*"(Meadows & Billington, 2005)

Measure of Equivalence. Most standardized tests have equivalent forms which can be used interchangeably. These alternate forms are generally selected on the basis of content and difficulty. The

correlation of pairs of scores of alternate forms for the same candidates gives another measure of consistency which is logically an extension of split-half reliability. Even with good testing experience, each test would vary slightly in content and difficulty level and confound the results. However, the use of different items in the two forms makes possible the examination of the extent to which group of items contributes to random errors in the estimates of test reliability. Unfortunately, Satterly (1994) said that the method of estimating reliability extolled by statisticians is to correlate at least two equivalent assessments though, the one- shot feature of UK examinations does not permit this method.

Theoretical Framework

The theoretical premise and the measurement tools for this study are based on the classical test theory. These include the standard error of measurement, the types of standard errors and the formulas for estimating them.

Classical Test Theory. The basic assumptions, principles and the statistical tools of this theory are reviewed in this section. These are as follows: True and measurement error scores, types of measurement errors,

procedures for estimating measurement error and the relationship between reliability and validity in classical theory.

True and Measurement Error Scores. Classical test theory presumes a mark or score as a composite of a true score and a measurement error score. The Standard Error of Measurement (SEM) conceptually is the standard deviation of the error component. Although it is not as popular as the reliability coefficient, it has two advantages: namely, it is not affected by the spread of scores and is more directly related to the likely error on an individual candidate's mark. The true mark will be within one standard error of the observed mark 68 *per cent* of the time and within two standard errors 95 *per cent* of the time' (Meadows & Billington, 2005).

Accordimg to (Meadows & Billington, 2005), scholars like Skurnik and Nuttall advocate for the use of the SEM as a measure of reliability. Recently, Cronbach also maintained that the SEM is the most valuable single piece of information reported about measurement instrument (Cronbach & Shavelson, 2004). They argued that the report on the uncertainty associated with each score, is easily understood by professional test interpreters, educators and lay persons. It is also argued that reliability must be expressed in terms of the level of errors in the

distribution of scores. William (1993), for instance argues that the consistency of classification is the only correct definition of the reliability of national curriculum assessment.

According to William (2000) the starting point for estimating the reliability of a test, in line with the classical test theory, is to assume that each candidate has a 'true score' on any particular test. A candidate's true score is the average score that the candidate would get from many repetitions of the same or similar test. A candidate's actual score in any occasion, according to the classical test theory, is the sum of his or her true score and some amount of error. On any particular day, a candidate might get something higher or a lower than his or her true score.

To obtain a reliability estimate one compares the sizes of the errors with the sizes of the actual scores. When the errors are small relative to the actual scores, the test is relatively reliable, and when the errors are large relative to the actual scores, the converse is true. It is impossible to use the average values for this comparison, because, by definition, the average value of the errors is zero. Instead, a measure of the spread of the values, the standard deviation (SD), is used (Meadows & Billington, 2005).

How high reliability should be is a function of the importance of the test. If the importance is high, as in public examinations, the internal consistency reliability, according to Wiliam (2000) should be above 0.90. When the importance is high, wrong grading due to measurement error must be reduced to the minimum.

The standard deviation of errors as mentioned earlier is known as the standard error of measurement (SEM). Satterly (1994) said that the purpose of a reliability study is to obtain an estimate for the SEM which allows the score user to express quantitatively the uncertainty associated with it and to state the range which the true scores lie.

Every time we engage in any form of measurement we are guided by two motives – precision and accuracy. These motives are predicated on our belief that exact or true value exists. It is this true value we earnestly strive to obtain all the time we have to measure one thing or the other. Every time we try, the quest for the true value remains insatiable.

Despite the difficulty of obtaining the true value, it is possible to get around the problem by determining the possible range of that true value when we are able to estimate the error value. This understanding could be summed up in the following equation

$$X = T + e$$

This simply means that every measurement is a composite of the true value and error component. Consequently the concept of measurement error is the real variation from the true score, (Syque, 2010) When we measure an object for instance and the observed value is say 50 and if statistically we estimate the error value to be 0.5 then the true value lies in the range of 50 ± 0.5

This is the reason, student achievement scores are graded in ranges e.g 100 – 70 is A, 60 – 69 is B e.t.c However when a student's score fall close to the boundary, the grade may not reflect the true score of the student

Measurement errors are generally classified into two, namely random error and systematic error. Consequently the earlier equation could be modified as

$$X = T + e_r + e_s$$

Where e_r and e_s are the random error and systematic error respectively

Random error is the error that inflates or deflates the true value of some of the objects or subjects of the entire sample being measured. According to Trochin (2006) the random error is caused by any factor that randomly affect measurement of the variable across the sample. In an achievement test for instance loss of marker's concentration can inflate or deflate the

performance of students. Scripts marked under loss of concentration will have higher marking error while those marked under good concentration will have lower marking error. The important thing about random error is that it does not have any consistent effects across the entire sample instead it pushes observed scores up or down randomly (Trochin,2006). The net effect of this random distribution is that there is a regression to the mean as some students lose marks while some gain marks. The net effect therefore does not affect the average score. Because of this, random error is sometimes considered noise (Trochin, 2006). Random error does however affect the spread of the scores by increasing the variances. This is shown in the figure 1

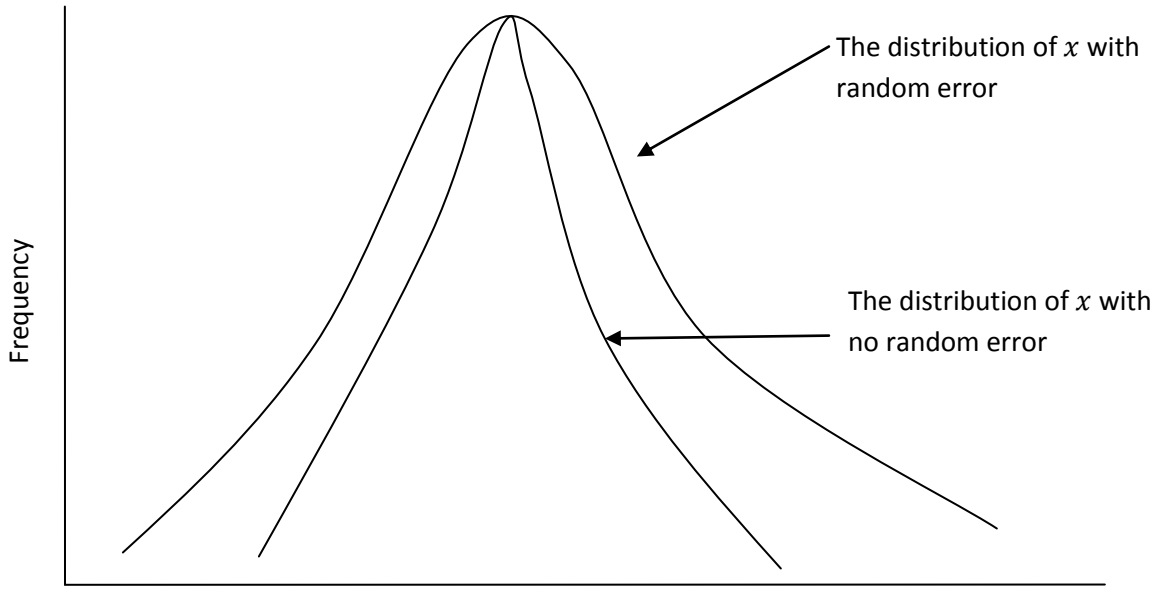

Figure 1: *Showing the Effect of Random Error on the Distribution of Scores*

Systematic error is an error that affects the entire scores positively or negatively. Its effect is consistent throughout the entire sample. According to Trochin (2006), systematic error is caused by any factor that systematically affects measurement of variable across the sample. An example of a systematic error is severity or leniency in marking. This attitude will affect all the examinee's performance and increased or lowered their scores respectively. As a result of the uniform impact, systematic error is sometimes referred to as a bias in measurement. Systematic error causes the average of the group to shift to the right or left

depending on whether it is negative or positive. This is shown in the diagram 2

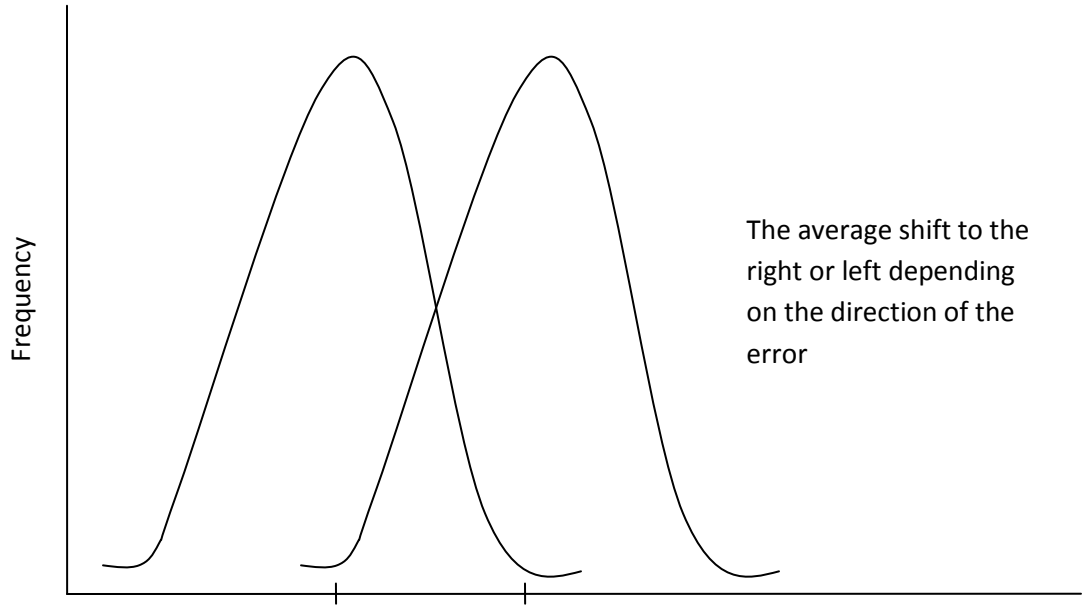

Figure 2: *Showing the Effect of Systematic Error on the Distribution of Scores*

Estimating Measurement Error. The best way of understanding the processes involved in the estimation of measurement error is to imagine that a single student were to take the same test several times without new learning and memory of questions effects. The standard deviation of his/her repeated test scores is referred to as the standard error of measurement.

*Standard Error of Measurement***.** While the standard deviation of scores on a test is a measure of the spread of scores between students, the

standard error of measurement on a test is a measure of the spread of scores within scores obtained by a single student when tested repeatedly. Since it is highly improbable to control or account for new learning and memory of question effects, we can estimate the standard error of measurement from the test scores of a population of students in a single test, by estimating the mean score, the standard deviation of scores from the mean score and the reliability of the scores. The standard error of measurement is computed thus:

$$S_E = S_X\sqrt{1 - r_{xx}}$$

S_E = Standard error of measurement

S_X = Standard deviation of the test scores

r_{xx} = reliability of the test scores.

This formula was derived as follows: Whether we give repeated test to one individual or we give one test to a group of individuals, variation in the error scores is equal to the variation in the observed scores and true scores as follows:

The basic assumption as mentioned earlier is that the observed score is equal to the true score plus the error score:

X=T+ E……,,,,,,……..(1)

According to classical theory one true score does not influence another true score and one error score does not influence another error score so if we add the variance of true scores and error scores of a group of testees it equals the variance of the individual observable scores :

$$\text{Var}(x) = \text{var}(T) + \text{var}(E) \quad \ldots\ldots (2)$$

To estimate the error component rearrange equation (2) to make var (E) the subject:

$$\text{Var}(E) = \text{Var}(x) - \text{var}(T) \quad \ldots\ldots (3)$$

To obtain var (E), we calculate var (x) from x_s and calculate var (T) by multiplying the reliability of x_s by var (x) i.e.

$$S^2_E = S^2_x - R\,[S^2_x] \quad \ldots\ldots (4)$$

We factorize the right hand side of equation 4, to have:

$$S^2_E = S^2_x[\,1 - R] \quad \ldots\ldots (5)$$

We take the square root of both sides of the equation 5 to obtain the SEM as follows:

$$S_E = S_X\sqrt{1 - r_{xx}} \quad \ldots\ldots (6)$$

Sygie, (2010).

The above calculation relates to the entire measurement process, which includes the test, candidate, administration of test and the marking. However, when the formula is used in comparing markers who mark the

same scripts, the differences in values will be a reflection of the differences in their marking reliability.

The validity of R [S^2_x] can be shown as follows: From equation (3) If var (x) and var (T) are equal, then var (E) equals zero meaning a perfectly reliable measurement, thus reliability is a function of how well the observed score variance approximates the size of the true score variance hence reliability is defined as the ratio of the true score variance to the observed score variance i.e. R = var (T)/var (x), Multiply the two sides of this equation by var(x) and reverse the equation to make var (T) the subject of the formula i.e. var (T)=R[var (x)].

*Standard Error of the Mean***.** In a survey where the objective is to estimate the sample mean in order to describe or infer the population mean the error associated with this measurement is the standard error of the mean. When repeated measurements of the sample means are done, the standard deviation from the average is referred to as the standard error of the mean, all other factors being equal.

As in the standard error of measurement, the standard error of the mean can be estimated in a single sample from the population using the formula below:

$$SE_{\bar{X}} = \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \ 1$$

Where:

S = is the sample standard devaiton (i.e) the sample based estimate of the standard deviation of the population

n = is the size of the sample

The formula above for the standard error of the mean is for small sample of means. When the sampling fraction is large (Approximately at 5% or more) the estimate of the error must be corrected using a finite population correction" (Isselis,1918)

$$FPC = \sqrt{\frac{N-n}{N-1}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ 2$$

The effect of this correction formula is that the error tends towards zero as the sample size n tends towards the population size N,

The above formulas (1 and 2) are for independent sample. When the sample is not independent the error estimate obtained by formula 1 should be corrected by factor f defined thus:

$$f = \sqrt{\frac{1 + (n-1)}{1-p}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . \ldots 3$$

Where:

p is the sample bias coefficient of the average of the auto correlation coefficient This value is between -1 and 1 for all sample points (pairs).

Relative Standard Error. When two or more standard errors are estimated for the same or different populations and we want to identify the survey that is more reliable, we can calculate the relative standard error of the different surveys by dividing the standard error of each survey by the respective sample mean and multiply by one hundred rendering the value in percentage.

The Relationship between Reliability and Validity in Classical Test Theory. In classical test theory predictive validity is the underlying motive of testing candidates. Though predictive validity is only one of a number of inter-related forms of validity, they are fundamentally the same, the issue according to Clark-Carter, (1997) is "*whether what is being measured is what the researchers intended*" (, p.28). According to classical test theory, the maximum validity for a test is the square root of the reliability. People often say that validity is more important than reliability, because it is useless to measure something reliably if one does not know what one is measuring. At the same time, reliability is a pre-condition for validity since no assessment can be said to be valid if the

mark a candidate gets varies significantly from one occasion to another, or is marker dependent **(**Clark-Carter, (1997).

Theoretical Studies

Thurstone Paired Comparison of Scripts is discussed in this section because of its major feature – comparative judgment of scripts. This feature is the major characteristic of the segmented scoring method. Also discussed in this section are the average mark change and the limitation of the correlation coefficient as a measure of reliability. These are as follows:

Thurstone Paired Comparison of Scripts. Pollitt (2004) and Pollitt and Crisp (2004) recommended that traditional marking be replaced with Thurstone paired comparison of scripts anchored on the examiners' impression of students' work. Instead of counting the number of correct points students make, the method employs the comparative judgment of the responses in whole scripts or work of each candidate.' This provides a method of constructing an interval scale from judgments". What this means is that a scorer will first make a comparative judgment of all the scripts and arrange them in a hierarchy according to the quality of the scripts and then determine the upper and the lower bound scores. The interval scores are then allotted accordingly.

Pollitt (2004) said that this is feasible because even if human judges are likely to have their own personal scale for rating quality, if they compare two things, their own standard becomes silent. A true measurement scale that shows the level of performances in relation to each other can be constructed. The method produces a measurement value for each script including the standard error of that estimate. "This method could also make awarding meetings (where grade boundaries are decided) redundant if some of last year's scripts were included" (Meadows & Billington, 2005).

Furthermore, scripts close to borderline (boundary between grades) which have standard error that transcends the boundary could be sent for further comparisons to minimize the risk of downgrading. This statistical procedure would also detect scripts with inconsistent judgments. Such scripts, could be solved by a senior examiner. The statistics also allow the assessment of the consistency of individual judges and this could be used in early disqualification of an inexperienced examiner.

Pollitt and Crisp (2004) demonstrated that this method would produce a more valid assessment by ameliorating the restrictions usually foist on the

format of questions in accord with traditional marking. This method however, is more expensive as it requires more than one examiner to make comparisons about the same script and each script would be compared to 20 other scripts normally.

> "Unless comparisons can be made quickly this could increase examiners' workload. Given that multiple assessments of scripts are required, the pros and cons of this approach compared to that of double-marking need to be investigated" **(**Meadow**s** & Billington, 2005**).**

The Average Mark Change. When the reliability of marking, as opposed to the reliability of the assessment as a whole, is the issue, Murphy (1982) argues that: the simplest way of describing the amount of variation in candidates' marks due to different examiners doing the marking, is the average mark change. This measure reports the mean of the variations in the marks awarded to the candidates in an examination. In the words of Meadows & Billington (2005) **"**The average mark change is expressed as a ratio of a fixed amount (say 100) for examinations that have similar distributions of marks. Such value can be used as a comparative index of marking reliability" (2005).

Presumably, Murphy (1982) intends that the mean mark variation should be calculated using absolute mark differences; otherwise, the positive and negative mark differences would cancel out and produce a misleadingly low mean mark change. What this means in effect is that one sums up all the deviations of each scores from their mean, ignoring the positive and negative signs and divide the sum by the number of cases.

The Limitations of the Correlation as a Measure of Reliability. The correlation coefficient is used by many researchers as internal consistency estimates of reliability. Coffman (1971) says that the correlation coefficient exaggerates reliability because it ignores the means and standard deviations of the scores.

Lunz, Stahl and Wright (1994) showed that even a perfect correlation may undermine systematic differences between raters. There is also the disadvantage that the correlations observed will be affected by the spread of performance in the sample of scripts examined.

Consequently researchers have searched for an alternative to the correlation coefficient (Meadows & Billington, 2005). Classical test theory has an alternative; it is the remediation of the reliability coefficient.

This approach is called the standard error of measurement (SEM) as discussed under the theoretical framework.

Empirical Studies

Marking reliability has been studied extensively in education covering various subjects and scoring methods. This section covers strategic studies to show the scope of research in this area and the levels of marking reliability reported by researchers in education. The subtitles in this section are: empirical evidences of unreliable marking in secondary school and higher education, the factors affecting marking reliability and the marking methods for improving marking reliability. These are reviewed as follows:

Empirical Evidences of Unreliable Marking. Murphy (1978, 1982) carried out detailed analyses of the reliability of marking in 20 O' and A' level examinations which were held between 1976 and 1979. Out of the eight subjects initially studied, the English A level held in 1976 was the least reliably marked and had a correlation coefficient comparing prime with re-mark of 0.73; the next two papers had slightly better coefficients of 0.85 and 0.76 respectively.

Murphy (1982) examined the reliability of marking in English O level between 1976 and 1979; in Britain, the coefficients of correlation for Paper 1 and Paper 2 were 0.75 and 0.91 respectively in 1976, while in 1979 they were 0.76 and 0.93 respectively. Murphy claims that these values are the reliability of marking of individual components. The overall reliability of an examination will depend on the marks obtained from all the examination papers. Thus, even if the highest coefficient of correlation for the three components of 1976 English A level was 0.85, the coefficient of mark re-mark of all subject was 0.91. This means that the greater the number of components the greater the reliability of marking of an examination.

Murphy (1982) also reported an analysis of the 1977 examinations in O level Mathematics and A level Pure Mathematics. The correlation coefficients comparing prime with re-mark for this two subjects were very high. Two of the three O level papers had a coefficient of 1.00 (although one of these was a computer marked objective test) and the other had a coefficient of 0.99. One of the three A level papers had a prime to re-mark correlation coefficient of 1.00, another had a coefficient of 0.99 and the third had a coefficient of 0.98" the figures of reliability for mathematics; were high which makes it the most reliably marked of all the subjects. It

was observed that the least reliably marked examinations appears to be those that are dominated by essay-type questions while the most reliably marked appears to be those that contain highly structured, analytically marked questions.

There has also been couple of studies reporting the reliability of marking among various subjects in Higher Education.

In 1970s it was already settled that marking reliability is relative to the subject area being examined. James (1974)and Mcvey (1975) studied the marking of scripts in physics electronic engineering. They found that in such examinations the correlation coefficients between markers were as high as 0.9 or above. Byrne (1979) reported a study carried out to unfold the reliability of tutor-marked assignments at the Open University. Inter-marker reliability was higher in assignments in mathematics, almost as good as in the physical sciences and related technology and lowest in the arts, social sciences and education faculties. Irrespective of the subject area, however, essay questions presented the greatest reliability problem.

Marking reliability study was also conducted by Engvik, Kvale and Havik (1970) at the Psychological Institute, Oslo. Essay and oral performances of candidates were assessed by committee of three experts. Significant

differences in the mean score awarded were found both within and between committees. The same essays had a wide variation of reliability coefficients from -0.16 to 0.90" when rated within a committee,

Laming (1990) studies the marks awarded (through blind making) by pairs of markers for answers in an unspecified university examination for two years. The correlations between the two marks varied from 0.47 to 0.72 for the first year and from 0.13 to 0.37 for the second year. Laming used the classical test theory to calculate the accuracy of the examination and found out that for the second year the level of precision was insufficient to back up the classes of degree awarded to candidates.

The Factors Affecting Marking Reliability. There are two main types of factors affecting marking reliability namely the voluntary and involuntary factors. The concerns of this study are the involuntary factors. The involuntary factors affecting marking reliability are reviewed in this section as follows:

Effect of Fatigue on Marking Reliability.

Morrissy (2000) carried out a study to find out whether the standards of examiners' marking will reduce before the end of the marking period with

the help of re-marking data from GCSE English and Geography, and GCE English and Theatre Studies. The study found no proof of significant difference on the marking reliability relative to the point of marking a script in the batch or the size of the batch. There was no evidence of variation in the leniency or severity of marking during the marking session. In a similar study by Pinot de Moira, Massey, Baird and Morrissy (2001) they found only little variation in the relative leniency or severity of examiners during the period of marking summer 2000 GCE English scripts. Lunz and Stahl (1990a), studied inter-judge differences in reliability, between examining sessions. They used a different method for eliminating variations in rater severity. They provided data from three different examinations namely: an English Literature essay examination, a clinical examination and a Health profession oral examination which revealed that raters manifested significant variation in severity in two of the three, over grading periods lasting from one to four days. They maintained that short-term effects like fatigue and attitude may have been responsible for the observed variations. They believed this is a normal human behavior and cannot be erased easily.

Humphris and Kaney (2001) examined the issue of fatigue in examiners in objective structured clinical examinations (OSCEs). Live patient-clinician

interactions were evaluated in the examinations. The purpose of the study was to find out if marking reliability changes during the single session of testing (two hours). They discovered little evidence of a systematic bias that may be attributed to fatigue or tiredness. The marks that were examined for bias were pooled marks from four examiners. "*Bias due to poor concentration, lack of vigilance or stereotypical judgments, which might be indicative of fatigue, may not be shown when the marks of four individual examiners are pooled"* (p. 448).

***The Effect of Contrast on Marking Reliability*.** The marking of a script has been shown to be influenced by the quality of the scripts marked before it (Meadows & Billington, 2005). Contrast effects have been reported in many marking exercises. Hales and Tokar (1975) carried out a study which shows that student teachers marked two essays of moderate standard significantly lower when they followed series of five good essays than five poor essays. Hughes, Keeling and Tuck (1980a and b) showed that good and poor essays were less vulnerable to contrast effects compare to moderate standard essays. They also established that contrast effects appeared to wane with time during marking session. Hughes *et al* said that as marking progresses marking standards would

emerged and as a result markers become less susceptible to contrast effects.

Daly and Dickson-Markman (1982) maintained that both the Hales and Tokar (1975) and Hughes *et al* (1980a and b) studies were constrained by the lack of effective control groups for comparative analysis that is a rating of the criterion essay by itself, unaffected by other papers and a rating of the criterion essay following a block of papers of variable quality. Their study included these conditions and replicated the finding that ratings of the criterion essay differed as a function of the quality of the previously read papers

In a similar study Spear (1996, 1997) also discovered that good work appeared to attract more favorable mark when it followed work of a lower quality compared to when it preceded such work. Poor quality work was marked more severely when it followed work of higher quality. Spear tried to improve on the design of previous research on contrast effects by using practicing teachers to mark live work (scientific reports). She maintained that the marking of this type of material would be more objective than marking of essays, naturally making contrast effects non effective.

She found evidence nonetheless of contrast effects when only two samples of work of contrasting quality preceded a criterion report. Two samples of work preceding a report of a contrasting quality elicited greater bias than a single sample. Spear in conclusion noted that the commonly adopted practice of reading through several pieces of work before marking appears incapable to prevent contrast effects unduly influencing the marks awarded to the first few pieces of work.

Vaughan (1991) showed qualitative evidence of contrast effects. In the study raters were made to read through and grade essays holistically while making necessary comments into a tape recorder. Analysis of the transcribed tapes were made, the revelation showed that the essays tend to become one long discourse in the mind of the raters. Raters made comparative statements such as: This essay is better or worse than the previous one.

Hughes *et al* (1980b) prefer to use 'context' instead of 'contrast' effect. They examined the impact of marking method and context essay position on essay marking. It had been argued that analytic marking method would be superior to holistic marking with respect to resistance to context effects. Analytic marking confines examiners to strict guidelines about

weightings to be awarded for specified essay competences like writing style, originality of ideas, grammar etc. In holistic scoring the marker makes a single global judgment based on general impression. In terms of context essay position, it was observed that if markers read and grade several essays varying in quality before marking the context block of essays context effects would reduce.

They discovered that both marking strategies were equally vulnerable to context effects. Similarly, putting the block of context essays late in the order of making did not reduce context effects in relation to placing the block of context essays early in the order of making. In a different study, Hughes *et al* (1983) attempted to prevent context effects by warning markers before commencement of marking about their influence. They asked markers to categorize essays qualitatively and read them again before awarding final grades. The marks of the three types of markers namely those who went through the above procedures those markers who were merely warned of the existence of context effects and those markers who were given no information about the influence of context were compared. Results revealed that the three groups had the same level of context bias.

Finally Hughes and Keeling (1984) gave model essays to markers to use as guide. Context effects were still detected notwithstanding. However it is still possible that the provision of models may reduce the influence of context on the marking of essays where factual accuracy instead of written communication is being assessed. In conclusion Hughes and Keeling (1984) noted that "*we may be forced to accept context effects as an unavoidable concomitant of essay scoring*" (p. 281).

The Effect of Hand Writing on Marking Reliability. Briggs, (1970, 1980); Bull and Stevens, (1979) provided evidence that good handwriting enables the teacher to make out easily the intent of pupil" responses, poor handwriting on the other hand makes the reading of their response difficult (Bull and Stevens, 1979).

Briggs (1970) observed the same phenomenon. Briggs (1980), conducted an experiment using practicing teachers of English to assessed copies of 16+ external examination scripts rewritten in five different handwriting styles, On the strength of this experiment he expressed "that handwriting may significantly affect the chances of some 16-year-olds passing or failing the 16+.

Markham (1976) carried out an experiment to examine the influence of handwriting quality on evaluation of written work by markers. The study used 45 teachers and 36 student teachers to rate descriptive paragraphs with varying quality in content and handwriting style. Multiple Classification Analysis revealed that the teacher characteristics of experience, level taught, degrees held, age, and the student teacher characteristic of level taught do not have a significant influence on the score given to a paper. In any case the analysis of variance showed that the variation in scores due to handwriting was significant. Papers with better handwriting consistently received higher scores than did those with poor handwriting regardless of content.

It seems that the handwriting factor is not straightforward, as other variables, such as gender and attractive photo of the student tend to interact with it. Bull and Stevens (1979), conducted a study in which an identical essay in content was marked by 72 raters who were mostly school teachers, with few students. Some of these raters have their essay scripts in typed form, some in good handwriting and some have scripts with poor handwriting. ”A photograph of the supposed author of the essay was attached to the essay. This photograph was of a male or a female who was either highly physically attractive or rather unattractive” (Meadows &

Billington, 2005).The result shows that when the authors were female the awards given to the essays were influenced by handwriting and attractiveness. But where the authors were male the influence of handwriting was absent. On this phenomenon Bull & Stevens say that

> it is possible that society expects females to have better handwriting than males and so when a female has poor handwriting the resulting variable impression created is poor. Similarly perhaps women are judged more on attractiveness than are men. (p. 58 cited by Meadows & Billington (2005)

Truly, good handwriting seems to favor pupils in the marks they get for their written work. Handwriting bias is magnified by gender and attractiveness.

***Effect of Teacher's Experience on Marking Reliability*.** Some studies seem to suggest that inexperienced markers tend to mark more severely and use different rating methods than experienced markers. (Cumming, 1990; Gordon & Kraemer, 1999; Huot, 1998; Ruth & Murphy, 1988; Shohmy, 1992; Weigle, 1994). In their study, Ruth and Murphy (1988) revealed that student teachers tended to award more

severe marks for essays than experienced markers, though the differences were insignificant. They opined that the markers' differences influenced their judgment of the essays. Weigle (1999) in a similar study showed that inexperienced examiners were more severe when compared to experienced markers. She discovered that, before their training, inexperienced markers are clearly more severe than experienced markers in some essay titles, but the differences in severity tend to disappear after training. She asserted that her findings *"underscore the complexity of the relationship between rater background, the scoring rubric, the prompt, and rater training in writing assessment."* (p.171,cited by Meadows & Billington, 2005)

Myford and Mislevy (1994) carried out a study on the Advanced Placement examination in Studio Art. They tried to find out background variables, including years of teaching experience, which are predictors of marker severity, instead they discovered that the variables studied had an insignificant effect on predictions of marker severity. Meyer (2000a, 2000b) revealed that the length of examiner experience and a senior examiner's "rating of the examiner's performance (from A - consistently excellent, to E - unsatisfactory not to be re-employed) rarely proved useful

as predictors of whether an examiner's marks would require adjustment" (Meadows and Billington, 2005).

Michael, Cooper, Shaffer and Wallis (1980) carried out a comparative study of the marks of two English essays awarded by professors in English who were labeled as experts and professors in other disciplines who were labeled as lay markers. The reliability coefficients for marks awarded by individual experts or pairs of experts were slightly higher than those awarded by lay readers or pairs of lay readers. The differences were not significant. Hence the authors concluded that the reliability of the two groups was similar. Differences in reliability coefficients were greater between essay questions compared to between types of marker indicating that reliability was more responsive to the type of question or variations in the mean performance of the candidates compare to the markers' factor. The same findings were reported for measures of concurrent validity of the essay assessment. Experienced examiners' marks had slightly higher validity compared to those of lay markers, but the changes in validity for the different essay questions were significantly greater.

Shohamy, Gordon, and Kramer (1992) examined marker reliability in the marking of English as a foreign language (EFL) among markers who were

professional, experienced EFL teachers or lay people (native English speakers). Half of the markers were given instructions in one of the three marking procedures namely holistic, analytic and primary trait scoring used in the study. Relatively high interrater reliability was achieved by the four groups of markers (trained/professionals, untrained/professionals, trained/lay and untrained/lay), irrespective of their training, but the overall reliability coefficients were higher for trained raters than they were for the untrained ones. As assessed by three criterion measures: Diagnostic Test of Written English; Test of Standard Written English; and grade point average across all college or university courses.

Training tends to have greater impact on marking, compared to the impact of markers' background. The same trend was found for all three of the marking procedures used. The implication of this study is that markers were able to score reliably, irrespective of background and training. Nevertheless reliability increases significantly when raters are given intensive procedural training. Thus Shohamy et al said that:

> the practical implication of this finding is that decision makers, in selecting raters, should be less concerned about their background, since that variable seems not to increase reliability. More emphasis,

> however, should be put into intensive training sessions to prepare raters for their task. (p. 31)

Another study on the scoring of English test by Lumley, Lynch and McNamara (1994) had doctors and trained Occupational English test makers rate the general communicative competences of 20 candidates of the Occupational English test. Difference between the two groups of raters in terms of leniency was not significant. Contrary to expectation the doctors were slightly more lenient. Generally all the doctors, except one, interpreted the scale reliably with the experienced raters.

The National Foundation for Educational Research (NFER) studied an online marking of pilots for Year 7 Progress Tests in mathematics and English. One of the issue they examined was the impart of using unskilled and semi-skilled examiners to mark some selected items (Whetton & Newton, 2002). The report showed that with some kind of intervention by supervisors would make this strategy technically robust.

Pinot de Moira (2003a) examined the relationship between background and marking reliability of examiners in seven GCE subjects. In the study reliability was defined as; the absolute difference between senior examiner and assistant examiner mark which entails whether an

adjustment is made to the assistant examiner's marks coupled with the rating of the examiner's performance (the rating scale range from A - consistently excellent, to E - unsatisfactory not to be re-employed). She discovered that the composition of an examiner's script allocation in terms of centre type had far more effect on accuracy than accessible aspects of an examiner's background, such as years since appointment. The only personal characteristic found to be significant in explaining examiner reliability was the number of years of marking experience. Royal-Dawson (2004) pointed out however that this characteristic was confounded because reliable examiners are engaged year after year and poor markers are not, so quality of marking and length of service are not mutually exclusive.

Royal-Dawson (2004) studied the marking reliability of four types of markers who have academic background in English with different amounts of teaching experience namely: English graduates, PGCE graduates, teachers with three or more years' teaching experience and experienced examiners.

Reliability was defined in the follow ways: the correlation between the marks awarded by the Lead Chief Marker to scripts, and the marker; the

agreement between the levels assigned by a marker and the Lead Chief Marker to pupils; the number of administrative errors. Generally there was no significant difference in the marking reliability of the different types of marker. Accurate markers were found almost equally in each of the groups, any other.

Marking reliability as defined and revealed in the study showed that some teaching experience was a contributing factor to higher reliability estimates only in some tasks. Besides the sub-test for reading where the experienced markers were more lenient than the other marker groups, there was no difference in lenience or severity between the marker groups. Royal-Dawson concluded that the criterion of teaching experience could be dropped so that markers with graduate-level subject knowledge could mark Key Stage 3 English tests.

Powers and Kubota (1998a) examined if individuals who are not teaching in tertiary institutions could reliably mark essays written by college students who are seeking admission into graduate programmes in business management. Thus they compared the marking reliability of experienced and inexperienced examiners. The experienced markers had previous experience in the holistic marking of essays for one or more Educational

Testing Service (ETS) programs; had graduate degrees and taught university courses with critical thinking skills or writing. The inexperienced group basically consists of persons with graduate degrees or persons without teaching experience in college level courses involving critical thinking skills or writing and had never participated in holistic scoring of essays. But all the members in this subgroup had a baccalaureate degree.

Essays were marked before and after training. After training, inexperienced markers in particular, improved considerably in their judgment of the 'correct' scores. Contrary to expectation many of the inexperienced markers were not lesser than the experienced markers in accuracy even before the training. The researcher in conclusion said that there were 'few significant relations between experience and accuracy' and "that the current pre-requisites for ETS essay markers would automatically disqualify a proportion of potential markers, who could, after training, mark accurately" (Meadows & Billington, 2005)

According to Meadows and Billington (2005) it is unfortunate that the design does not extricate teaching experience and subject knowledge. It is likely that these are differentially important" as revealed in the study of

moderation systems in New Zealand by Ham (2001) that moderator or assessor experience was relatively more important than subject experience for marking reliability.

Powers and Kubota (1998b) carried out a second study, building in their previous study essay writing prompt – 'analysis of argument' which is used to select applicants for graduate programs in management. Like the previous study the results showed that inexperienced markers without the usual requirement can be trained to score 'argument' essays with a high marking reliability. From the logical reasoning scores they award to markers, they saw a possible relationship between logical reasoning and markin reliability.

Effect of Teacher's Personality on Marking Reliability. Efforts to see if there is possible relationship between personality traits and marking reliability have been made. Painfully however, the scale of work done is small and the problem of the lack of clear-cut personality measures, make it difficult to make a rational interpretation of the possible relationship between examiner characteristics and marking reliability.

Branthwaite, Trueman and Berrisford (1981) investigated the relationship between markers' personality rating on the Eysenck Personality

Questionnaire and the marks they awarded to essays. The marks given did not correlate with extroversion, neuroticism or psychoticism scores but correlated positively with scores on the lie scale. This was interpreted as indicating that low marking reliability might be attributable to the different levels or types of desire among tutors for social acceptability.

Pal (1986) carried out a study and compared the Meenakshi Personality Inventory scores of two groups of four examiners who were catergorized as efficient and inefficient on the basis of their marking reliability in marked awarded to twenty scripts of high school students in the subject of Hindi. Efficient examiners relatively had high needs for achievement and dominance, but low needs for affiliation.

Greatorex and Bell (2002a and b) requested some examiners of GCSE English, Food Technologies and History to complete the Bem Sex Role Inventory. This instrument gives a measure of self-reported possession of socially desirable, stereotypically masculine and feminine personality traits. Examiners who rated themselves highly on the masculinity scales were more likely to be Team Leaders. The masculinity scales are made up of dominant/assertive traits and self-sufficiency/decisive traits. Greatorex and Bell saw this as unsurprising since Team Leaders need to be decisive.

The appointment of Team Leaders is under the control of awarding body staff, who presumably perceive these traits to be important in fulfilling the Team Leader role. Team Leaders did not however rate themselves highly on traits that could be useful for developing people skills, which is another important aspect of the role.

Effect of Teacher's Mood on Marking Reliability. Mood is another feature of the marker that may have significant influence on marking reliability. Townsend, Yong Kek and Tuck (1989) put some markers through a film show designed to create a positive or a negative mood state. The markers thereafter graded nine essays, comparing some target essays. Secondary school students wrote the essays entitled "our hopes and aspirations in the next decade". Mood affected only the grading of the first essay. Even if there was no significant influence of mood on scoring (with the exception of the first essay scored) there was a pattern though relatively short lived, for higher grades to be awarded in the negative mood condition. Townsend *et al* justified their findings with the theory that helping or pro social behaviour have a self-gratifying function that allows people to relieve their own wories (Cialdini, Darby and Vincent, 1973).

***The Effect of Question Format on Marking Reliability*.** Many studies have revealed that more closely defined questions format, which call for definite answers, tend to have higher marking reliability. Hill, (1973); James, (1974); Murphy, (1978), found that the multiple choice test is the highest closed response test. Multiple choice tests are referred to as objective tests' simply because no form of judgment is required for the assessment. This implies that they can be marked with perfect reliability.

According to Meadows and Billington, (2005), the US has the most elaborate development and the highest use of objective tests at all levels of education while its development in the UK has been cautiously slow. Pillner, (1968) and Wolf (1995) justified the US reliance on the multiple choice test on the grounds of its strengths. Such tests according to them ensure fairness or objective testing on a large scale with minimum cost. "Moreover, since results do not vary according to marker, there is less scope for candidates to appeal (a factor that is particularly important in a country where litigation is widespread)" (Meadows & Billington, 2005).

According to Pillner (Meadows & Billington, 2005) remarked that the one-way nature of objective tests is demonstrated by 0.95 or above correlation between two NFER or two Moray House reasoning or

attainment tests administered with an interval of forty days and the coefficient of equivalence in line with the notion of substituting one test for another on the single occasion of testing is rarely below 0.98

He seemed to have little reservations against the use of objective tests for educational measurement, thus he says:

> where the nature of the domain examined allows of it, 'objective' questions which require no evaluative judgement in their marking should be used. Where the nature of the domain calls for extended writing, the attendant difficulties of marking consistently have to be accepted. (p. 170 cited by Meadows & Billington, 2005)

Objective tests gives more reliable results always than short-answer or essay questions though, it is strongly debated if such tests achieve high reliability at the cost of invalidity. Objective tests are often seen as an invalid way to measure writing ability, and all other theoretical skills. In any case research has shown that there is a correlation between holistic ratings of essays and objective test marks (Charney, 1984), and has also revealed that objective tests are better predictor of the quality of essays compare to other essay tests (Breland, 1977). Wilmut, Wood and Murphy

(1996) suggested that the usefulness of objective tests be re-evaluated in view of the increased ingenuity and sophistication of response formats (Case & Swanson, 1993).

The quest for validity of assessment has restricted the use of these types of questions to certain subject areas. Examinations in such subjects that are mainly quantitative require definite questions and precise answers as in objective questions which tend to lead to high interrater correlations. More rigid disciplines also use more comprehensive mark schemes which further enhance marking reliability. Definite mark schemes reduce the need for judgmental marking.

"It is often the case, however, that in subjects such as English it is not possible to specify precisely how each mark will be allocated. In these subjects, the task of the examiner is to interpret the quality of candidates' work" (Meadows & Billington, 2005). Essays by their nature of being extended, free-response items they make reliance on a detailed mark scheme that can be prepared before marking, used systematically devoid of examiner's professional judgment difficult. The inevitability of interpretation and the subjectivity associated with it are the major causes of low reliability of marks.

Hill (1975) examined the marking reliability of final examinations in B.Sc Engineering. He discovered that the correlations between marks awarded by different examiners were far higher with 'problem' type questions compared to essays.

In a similar study, conducted by Murphy (1982), on marking reliability among some O and A level subjects, he found, that examinations that have more of essay-type questions were the least reliably marked while examinations that have highly structured, analytically marked questions were the most reliably marked

James (1974) examined the marking of physics scripts, which are highly quantitative. Fifteen papers were rated by six examiners. The standard deviation of the marks of the six examiners from the mean mark for all candidates was just 4.1 marks on a paper whose maximum obtainable mark was 100. The mean value of the correlation coefficient was very high (0.94), implying high interrater reliability.

Murphy (1978, 1982) carried out extensive meta-analysis of the reliability of marking for the AEB on twenty, O and A level examinations. A senior examiner for each of the subjects re-marked scripts from a randomly

selected sample of about 200 candidates. Out of the eight subjects he examined English was the least reliably marked and Mathematics the most reliably marked.

In another study, Newton (1996) sought to know if standards of marking reliability had been maintained in view of changes to assessment such as the replacement of O level with GCSE. He carried out a comparative study of the marking reliability of the 1994 SMM GCSE examinations in mathematics and English. He discovered the marking reliability of mathematics was very high while that of English was conspicuously lower. He asserted that the cause of this significant difference in reliability was the contrasting nature of the two subjects.

As Newton, says the highly detailed mark schemes in mathematics are partly the cause of the high reliability obtained. He did not infer that the awarding bodies were not leaving up to the standards of assessment of English. Instead he argued that awarding body must take into cognizance the implication of reliability of assessment on validity and cost-effectiveness. In conclusion, He says that problems of low reliability in English cannot be completely avoided in the face of current assessment formats. In view of this problem, some authors maintained that essay

represents a serious problem to examination reliability. “Both the interrater reliability and intrarater reliability of essays have been shown to be problematic” (Meadows and Billington, 2005).

Akeju (1972) photocopied the same 100 West African Examinations Council GCE English composition scripts and were given to ten different examiners to mark. The correlations between examiners were between 0.51 to 0.76.

Lucas (1971) studied the interrater reliability of essay tests under real conditions, using part scripts of an Australian Matriculation Biology examination. The experiment consisted of six examiners marking the same 44 scripts during their official examination marking session.

Consequently tendencies for wide discrepancies between markers were reduced. For instance, a restricted mark range of 0-6 was used, with 0 reserved for candidates who failed to answer the question or whose answer was wrong. Furthermore, the markers were required to mark to a distribution of scores. Despite these efforts to reduce marker inconsistencies, the results revealed that only one out of the 44 scripts had been awarded the same score by all six markers; 19 scripts had a range of

two marks; 12 had a range of three marks; 12 scripts were awarded a mark of 0 by one examiners and 3 by another. Another script was awarded both a 0 and a 4. Lucas said "*Clearly there was no agreement on what constitutes completely false interpretation of biological concepts or complete irrelevance*" (p. 82).

Lehmann (1990) conducted a study to examined four sources of variance namely between and within markers, between topics and within students in the assessment of writing achievement. Despite the use of clear-cut criteria for assessment, about 12 per cent of the variance in final scores was attributed to the variances between and within markers. The low levels of inter-rater reliability and intra-rater reliability in the assessment of essays is not a new development.

The Effect of Candidates' Choice of Essay Question on Marking Reliability. It seems that the problem of low marking reliability of essays is magnified by the candidates' choice of essay topic. Coffman (1971) revealed that the marking reliability of essay will be lower if the subject matter is discursive and indefinite. Hake (1986) discovered that essays of pure narratives of personal experience were misgraded much more frequently compared to expository essays using personal narration to

buttress or defend an assertion. Hamp-Lyons and Mathias (1994) revealed that essay topics that were judged more difficult by composition specialists appears to get higher marks compare to those judged to be easier, and suggested that raters may be instinctively rewarding test takers who choose the more difficult prompt or were surprised by the testees' level of responses to such question

The Effect of Marking Scheme on Marking Reliability. Research has shown that substandard mark scheme can be the major cause of inconsistency in marking. Delap (1993a, 1993b) carried out marking reliability studies in 1992 AEB GCSE Business Studies and Geography examinations. The studies were conducted to find out the extent of any inconsistency in marking and to work out ways of ameliorating the problem. Meetings were held with examiners after the re-marking of scripts, to discuss the results and the problems encountered during marking. The source of the major problems in the two subjects was linked to the mark scheme.

Consequently improvements to the mark scheme have been cited more often than not as the solution to the problem of inconsistency in marking. Price and Rust (1999), maintained that, the use of detailed assessment

criteria increases in marking consistency excepting few cases. Moskal and Leydens (2000), had a similar position that improving the scoring rubric will improve both interrater and intrarater reliability. They gave some questions that might be helpful in assessing the goodness of any particular rubric. These are as follows:

1. Are the scoring categories well defined?
2. Are the differences between the score categories clear?
3. And would two independent markers arrive at the same score for a given response given the scoring rubric?

Positive answer to these questions is the only guarantee, according to them, for consistent marking among examiners. The use of exemplars was also part of their recommendations. The marker may refer to the exemplars during the scoring session to keep in focus the differences between the score levels. They also maintained that the rubric should be pretested. Any ambiguity in interpretation and amendments should be discussed by markers. Although this may take considerable time it will yield significant reliability (Yancey, 1999).

Despite their emphasis on the importance of the scoring rubric in producing reliable marking, Moskal and Leydens maintain that teachers

who depend solely upon the scoring criteria during the evaluation process may be less likely to recognise inconsistencies between the observed performances and the final score awarded to the candidate. In other words, unexpected but correct responses may be mistakenly marked down.

Saunders and Davis (1998) investigated the making and use of assessment

criteria for the undergraduate projects of management students. With data from two workshops markers evaluated the same undergraduate projects, using the criteria, with lecturer and student feedback, the authors made some suggestions for good practice as follows:

1) First, they maintained that the collective development of criteria by the markers will ensure a good beginning for each lecturer to have a common understanding of the criteria. This would be helpful because it is the only way to create community of assessment practice.

2) They say that since *"over time understanding and application of criteria will alter" (*p.167), criteria need to be debated occasionally avoid departure from standards

3) They extolled the imperative of clear-cut assessment procedures and the belief that these procedures need not be a cog in the wheel of progress. Research according to them as shown that carefully constructed mark schemes and use diligently reduces inconsistencies. *"They enable*

lecturers to be more certain they are following the same process and judging each piece of work against the same criteria, thereby assessing each student the same way." (p. 165)

4) Particularly, they suggested that procedures should not only relate to administrative protocols, but also to matters such as time spent marking each piece of work. Their study shown that spending too much time over assessing a student's work may result in a lower grade. Although Saunders and Davis' study was on the consistency of assessing projects from a lecturer's point of view, their suggestions are clearly useful to the use of mark schemes in evaluating examination scripts.

"Despite the pervasive view that a clear and detailed mark scheme results in higher marker reliability, intended improvements to the mark scheme do not always bring about expected improvements in reliability" (Meadows & Billington, 2005).

Subsequent studies reported similar futility in attempt to increase marking reliability. Baird and Pinot de Moira (1997) altered the GCE Business Studies mark scheme in order to assess its impact on the marking process. Baird, Greatorex and Bell (2002, 2003) carried out further study to measure the impact of increasing the specifications in the mark scheme

while varying the styles of standardisation meeting. *Neither analysis supported the hypothesis that marking reliability was affected by the different conditions applied* (Meadows & Billington, 2005).

Moreover, there is evidence supporting the fact that consistency in marking sometimes can be achieved only when assessors use their own criteria without assessment criteria. Wiliam (1996) say that teachers from a 100 *per cent* coursework in GCSE English learned to agree on the grade appropriate for a given student's work, despite the fact that there were no specific criteria and agreement on which aspects of the work were most significant for receiving a particular grade. This is described as construct referencing. Mark schemes differ in teachers' perception as to the degree they support objective or subjective methods of scoring. If judgment is not required by the scorer, in scoring, it is deemed as objective. This type of scoring as indicated earlier is possible in multiple choice items. Where marking is subject to interpretation is required, as in short answer responses and extended writing, scoring is considered to be subjective. "*In general, the less subjective the scoring, the greater agreement there will be between two different scorers (and between the scores of one person scoring the same test paper on different occasions).*" (Hughes, 2003, p. 22)

However, some experts maintained that no test is completely objective. Hamp-Lyons (1990) asserted that "'*Objective scoring' can be carried out only when humans have decided what the correct answers are"* (p.78). Pre-defined mark scheme (scoring rubric or rating system) minimize the subjectivity involved in rating short answer questions and essays, Hence increasing rater reliability (Moskal, 2000).

Effect of Training and feedback on Marking Reliability. Most examination bodies use a couple of days before commencement of live marking of scripts to train markers on the scoring rubrics and other necessary instructions they need to know. Many persons believe that Training is a 'crucial component' of the marking process because it compensates for different examiner backgrounds, moderates examiner's differences so that any variation in the process resulting from varying expectations is reduced (Charney, 1984; Huot, 1990).

Few studies though have experimentally control examiner training to find out which aspects of a training programme are fruitful and why, there are certain studies carried out to assess the overall effectiveness of different types of examiner training. An analysis of the marking and verbal

protocols of four inexperienced markers of the ESL composition placement test at UCLA, before and after training was conducted by Weigle (1994). She discovered that: "training was effective in bringing the four new, initially aberrant raters 'more or less in line with the rest' in terms of both marks and the procedures by which they arrived at those marks". Other training attempts have not been as successful as this.

Rasch multi-faceted analysis was used by Lumley and McNamara (1995) to compare ratings given on three different occasions, before and after training, by experienced raters for the speaking sub-test of the Occupational English Test. They discovered that "a substantial variation in rater harshness, which training has by no means eliminated, nor even reduced to a level which should permit reporting of raw scores for candidate performance" (p.69).

Considering the research evidence of differences in severity between raters after training, McNamara (1996) maintained that "assessment procedures which rely on single ratings by trained and qualified raters are hard to defend" (p. 235). He declared that the usual purpose of rater training "to eliminate as far as possible differences between raters – is unachievable and possibly undesirable" (p. 232). The true aim of training,

according to him, is to make new raters concentrate and become self-consistent. There is overwhelming empirical indication in favor of McNamara's (1996) viewpoint.

Lunz, Wright, and Linacre (1990), Stahl and Lunz (1991), and Weigle (1998) maintained that though rater training cannot make raters replica of each other, it can make them more self-consistent. Weigle (1998) asserts that the self-consistency will lead to improved accuracy in examinee's performance since variations in severity among markers can be predicted and modeled for compensation mathematically. Freedman (1981) discovered that a few strategic words by the head examiner at the beginning of a session could influence the marker consistency significantly.

Another form of education in the marking process that may impact on marker consistency is Feedback from senior examiners. Wigglesworth (1993) found some evidence that examiner biases were reduced and that interrater reliability improved because of feedback. Breland and Jones (1988) gave essays written by undergraduate students to two sets of markers to mark. One set is examiners working in a conference setting and second set of examiners are those working in their own homes or

offices. The conference markers were trained on the specific rubrics for scoring and were monitored by group leaders. The markers at home received written instructions by post without monitoring of their scoring, hence no opportunity, for discussion and feedback, at the conference. Reliability was 0.75 for conference scoring and a mean of 0.62 for three examiners at home. This result shows that interactions with group leaders and other markers at the conference increased marking reliability.

Shaw (2002) studied if an iterative standardisation practices will improve the interrater reliability of multiple rating of the same set of scripts. The markers received their first training in an interactive meeting in hierarchical co-ordination meeting. As part of the training markers were made to mark a set of scripts. The markers were given training materials with each batch of scripts sent to them. There were clearly written feedback notes on each script in the batch previously marked. The expectation was that a gradual improvement in interrater correlation would accompany each successive feedback. The results showed however, that though the interrater reliabilities were moderately high (0.77) they did not increase with time and standardization but fixated. And interestingly the before and after training in standardization coefficients did not differ significantly. Thus Shaw claimed that "the mark scheme, comprising a set

of detailed and explicit descriptors, engenders a standardizing effect even in the absence of a formalized training programme" (p.16).

Furneaux and Rignall (cited by Meadows & Billington, 2005) examined the judgments of twelve trainee examiners for an International English Language Testing System writing module. The report goes as follows: On successive occasions, before and after training, the examiners rated a set of eight scripts and wrote brief retrospective reports about their rating of four of the scripts. The examiners' scores before training did not differ as greatly from the standard as might have been expected.

Furneaux and Rignall drew a conclusion similar to that Shaw (2002). They affirmed that standardizing effect may have resulted from the use of a rating scale with detailed band descriptors. Furthermore, they claimed that the examiners' similar professional background may have contributed. The need for examiner training may be undermined by the use of an explicit mark scheme. Examiner training, however, often occurs in groups. It is an opportunity for examiners to meet together and discuss issues related to marking or related to their subject area. These meetings may help engender a 'community of practice', which some believe to be crucial for reliable marking.

Marking Methods for Improving Marking Reliability in Essay.

Some suggested measures to improve the marking reliability of essay have been empirically tested. In the US, for instance Quality Rating Scales were produced to provide exemplar material at various levels. Examiners were requested to make final awards by comparing scripts against the exemplar material. In England, efforts to improve marking reliability included identifying the criteria for assessment "and assessing each one separately. The final mark was therefore the product of several separate assessments, all made by the same assessor. Unfortunately none of these initiatives yielded the desired increase in interrater reliability" (Meadows and Billington, 2005)

In line with this desire many scoring methods are currently being use by markers. The two most popular ones are: holistic and analytic

.

Holistic and General impression Scoring. Holistic scoring entails the rating of a piece of work informed by an overall impression of it Individual characters of the work, like grammar, spelling, ideas and organisation are not evaluated as distinct parts Hamp-Lyons (1990)

In contrast, analytic scoring procedures require markers to assign a discrete score to each of a number of aspects of a task. In an essay, for

example, these might be as follows: grammatical accuracy, vocabulary, idiomatic expression, organization, relevance, or coherence. Thus, analytic scoring is slower than holistic, but provides more diagnostic information about candidates' ability.

Some research studies have been conducted to assess the reliability of general impression or holistic and analytic methods of marking. Kaczmarek's (1980) work on the marking of essays written by students learning English as a second language, the scores awarded by examiners using holistic or analytic scoring rubrics correlated highly. Kaczmarek (1980) in conclusion noted that subjective methods of marking essays 'are almost as good as objective scoring methods for students learning English as a second language.

Nevertheless, the holistic scoring of writing has also been attacked on the premise of invalidity. Charney (1984) said that:

> Early attempts at qualitative evaluation of writing samples were abandoned because they were unreliable, not because they were invalid. However, the widespread confidence in the validity of current qualitative assessments must surely be tempered by considering the method of obtaining those

> assessments. Not any qualitative method will automatically be valid, even if it produces reliable results." (p. 77-78)

He continues as follows:

> The validity of holistic scoring remains an open question despite such widespread use [;] the question of whether holistic ratings produce accurate assessment of true writing ability has very often been begged; their validity is asserted, but has never been convincingly demonstrated." (p. 206)

In the opinion of Charney holistic ratings may have high interrater reliability *"largely because they depend on characteristics in the essays which are easy to pick out but which are irrelevant to 'true' writing ability"* (p. 75). According to her such characteristics included the following: quality of handwriting, word choice, length of essay, and spelling errors. In a study to examine what goes on in a rater's mind when they mark essay holistically. Vanghan (1991) also threw his weight against holistic scoring. Her Think-aloud protocol analysis revealed that:

> *raters are not a tabula rasa, they do not, like computers, internalize a predetermined grid that they apply uniformly to every essay. Despite their similar training, different*

raters focus on different essay elements and perhaps have individual approaches to reading essays. (p.120)

In addition, the academic environment in which holistic scoring usually takes place is not free. Thus Charney (1984) evaluate the methods used for controlling examiners using holistic rating as *"peer pressure, monitoring and (insistence upon) rating speed."* (p.73)

Analytic Scoring Method. There is some evidence that under certain situations the analytic method might be more reliable compare to holistic or impression marking though it is more demanding and time-consuming. Wood (1991) acknowledged this long time factor of analytic scoring method.

On the contrary, it is argued that if a number of holistic readings can be given to a script in the same time that it takes to award an analytic score, it is better to use the judgement of several examiners rather than compounding the error of a single examiner assigning three or more different scores (Cooper, 1984).

Doubts among some researchers whether analytic scoring methods is truly effective have been expressed (Meadows & Billington, 2005). Moreover, Hamp-Lyons, (1986: cited by Park n.d.) argued that some analytic scoring

schemes sometimes pose difficulty to even experienced essay markers trying to award scores in line with certain descriptors. This problem was manifested in the study published by Delap (1993a and b) and in a study carried out by UCLES (2000). Their study examined three possibilities of ensuring consistency between markers of Key Stage 3 English by comparing the marking of four different groups of experienced markers who marked the same scripts. The marking of the group revealed that the analytical method was greatly affected by this procedure; their marking reflected some measure of severity and instability.

Hughes (1989) maintained that to use an analytical mark scheme compels markers to focus on different features of the work and that this might distract attention from the overall quality of the write up. And since the whole is often greater than the sum of its parts, a composite score might be very reliable, and yet invalid.

Foley (1971) also recommended that markers adopt a global or holistic method in awarding scores, instead of an analytic one to take into cognizance the whole rather than the part. "The analytic method of scoring may fragment effects that remain intact in global reading" (Meadows & Billington, 2005).

Scoring with Exemplar Material. Exemplars are used as a means of introducing the inexperience markers into an academic community. Wolf (1995) asserted that standards are enforced by using examples of students' work instead of assessment criteria. She maintained that in the absence of assessment criteria from work candidates could be rated differently.

For Wenger (1998) "the process of reification means that examiners must discuss exemplar material" (Meadows & Billington, 2005).He defines reification as imputing the status of an object on human experience, for example, treating the concept of mathematical ability as though it is an object in human experience. In creating mark schemes, we reify the constructs we are measuring and it is important for examiners to discuss examples before they can have a common knowledge of the concepts. Lave and Wenger (1991) recommended that the final work of 'masters' can be used as exemplars in the process of 'apprentices' becoming full participants. This means that assistant examiners should follow more experienced examiners' marked scripts as an example. This is actually the technique of improving marking reliability used by awarding bodies in the UK.

Irrespective of some commentators' demand for the discussion of exemplars by examiners during the marking session, it is necessary to acknowledge that there are some problems with this method of acquainting examiners with standards. Different examiners comprehend exemplars differently (Baird, Greatorex and Bell, 2002, 2003). More still Sadler (1987) shows that exemplars of the same standard vary and they are weak as an index of standards because they can accommodates" factors like cultural tradition and current technology, which implies they quickly become outdated. In conclusion, he said that a small number of exemplars alone cannot accurately define a standard if multiple criteria are involved.

In response to Wolf's (1995) argument, there are unexpectedly little empirical research on the value of exemplars in promoting common standards. Baird, Greatorex and Bell (2002, 2003) examined the impact of exemplar materials on marking reliability. The impact of prototypical band exemplar scripts and cut score exemplar scripts were compared to find out if scripts of prototypical examples of a particular band are better than scripts that serve as examples at the cut score between bands. Three groups of examiners were used; one marked without exemplar scripts, one

used prototypical exemplar scripts while the third one used cut score exemplar scripts.

Contrary to expectation, the most accurate marking was achieved by the group who did not use exemplars. Examiners who used prototypical exemplars had more severe scores than those who did not use exemplars or cut-score exemplars. Baird *et al* said that examiners may be used to thinking about cut-scores and cut-score performances and perhaps the prototypical exemplars were read as cut-score performances by the markers. Since the prototypical exemplars had higher marks than the cut-score, this would make their marking more severe. Baird *et al* advocated that examiners should be given exemplars which portray the range of achievement for each mark band.

Double and Multiple Marking. **D**ouble marking entails that two examiners seperately, assess the same scripts. The final mark is pooled from the two individual marks while 'Multiple marking' means that two or more examiners separately, assess the same scripts and the final mark is pooled from the individual marks. As far back as 1949 Wiseman published the findings of a study on multiple marking of English composition scripts of 11-plus candidates.

Groups of four markers rated each script separately and the final mark for each script was the total of four separate marking. He posited that the multiple marking elicited reliability coefficients as high as 0.95 thus comparable with those of objective tests. Other researchers such as Lucas, (1971) raised objections, pointing out whether Wiseman was in fact assessing interrater reliability or the mark-re-mark reliability of the markers. Markers were trained by Wiseman to use general impression marking to ensure that marking is fast to make it worthwhile.

Importantly markers were not selected for interrater agreement but for their high levels of self-consistency in fact they had to attain a mark re-mark correlation of 0.7 or above in pretest as precondition. Wiseman may have been the first to attest to the value of individual markers differences as expressed in the following statement:

> *Provided markers are experienced teachers, lack of high inter-correlation is desirable, since it points to a diversity of view-point in the judgement of complex material, i.e., each composition is illuminated by beams from different angles, and the total mark gives a truer 'all-round' picture* (p. 206).

Wood and Quinn (1976) noticed that since the work of Hartog and Rhodes, disagreement between examiners was no longer acceptable. In any case, Britton Martin and Rosen (1966), however, defended Wiseman, that differences arise from the most delicate areas of individual differences, which ought to be part of assessment.

When there is a reasonable measure of agreement among individual markers about the scripts' merits, the aggregated marks from a team of markers will be a valid expression of the team's consensus of opinion, the reliability of which will increase as the size of the team increases. The high level of reliability obtained from double marking as revealed in literature has prompted its use by awarding bodies in examinations involving subjective assessment and new subjects, between 1969 and 1980, Brooks (1980) for instance showed that in the late 1970s a good number of GCE and CSE examination boards were using more than one marker to rate English Language composition scripts.

The awarding bodies carried out some unpublished studies of the increase in reliability obtained through double marking. The Joint Matriculation Board (JMB) Research Unit (Meadows & Billington, 2005) carried out an evaluation study of double marking of two A level General papers and

two O level English Language papers. In English one examiner used impression marking while one used analytic scoring method. The final mark awarded to the candidate was the sum of the two separate marks. The result shows that there was no difference between the mean marks assigned by the two marking methods. The correlation between the two markers was 0.45 and 0.60 respectively. *"If just the analytical marks had been used then 6.1 per cent of candidates would have changed grade on one paper and 6.4 per cent on the other paper" (Meadows & Billington, 2005).*

For the general papers each essay was marked twice and assigned an impression mark on a scale of 1 to 9. The aggregate marks make up the final mark awarded to the candidate. The correlation between marks was 0.70.

> *If just the first impression marks had been used then 6.9 per cent of candidates would have changed grade; if just the second impression marks had been used then 7.3 per cent of candidates would have changed grade. The research concluded that double marking continue (Meadows & Billington, 2005).*

Lucas (1971) also examined double marking with Biology essays, under operational rules. In the course of official marking, six markers also marked the same 44 scripts by general impression with a scale from 0-6. Interrater reliability was estimated on the basis of whether one, two, three or four separate marks contributed to the final mark. This allowed the relative benefit from scaling up from single, to double, to multiple marking to be measured. Lucas discovered that multiple marking significantly increased the reliability of the marks assigned, though the highest increase in reliability came from an increase from one to two raters. The increase in reliability arising from each additional marker waned as the number of raters increased. Any additional benefits obtained from using groups of three or four markers were though significant was in descending order. This phenomenon was confirmed by Akeju (1972). Lucas opined that the increase in reliability has to be weight against the additional sources required.

Wood and Quinn (1976) examined whether these benefits in reliability from multiple marking would be a general phenomenon in all subjects. English Scripts from O-level English Language were rated by markers under the same conditions as in operational marking. The scripts comprise essays and summaries. Prior to their briefing in analytical marking, the

method used by the board, ten examiners rated the same 100 scripts by general impression marking on a nine-point scale.

Wood and Quinn firmly believed that though reliability can be reduced by marker bias and inconsistency, the former can be easily corrected, the real problem is the latter because it is not easy to correct. They observed that double marking result in greater consistency compare to single marking.

They also investigated the impact of pairing examiners according to known characteristics of their marking behaviour but discovered little advantage of systematic ordering over a random approach. They opined that in - between marker correlations of 0.50 to 0.60 are adequate because a minimum level of disagreement is helpful. They maintained that the advantages of double marking in respect of increased reliability according to Meadows and Billington, (2005) offset the reduced spread of marks caused by regression to the mean and the consequent reduced discrimination between different levels of achievement"

In addition Wood and Quinn say that the impact of switching from analytical marking to impression marking without multiple marking affects a candidate's mark just as if a different examiner marked him or her. Double marking among awarding bodies is no longer practiced,

mainly because of the increasing problems with the supply of markers. Awarding bodies manage to recruit sufficient markers to mark scripts once, let alone twice." Double marking of all examination papers is not a feasible option. There were approximately 5,712,588 GCSE and 2,794,188 GCE examination scripts marked in summer 2004 by the AQA alone" (Meadows & Billington, 2005).

Double marking is, however, common in Higher Education and there has been some assessment of its effectiveness. Partington, 1994; Smith, Sinclair, Simpson, van Teijlingen, Bond & Taylor, 2002; Sparks & Ballantyne, 1997 are examples of works in this area of research (Meadows & Billington, 2005).

In this regard, Chaplen (1969) examined the blind double marking of an essay as part of a university entrance examination in English for foreign speakers of English. The examiners re-marked the essays after three months and the two sets of marks were then correlated to obtain a coefficient of stability of each marker. The essays were rated by impression on an eight point scale. The points on the scale were described in detail. When examiners mark two instead of one essay reliability was

increased and when examiners double mark reliability was also increased. The overall reliability when examiners double mark two essays was 0.92.

The assessment of the effectiveness of double marking is imperative because it sometimes lead to unexpected results. Newstead and Dennis (1994) requested 14 experienced Psychology examiners, who were external examiners to other courses, to rate the same six students' scripts. The variation in marks was notorious, A typical example was an essay that was awarded a first class from one examiner and the same was on the point between second and third class from another. Newstead and Dennis (1994) argues as follows: *since students' degree classes are based on a number of examinations, measurement error like that may likely lead to misclassification only for students whose GPA are close to the borderlines*

Partington (1994) examined the worth of double marking in Higher Education. He maintained that double marking is not the solution for clear evaluation guidelines and marking rules. More still, double marking would not be adequate without the latter.

Not long ago, Smith, Sinclair, Simpson, van Teijlingen, Bond and Taylor (2002) carried out a study of double marking of an essay on a medical science course. The correlation between the two markers was low. The

markers were either academics who are not teaching the course or generalist teaching the course. Correlation was poor irrespective of whether the two markers were the same or different. The way to remove disagreement between markers was not in view. Many students would have received clearly different grades if the scripts were marked by one examiner rather than two.

In spite of resource constraint, the double marking of public examinations has currently received a boost. In 2002, QCA released the report of an independent panel of experts charge with the maintenance of standards at A level. With respect to quality of marking, the report suggested "*limited experimental double marking of scripts in subjects such as English to determine whether the strategy would significant reduce errors in assessment*" (p. 24,)

Newton (1966), said however that it is not certain which papers would benefit from double marking to compensate for the increased costs. Certainly GCSE mathematics, is not one of them. If the marking of two examiners were perfectly accurate the correlation coefficient between each set of marks would be +1.00. A high correlation in the neighborhood of +0.80 or 0.90 would indicate that the order of merit of the candidates is

almost the same for both markers. It is either the candidates are getting similar marks or ranks. The implication of this is that double marking is unnecessary.

A low co-efficient, less than +0.30 would indicate little relationship between the marks in most pairs and suggests that examiners are not assessing the same criteria. Using an aggregate of the two marks under these circumstances may bunch the candidates about the mean. Double marking strategies may be most appropriate when the coefficient is intermediate in value

The introduction by awarding bodies of double marking would change the entire philosophy of the whole exercise. With the current hierarchical system, the assistant examiners is under the oversight of senior examiners who are answerable to chief examiners, whose years of experience and good record makes them the custodian of standards for particular examinations. The philosophy is based on the assumption that marks are more accurate the higher the marker is in the hierarchy.

Double marking rests on a different view of what constitutes a ‘true mark’. Wiseman argued that the ‘true mark’ would be that given by the pooled judgement of an infinite number of markers. Wood and Quinn

agreed; defining the true mark as the average mark awarded by all the examiners. The best way of combining the marks generated by multiple marking has also generated some discussion in the literature (Meadows & Billington, 2005).

Cresswell (1985) mentioned four types of double marking. Firstly, the most ideal double marking according to Cresswell is a separate re-marking of the first marking, using the original marking scheme, without knowledge of the original marks awarded to the scripts. The second, is re-marking with the original marking scheme, with knowledge of the original marks. The Third, is a re-marking of the scripts by impression method using assessment criteria used in the original marking scheme without knowledge of the original marks. Finally, it could be re- marking by impression method with knowledge of the original marks.

Smith *et al* (2002) outline the options for combining the marks awarded to medical students' essays. These are as follows: The first to take average of the two marks; A second option is to use a third marker; to mediate between the two markers. The usual recommendation however is to add the marks from more than one marker as composite mark to form candidates' final scores(Coffman, 1971). Wiseman (1949, 1956) and

Pilliner (1969) demonstrated that an average mark when there is 'fair' measure of interrater correlated, gives a better mark. Cresswell (1983a) adopted a more statistical approach. He proved that the simple addition of the two examiner's marks will scarcely produce an aggregate score with the greatest possible reliability. Consequently he derived formulae for the weights that should be used in summing up the composite mark with optimum reliability.

Whatever the improvements in reliability brought about by double marking, the resource implications of its introduction may make it impossible to implement in the public examination system. Lamprianou (2004) suggested that a solution might be to have each script marked by a human marker and by software. In the case of a marking discrepancy, a second human marker would be called in for a second blind marking. This solution may be made possible by the range of writing assessment programs available: Project Essay Grade (Page, 1966), Intelligent Essay Assessor (Landauer and Dumais, 1997) and E-rater (Educational Testing Service), for example (Meadows and Billington, 2005). The argument for and against computer marking is discussed in the latter section of this review.

E-Marking. *I*deally the use of e-marking should improve marking reliability than traditional marking methods. E-marking warrants more effective monitoring of examiner reliability while marking is in progress thus permitting early intervention if any problem is detected in the exercise. Again the e-capture of marks deters examiners from recording marks outside the range allowed by the mark scheme. Considering the level of e-marking occurring in the US and the astronomical increase in the level of e-marking envisaged by UK awarding bodies, there are ironically few published studies of the relative reliability electronic marking. "Available studies show small and inconsistent differences in the reliability of the marking methods" (Meadows & Billington, 2005).

Twing & Harrison (2003) carried out a comparative study of paper-based and image-based marking of a writing test in the US. The marks obtained under the paper-based system were slightly more stable compare to the marks obtained under the image-based system. This was the case in all measures of reliability in the study:

> grades were the same or adjacent in 90.1 *per cent* of cases for image-based marking and 91.8 *per cent* for paper-based marking; the correlation between marks assigned by the first and second marker was 0.64 for image-based marking and

> 0.70 for paper based marking; the Kappa coefficient (which adjusts the measure of reliability for chance agreement) was 0.32 for image-based marking and 0.35 for paper-based marking. The authors described the differences in reliability between the two methods of marking as statistically significant, but not practically meaningful (Meadows and Billington, 2005).

Sturman and Kispal (2003) carried out a similar study electronic and paper based marking of three papers namely reading, writing and spelling in pupils between aged seven to ten. The scores awarded was influenced by paper, age and method of marking. Sometimes paper marking was more generous, and at other times e-marking was. They claimed that different issues of marker judgment arise in particular aspects of e-marking and conventional marking, but will not have a positive and negative effect on pupils in a consistent way. At the test level, analysis revealed highly comparable outcomes between the methods. Unfortunately double marking using each method was not included so no comment on the relative reliabilities of each method.

Raikes (2002) examined the reliability of paper-based and image-based marking of GCE Mathematics, Geography and English Literature scripts. Two types of on-screen marking were studied - whole script marking and individual question marking. In English Literature markers were slightly more severe on screen compare to marking on paper. They were most consistent on paper and least consistent when marking individual items on screen, probably because the scripts were split by question examiners could not be influenced by a candidate's performance on other questions. In Mathematics, markers used similar criteria and were equally consistent in the three methods. In Geography one marker was slightly more severe on screen and one was less consistent on screen than on paper. The increased severity with on-screen marking is not a problem since it affects all candidates equally. In conclusion Raikes says that on-screen marking of whole scanned paper scripts may be reliable as conventional marking, but individual question marking requires further investigation.

Fowles (2002) examined e-marking and conventional marking in GCE Chemistry. The relationship between the two sets of marks was high. Markers were no more severe or lenient when on-screen. The mean difference in total marks in all the scripts was only 0.13 marks. There was also a high correlation (0.99) between the two sets of total marks. E-

marking often entails examiners marking individual items instead of whole scripts. “Although part versus whole marking is a topic that might be expected to have received research attention, Fowles (2005) found little reference to marking individual items. She said that as e-marking becomes common there will be increased opportunities for empirical study of the belief that segmentation can ‘add to the objectivity of the marking’ (Bakker and van Lent, 2003). Williams and van Lent (2002) pointed out three specific factors that will contribute hopefully to the fairness of e-marking of parts: These are as follows:

i. the use of blind marking;

ii. little chances to build up a ‘halo’ effect’; and

iii. random distribution of a candidate’s responses to a group of markers, so that each examiner’s marking error will be randomly distributed among individual candidates.

***Computer Marking*.** Computer marking of candidates’ answers to closed questions is used regularly, but automated scoring of open responses is the interest of current research. Some approaches have been taken to automatic marking. Cohen, Ben-Simon and Hovav (2003) adopted the approach of using the computer analyse the mechanical features of the response, like the number of characters entered, the

number of sentences, sentence length, the number of low-frequency words used e.t.c.. The success of methods like this has been to compare the correlation between it and human markers, and the correlation between marks awarded by two groups of human markers. Cohen *et al* examined the marking of different types of essay by humans and computer, and said that the correlation between the number of characters entered by the candidate, and the marks awarded by human markers are as high as the correlation between marks awarded by human markers.

Ridgway and McCusker (2004) maintained that it is improbable that computer marking of this type would be used in the UK, because the UK regulations demands that marking schemes be made candidates and teachers' friendly. Moreover the attendant validity of these kind of marking systems would be "*dire*" (p.23). "The advice to candidates would be to improve their scores simply by using more keystrokes".

A second approach to automated scoring rate student responses on tasks where the range of acceptable responses can be deadly speak out; as in short answer science tasks (Sukkarieh, Pulman and Raikes, 2003,). From the analyses of a good number of student responses, an outline of correct and incorrect responses, synonyms for nouns and verbs and alternative grammatical forms are elicited. Student responses are parsed with

techniques from Natural Language Processing, and are compared with stored correct and incorrect responses, with the help of many forms of Information Extraction techniques (Cowie and Lehnert 1996).

A similar technique of to marking tasks with a more restricted range of correct responses was used by AQA (Fowles, 2005) which they called 'automatic marking'. Responses to items stipulated for automatic marking are all double-keyed. A list of responses with their frequencies is given to the senior examiner, who will mark each response on the list. The computer will allocate the mark determined for each candidate's response in line with to the senior examiner's marking rules. Fowles (2005) asserts that automatic marking is completely reliable because it will elicit the same set of marks on a second occasion of marking. However, a second set of scores might differ if another examiner provides the marking rules.

"It is doubtful if marking solely by computer will be acceptable in the foreseeable future" (Meadows & Billington, 2005). It is recommended by Lamprianou, (2004,) that a realistic and effective way of increasing marking reliability might be to mark each script by a human marker and by computer. In the case of significant differences, a second human marker would do a second blind marking.

Whole Script Marking and Individual Question Marking Methods. Meadows and Billington, (2005) said that: "Although part versus whole script marking is a topic that might be expected to have received research attention, Fowles (2005) found little reference to this aspect of marking". According to Bakker & van Lent, (2003), as e-marking becomes common there will be increased opportunities for empirical study of the belief that segmentation can 'add to the objectivity of the marking'.

However a report published by the university of Cambridge local examinations syndicate, according to Ucles, (2002) shows that Mathematics examiners appear to find question marking boring or less rewarding than marking whole scripts. Similarly some English Literature examiners claimed that marking a whole script enabled them to award a fair mark; Generally the report shows:

> For the Mathematics component, examiners applied similar standards and were similarly consistent across the three marking methods (on paper, whole scripts on screen, and individual question marking on screen);

> For Geography, although most of the marking was satisfactory, one examiner was a little more severe when marking on screen and one examiner, whose paper based marking was reasonably consistent, was inconsistent when marking on screen;

For English Literature, two examiners were a little more severe on screen than *on paper (it made little difference whether they were marking whole scripts on screen or individual questions). Examiners tended to be most consistent when* marking on paper and least consistent when marking question apportionments (on screen). This may have been because examiners were unable to be influenced by a candidate's performance on other questions when the scripts were split by question. The results indicated that with suitable modifications to the software used by examiners, screen based marking of whole scanned paper scripts would be likely to be as reliable as conventional marking. Individual question marking required more investigation, particularly for English Literature. Ucles, (2002)

More recently, an evaluation for QCA was carried out by AQA using CMI+ e-marking, in January 2005, for GCSE French Specification B listening tests (Fowles, 2005). Scripts were segmented and the questions

which required human marking (i.e. expert or general marking) were double marked. The findings revealed a high level of agreement (98.4%) between examiners in many of the responses. The question papers used in this study, however, tend to involve short response questions, meaning high levels of agreement between examiners is perhaps not unexpected, Ucles, (2002). This is similar to the GCE Chemistry paper marked on screen (Fowles, 2002), and those now e-marked in practice. According to Fowles (2006b) electronic marking with CMI+ raises the issue of differences, between whole paper and segmented marking, if any,

Some of the GCSE English examiners suggested that they normally keep an awareness, as they are marking, of the candidate's performance on the whole paper and they become 'familiar' with a candidate, including their handwriting, in Section A, which, it was claimed, helps them in reading and marking the essay in Section B. Ucles, (2002). The above advantages, they argued are not there in segmentation marking with CMI+ and therefore they claimed that segmented marking would not be more accurate than the whole paper marking. "*However AQA continues to prefer segmented e-marking (within CMI+) to whole paper e-marking because it has greater potential to monitor examiners' marking on*

individual questions and to direct them to specialist questions, as well as being considered more objective and reliable" Ucles, (2002).

Sequel to the above report Raikes according to Ucles (2002) concluded "*that on-screen marking is likely to be as reliable as paperbased marking, though it was suggested that segmentation would require further investigation, which still appears to be the case".* Ofqual (2014) says *"We only identified a very small number of relevant research papers, which suggests this is a topic yet to be fully explored".*

With respect to the e – marking version of segmented marking Wheadon and Pinot de Moira (2012) found that this method of marking, relative to whole script marking seems to improve the reliability of marking particularly for the high achieving students. Similarly, Black and Curcin (in preparation) reported evidence of halo effect in the whole script marking method which was not present in item level marking.

Summary of Literature Review

The researcher reviewed in this chapter the concepts of reliability. It was shown that the correlation coefficient is inadequate for estimating marking reliability. The researcher also showed that the Classical test theory on

marking reliability called the standard error of measurement is preferred by experts in measurement literature. The literature reviewed showed clearly that there is inherent unreliability associated with essay marking. The degree of this unreliability varies across subjects and assessment formats. This study also covered the various factors that affect marking reliability. Thereafter the researcher reviewed the various methods suggested in literature for reducing marking reliability. Although unreliability in essay marking may be reduced through training, strict adherence to marking schemes, marks or grades awarded to candidates will not be perfectly reliable according to experts.

In view of this problem some experts suggested that we should report the level of reliability associated with marks/grades, or seek alternatives to marking. In respect of these Satterly (1994) argues that the reliability of scores and grades in many external examinations will remain unknown to users and candidates since publishing low reliabilities and large measurement error associated with marks or grades would create poor public image to awarding bodies. In fact it does not seem feasible for an awarding body to take the sole responsibility of reporting reliability estimates or "that any individual awarding body would be willing to

accept the burden of educating test users in the meanings of those reliability estimates" (Meadows & Billington, 2005)

Another suggestion for resolving the problem of marking unreliability of essay was the use of computers. Ridgway and McCusker (2004) responding to computer marking as an alternative to human marking says the consequential validity of such marking systems would be "*dire*" (p.23). The advice to candidates would be to improve their scores simply by using more keystrokes. Fowles (2005) points out that automated marking is perfectly reliable in the sense that it will produce the same set of marks on a second occasion of marking. Nonetheless, a second set of marks might differ if a second examiner were to provide the marking rules.

It is unlikely that marking solely by computer will be acceptable in the foreseeable future. It has been suggested (Lamprianou, 2004, for example) that a pragmatic and effective way of improving marking reliability might be to have each script marked by a human marker and by software. In the case of a marking discrepancy, a second human marker would be called in for a second blind marking.

In any case marking by humans, it seems would always be the mode or at least part of the assessment system in many years to come. Therefore improving the quality of human marking remains a feasible option. In this regard research points to double marking as a credible strategy whose, only constraint is the high cost implication.

But one of the silent points about improving marking reliability not yet dismissed in literature is the part or segmented marking versus whole script marking. There is no significant experimental evidence of the effectiveness of segmented scoring method in literature. So it was suggested that this form of marking requires further investigation (Raikes,2002; Ucles, 2002; Bakker & Van Lent,2003; Meadows & Billington, 2005).

CHAPTER THREE

METHOD

This Chapter describes the method used in the study. The description is as follows:

Research Design

The post test - only control design was used in this study. This design is the use of a single test to compare the effect of different treatments (different levels of the independent variable/s) on particular groups whose members are selected randomly from the same population. The effects of the treatments are compared with one another so that each group is a control group to each other. (Garson, 2010)

The illustration is as follows:

R	W	--	X_1	Q_2
R	S	--	X_2	Q_2

R ---- Randomisation

W --- Whole Script Marking Group

S ---- Segmentation Marking Group

X_1 ---- Treatment (Manipulation) : Whole Script Marking Method

X_2 ---- Treatment (Manipulation) : Segmentation Marking Method

Q_2 --- Post - Test

This design is chosen because of the one shot test. Using a single test for this study instead of pretest and posttest excludes the problems of memory

effect, the effect of new learning and serious constraint that might arise, in the event of discontinuity of some participants after the first marking session. Thus a single test is likely to yield better validity than a double test, so it is more suitable for this study considering the high level of objectivity that is sought in this research.

Area of the Study

This study was carried out in Benin City. Benin City is the administrative, educational and commercial center of Edo State of Nigeria. The marking of all external examinations in the state are co – ordinated in the city. The city has more than two hundred private and public primary and secondary schools, with one public University, two private universities, one college of education, one institute of continuous education and several private post secondary educational training centers.

Population of the Study

The population of study was one hundred and seventeen (117) graduateteachers of Economics, who were shortlisted NECO Examiners

for 2012 NECO examinations in Edo State. Their marking experiences ranged between 0 – 12 years. The population consists of 42 inexperienced and 75 experienced markers.

Sample and Sampling Technique

A total of 48 NECO examiners constituted the sample for the study made up of 16 less – experienced and 32 more experienced markers. Since marking experience was considered a factor in this study, two groups of markers matched in terms of marking experience were obtained. Thus equal numbers of markers for the two groups were obtained through stratified random sampling from each category of markers of the same experience. The distribution of markers according to marking experience is shown in table 1

Table 1*: Distribution of Members in two Groups According to Years of Marking for NECO*

Number of Years of marking for NECO	Segmentation Group (SMM)	Whole Script Marking Group (WSMM)	Total
0	2	2	4
3	1	1	2
4	3	3	6
5	1	1	2
6	1	1	2
7	1	1	2
8	2	2	4
9	1	1	2
10	2	2	4
11	5	5	10
12	5	5	10
Total	**24**	**24**	**48**

Table 1 shows the distribution of the markers in respect of years of experience. Markers of the different marking experiences were equally distributed in the two groups.

The researcher used the scripts of 32 SSIII students of Kiddies College, Ikpoba slope, Benin City on who were administered by the researcher, economics examination for the purpose of this study, using the 1998 SSCE Economics theory examination paper.

Instrument

The instrument for the study was the marking scheme for ten essay test items in the 1998 SSCE Economics 2 Theory Questions. The instrument was adapted from Anyaele (2009) Senior School Certificate Past Questions and Answers; and marks were allocated to each unit by the researcher who is a graduate of Education/Economics.

Validation of Instrument

The marking scheme was screened and discussed by the 48 markers, for corrections and necessary adjustments during the coordination meeting, prior to the marking session. Collective agreement was the yardstick for validity.

Method of Data Collection

The data for the study were obtained through conference marking in which the two groups, of 24 markers each, marked the same 32 scripts. The 32 marks awarded by the 48 markers constituted the data for the study. The marks were used to compare the two groups.

The markers were invited for one day conference marking through correspondence by mail and telephone. Two centers namely Benson Idahosa University and Our Lady of Fatima at Auchi were used for the exercise. Upon arrival they were given attendance sheet to fill, each of them was given a parcel containing 32 clean photocopies of the students' answer scripts in the essay test, mark sheet and the question paper. Each parcel had one of the two methods and specified years of experience labeled on it. These parcels were distributed to each examiner according to their years of experience in line with table 2. The two groups were not separated, they all sat together.

Before the commencement of marking, the researcher spelt out the financial terms, explained the two methods and the purpose of the research. The marking guide was then discussed, item by item. Corrections that were accepted by all were effected in the guide and the corrected version was used for the marking. The marking exercise was co

– ordinated by the researcher and two assistants. Furthermore all the markers were instructed to write their names and the time it took each of them to finish marking the 32 scripts, on the parcel that was used for packaging of the scripts. In respect of students' handwriting, the researcher gave the 32 scripts to 20 research assistants from Benson Idahosa University to rate them according to the clarity of their handwriting.

Method of Data Analysis

The researcher used excel software and SPSS package to estimate the relevant statistics and charts to provide answers to the research questions and test the hypotheses. The standard error of measurement (SEM) was computed for each marker across the two groups using the formula for the standard error of measurement (see appendix i)

To compute the average mark changes (AMC) for each script the researcher summed the absolute difference between the mean and each mark and divided by 24 (number of marks for each script).

The mean rating of examinees' hand writing (MREH) was computed for the 32 scripts.

To compare the mean standard errors of measurement (SEM) in the scores awarded by the two groups as sought in research question 1, the mean

SEM for the two groups were computed. The maximum SEM, minimum SEM, SEM range, the mean SEM difference and percentage mean SEM difference were also computed for the two groups.

To compare the average mark changes (AMC) in the scores awarded by the two groups in line with research question 2, the mean AMC, for the two groups were computed. The maximum AMC, minimum AMC, AMC range, for the two groups, the mean AMC difference and percentage mean AMC difference were also computed for the two groups.

To address research questions 3 and 4 which sought to find the effect of examinees' handwriting on AMC_s of the two groups of markers, the mean ratings of examinees' handwriting for the 32 scripts were correlated with their average mark changes, using the Pearson correlation statistics. The linear relationship between AMC and mean rating of examinees' hand writing was also computed.

To solve research question 6 and 7 which addressed the effect of marker's experience on the SEM in the two groups of markers, the researcher correlated the individual markers' SEMs with their years of marking for NECO in the two groups, using the Pearson product moment correlation

statistics. The linear relationship between SEM_S and markers' experience in the two groups were also computed.

The average marking times used by the two groups sought in question 9 was computed by calculating the arithmetic mean of the total hours spent by each group.

For hypotheses 1 and 2, the t- test for independent samples was used to test the significant difference between the two groups in their mean SEM and in their mean AMC respectively. To test hypotheses 3and 4, SPSS Pearson correlation with flagged significance (p – level) was used.

The group that has the lower co-efficient of correlation between students' handwriting and their average mark changes is less sensitive to handwriting

To test hypotheses 6 and 7 the SPSS Pearson correlation with flagged significance (p – level) was used. To address hypothesis 8, the Fisher's z- transformation statistics was used to test for significant difference between the correlation coefficients of the two groups at alpha level of 0.05. The group that has the significantly lower co-efficient of correlation between marking experiences and their SEM_s is less sensitive to marker's

experience. The t – test for independent samples was used to test hypothesis 9, for significant difference between the means of the marking times used by the two groups at .05 alpha level.

CHAPTER FOUR

PRESENTATION AND ANALYSIS OF DATA

The data obtained from the field by the researcher are presented and analyzed in relation to the research questions and hypotheses in this chapter as follows:

Research Question 1:

The research question sought to ascertain the mean SEMs in scores awarded by the segmentation and whole script marking groups. The result is presented in table 2.

Table 2 - Standard Errors of Measurements (SEM) in Scores Awarded by the Segmentation (SMM) and Whole Script Marking Groups. (WSMM)

	N	Range	Min	Max	Sum	Mean	Std. Dev	Mean Diff	% Diff
SMM	24	11.50	6.00	17.50	274.60	11.44	2.64		
								0.00	0.00
WSMM	24	9.60	7.70	17.30	274.50	11.44	3.01		

Table 2 shows the major levels of SEM in the scores awarded by SMM and WSMM groups. The mean SEM for the segmentation marking group is **11.44** while the mean for the whole script marking group is **11.44.** The mean SEM difference between the two groups is **0.00** representing **0.00** percent.

Research Question 2:

The concern of the research question was to establish the mean average mark changes in scores awarded by the segmentation and whole script marking groups. The result is presented in the table 3.

Table 3 – Mean Average Mark Changes (AMC) in Scores Awarded by the Segmentation and Whole Script Marking Groups.

	N	Range	Min	Max	Sum	Mean	Std. Dev	Mean Diff	% Diff
SMM	32	6.20	5.64	11.84	277.44	8.670	1.3036		
								0.362	4.36
WSMM	32	4.16	6.38	10.54	265.87	8.308	1.1738		

Table 3 shows the major levels of AMC in the scores awarded by SMM and WSMM marking groups. The mean AMC for the segmentation marking group was **8.70** while the mean for the whole script marking group method was **8.30**. The mean AMC difference between the two groups is **0.40** which is 4.36 percent.

Research Question 3:

This question was aimed at determining the nature of the relationship existing between students' handwriting and the AMC in the scores awarded by the segmentation marking group. The study reveals that the relationship between the mean rating of examinees' handwriting and the AMC in the scores awarded by the segmentation marking group has

low negative correlation coefficients. The summary of the computation of the Pearson correlation is shown in table 4.

Table 4: Pearson Correlation between the Average Mark Changes and the Mean Rating of Examinees' Handwriting in the Scores Awarded by the Segmentation Marking Group.

No	Sum of Squares and Cross-products	Covariance	R
32	-40.063	-1.292	-0. 185

The computation of the coefficient of relationship between the average mark changes in 32 scripts and the mean rating of examinees' handwriting as shown in table 4 reveals a low negative correlation coefficient of **-0.185.** A graphic illustration of the linear relationship is shown in figure 3.

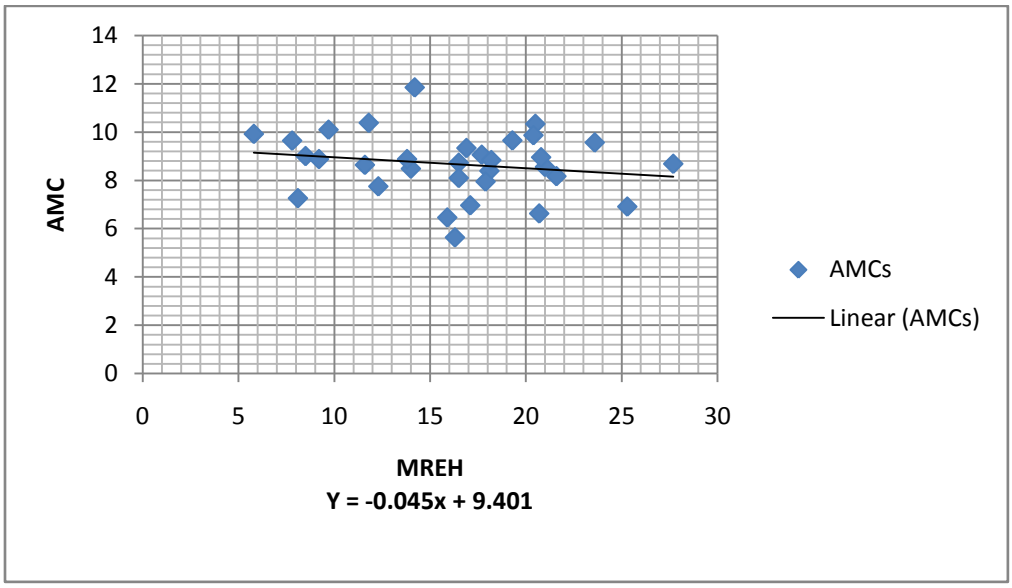

Figure 3: A Graphic Illustration of the Linear Relationship between AMC and MREH in the Segmented marking group

The regression equation is: **y = -0.045x + 9.401** Y and X represent the mean rating of examinees' handwriting and the average mark change. The lowest AMC is 5.6 with an MREH of 16.3 while the highest AMC is 11.8 with an MREH of 14.2. (See distribution in appendix vi serial numbers 21 and 26 respectively of page 169)

Research Question 4.

The research question on the relationship between students' handwriting and the average mark changes in the scores awarded by the whole script marking group was aimed to establish the nature of interaction between the two variables. The study shows that the relationship between the mean rating of examinees' handwriting and the AMC in the scores awarded by the whole script marking group has a low negative correlation as shown in table 5.

Table 5: Pearson Correlation between AMC and MREH in the Scripts Marked by the Whole Script Marking Group.

No	Sum of Squares and Cross-products	Covariance	r
32	-0.432	-0.014	-0. 002

The coefficient of relationship between the average mark changes in 32 scripts and the mean rating of examinees' handwriting is **-0.002,** a low negative coefficient of relationship as table 5 shows. A graphic representation of the linear relationship between AMC and MREH is shown in figure 4.

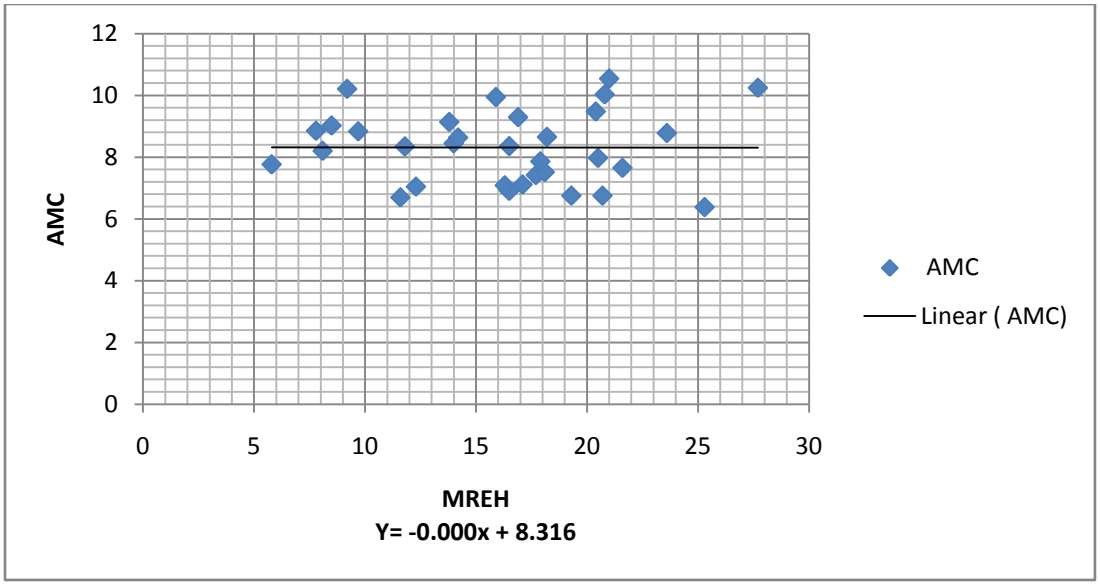

Figure 4: A Graphic Illustration of the Linear Relationship between AMC and MREH in Whole Script Marking Group

The coefficient of the linear relationship is depicted by the slight gradient of the slop of the regression line. The equation of the line is: **y = -0.000x + 8.316.** Y and X represent the average mark changes and the mean rating of examinees' handwriting respectively. The lowest AMC is 6.3 with an MREH of 25.3 while the highest AMC is 10.5 with an MREH of 21. (See

distribution in appendix vi serial numbers 7 and 25 respectively of page 169)

Research Question 5:

The research question 6 intended to ascertain the relationship between marking experience and the standard errors of measurement in the scores awarded by the segmentation marking group. The study shows that the relationship between marking experience (MEXP) and the SEM in the scores awarded by the segmentation marking group has a low negative correlation. This is shown in table 6.

Table 6 – Pearson Product Moment Correlation between Standard Errors of Measurement (SEM) and the Marking Experience (MEXP) in the Segmentation Marking Group.

No	Sum of Squares and Cross-products	Covariance	R
24	-46.542	-2.024	-0.197

Table 6 shows that the coefficient of relationship between SEM in 32 scripts and the marking experienced of the SMM marking group is **-0.197**. A graphic representation of the linear relationship between the two

variables is shown in figure 5.

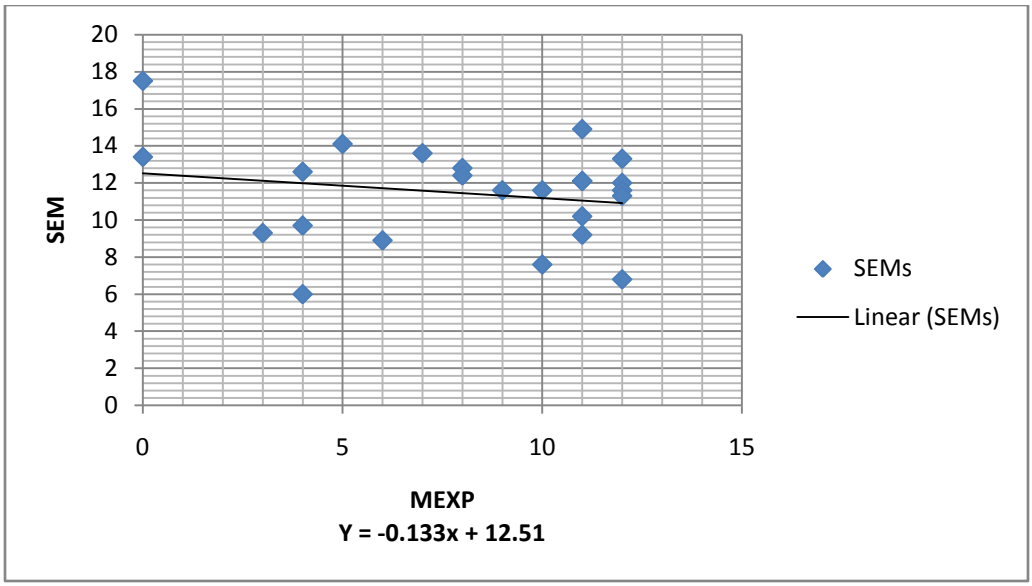

Figure 5: Showing the Linear Relationship between SEM and MEXP in Segmented Marking Group

Figure 5 shows that the slope of the line, the coefficient of the linear relationship between SEM and marking experience is **-0.133.** The regression equation is: **y = -0.133x + 12.51.** Y and X represent standard errors of measurement and marking experience respectively. The lowest SEM is 6.0 with an MEXP of 4 while the highest SEM is 17.5 with an MEXP of 0. (See distribution in appendix iv, serial numbers 14 and 20 respectively of page 167)

Research Question 6:

The research question was aimed at finding out the relationship between marking experience and the SEM in the scores awarded by the whole

script marking group. The study reveals that the relationship between marking experience and the standard errors of measurement in the scores awarded by the whole script group has a low negative correlation. This is shown in table 7.

Table 7 –Summary of the Pearson Product Moment Correlation between the Marking Experience (MEXP) of the WSMM Marking Group and Standard Errors of Measurement (SEM) in the Scores they Awarded.

No	Sum of Squares and Cross-products	Covariance	R
24	4.963	0.216	-0.018

Table 10 shows that the coefficient of relationship between SEM and the marking experienced of the WSMM group is **-0.03**. The linear relationship between the two variables is shown in figure 6.

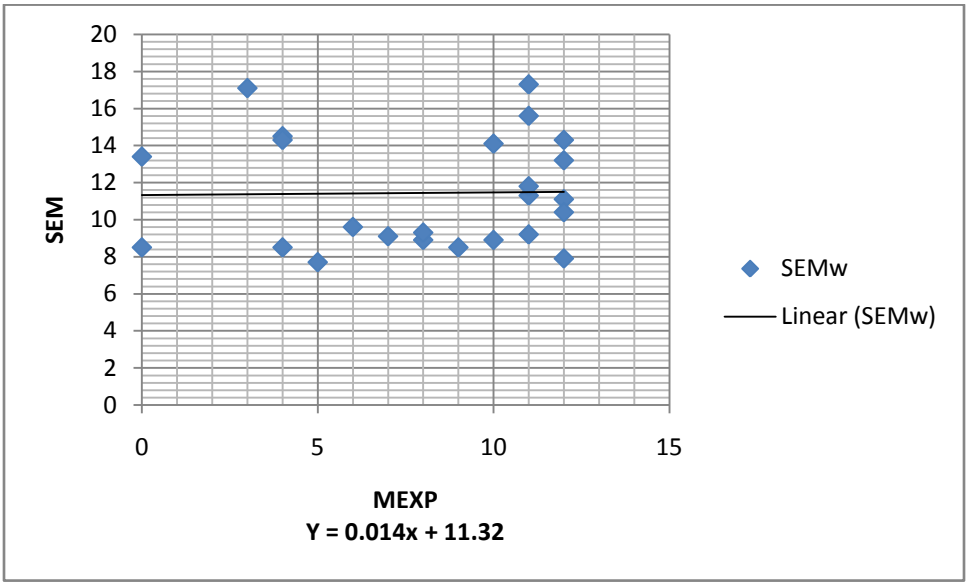

Figure 6: Showing the Linear Relationship between SEM and MEXP in the Whole Script Marking Method

The slope of the equation line in figure 4 representing **0.014** is the coefficient of the linear relationship between SEM and MEXP in the WSMM marking group. The regression equation is: **y = 0.014x + 11.32. Y** and x represent SEM and MEXP respectively. The lowest SEM is 7.7 with an MEXP of 5 while the highest SEM is 17.3 with an MEXP of 11. (See distribution in appendix iv, serial numbers 18 and 17 respectively of page 167)

Research Question 7:

The research question sought to determine the marking time efficiency of the SMM and WSMM marking groups. The average times used by the two groups are shown in table 8.

Table 8: Mean Marking Hours Used by the Segmentation and Whole Script Marking Groups

	N	Range	Min	Max	Sum	Mean	Std. Dev	Variance	Mean Diff
SMM	24	3.80	1.80	5.60	94.70	3.95	1.05	1.097	
									0.25
WSMM	24	4.00	1.80	5.80	88.90	3.70	1.12	1.264	

As indicated in table 8, the mean marking time used by segmentation and the whole script marking groups were **3.90** and **3.59** respectively with a mean difference of **0.31**.

Test of Hypotheses

The nine hypotheses formulated as operational guide to this study by the researcher were tested as follows:

Hypothesis 1

There is no significant difference between the mean standard errors of measurement in the scores awarded by the segmentation and whole script marking groups. The result of the test is presented in table 9.

Table 9 ***–Summary of the t- test Analysis of the Difference between the Means of the SEM in the Marks Awarded by the SMM and WSMM Groups on the same Scripts.***

Group	No of Examiners	Mean	Variance	df	t-cal	P-level (2 tailed)
SMM	24	11.442	6.940	46	0.005	0.996
WSMM	24	11.437	9.069			

The segmentation marking group has the similar level of SEM with the whole script marking group, with a mean difference of **0.005** as shown in table 9. This difference is not significant at 0.05 alpha level since the alpha level (**0.05)** is less than the p- level **(0.996).** The hypothesis is therefore not rejected.

Hypothesis 2

There is no significant difference between the mean average mark change in the scores awarded by the segmentation and whole script marking groups. The result of the test is presented in table 10.

Table10 –Summary of the t- test Analysis of the Difference between the Means of the Average Mark Change in the Marks Awarded by the SMM and WSMM Groups to the Same Scripts

Group	No of Examiners	Mean	Variance	df	t-cal	P-level (2 tail)
SMM	32	8.67	1.70	62	-1.17	0.25
WSMM	32	8.31	1.38			

Table 10 shows that the segmentation marking group has a higher AMC than the WSMM group with a mean difference of **0.36**. This difference is not significant at **0.05** alpha level since the alpha level (**0.05)** is less than the p- level (**0.25**). The hypothesis is therefore not rejected

Hypothesis 3

There is no significant relationship between students' handwriting and the average mark changes in the scores awarded by the segmentation marking group. This hypothesis was tested using the SPSS Pearson correlation test significance. The result is presented in table11.

Table 11 –Summary of the SPSS Test of Significance of Pearson Correlation Coefficient between Examinees' Handwriting and the AMC in the Marks Awarded by the *Segmentation Marking Group*.

No	Sum of Squares and Cross-products	Covariance	R	Sig. (2-tailed)
32	-40.063	-1.292	-0.185	0.310

The Pearson correlation coefficient (r) of the relationship between the clarity of examinees' handwriting and the average mark changes in the marks awarded by the segmentation marking group to the same scripts in table 11, is **-0.185.** r is not significant at **0.05** alpha level since p- level (**0.310**) is greater than the alpha level. Hypothesis 3 is therefore not rejected

Hypothesis 4

There is no significant relationship between students' handwriting and the average mark change in the scores awarded by the whole script marking group. The result of the SPSS test of significance of Pearson correlation is presented in table 12.

Table 12 –Summary of the SPSS Test of Significance of Pearson Correlation Coefficient between Examinees' Handwriting and AMC in the Marks Awarded by the WSMM

No	Sum of Squares and Cross-products	Covariance	R	Sig. (2-tailed)
32	-0.432	-0.014	-0.002	0.990

The Pearson correlation coefficient (r) of the relationship between the clarity of examinees' handwriting and the average mark change in the marks awarded by the whole script marking group is **--,0002.** r is not significant at **0.05** alpha level since p- level (**0.99)** is greater than the alpha level. Hypothesis 4 is therefore not rejected.

Hypothesis 5

There is no significant relationship between the marking experience of the segmentation marking group and the standard errors of measurement in the scores they awarded.

The result of the SPSS Pearson correlation test of significance is presented in table 13.

Table 13–Summary of the SPSS Test of Significance of Pearson Correlation Coefficient between the Marking Experience of the SMM group and the SEM.

No	Sum of Squares and Cross-products	Covariance	R	Sig. (2-tailed)
24	-46.542	-2.024	-0.197	0.356

Table 18 shows that the Pearson correlation coefficient (r) of the relationship between the standard errors of measurement in the marks awarded by the segmentation marking group and their marking experiences is **-0.21.** r is not significant at 0.05 alpha level since p- level (**0.66** is greater than the alpha level, thus hypothesis 6 is not rejected.

Hypothesis 6

There is no significant relationship between marking experience of the whole script marking group and the standard errors of measurement in the scores they awarded.

The result of the test is presented in table 14

Table 14 –Summary of the SPSS Pearson Correlation Test of Significance of the Coefficient between the Marking Experiences of the WSMM Group and the SEM

No	Sum of Squares and Cross-products	Covariance	R	Sig. (2-tailed)
24	4.963	0.216	0.018	0.932

The Pearson correlation coefficient of the relationship between the standard errors of measurement in the marks awarded by the whole script marking group and their marking experience is **-0.04.** r is not significant at 0.05 alpha level since p- level (**0.93)** is greater than the alpha level. Hypothesis 7 is therefore not rejected

Hypothesis 7.

There is no significant difference between the average times used by the segmentation and whole script marking groups.

The result of the test is shown in table 15.

Table 15 –Summary of the t- test analysis of the Difference between the Means of the Marking Durations of the SMM and WSMM Groups.

Group	No of Examiners	Mean	Variance	df	t-cal	P-level (2 tail)
SMM	24	3.95	1.10	46	0.77	0.4
WSMM	24	3.70	1.26			

Table 15 shows that the segmentation marking group has a higher average marking duration than the whole script marking group with a mean difference of **0.31**. This difference is significant at **0.05** alpha level, since p - level (**0.25)** is greater than the alpha level .Hypothesis 9 is therefore not rejected.

Summary of Major Findings

The major findings of the study were as follows:

1. SMM has similar mean SEM with WSMM, but has a higher mean AMC, higher mean marking duration but as marking experience increases, SEM reduces in SMM and increases in WSMM and as handwriting clarity increases, AMC becomes less in SMM than in WSMM.
2. There is no significant difference between the means of the SEM in the scores awarded by the SMM and the WSMM groups

3. There is no significant difference between the means of the AMC in the scores awarded by the SMM and the WSMM groups.
4. There is no significant relationship between students' handwriting and the AMC in the scores awarded by the SMM group.
5. There is no significant relationship between students' handwriting and the AMC in the scores awarded by the WSMM group.
6. There is no significant relationship between the marking experience of the SMM group and the SEM in the scores they awarded.
7. There is no significant relationship between the marking experience of the WSMM group and the SEM in the scores they awarded.
8. There is no significant difference between the average marking times used by the SMM and the WSMM groups.

CHAPTER FIVE

DISCUSSION OF RESULTS, CONCLUSION AND RECOMMEDATIONS

Discussion of Results

The hope that Segmented marking method (SMM) will yield higher marking reliability relative to the Whole Script marking method was tested in this study. The results of this study are discussed as follows:

Standard Error of Measurement in the Marks Awarded by the segmentation and whole script marking groups. One thing to keep in mind, in discussing the result of this study is the fact that the standard error of measurement (SEM) and the average mark change (AMC), though are conceptually similar, are not equal in size as SEM only covers the random portion of measurement error excluding the systematic errors while the AMC covers both components. This study shows that the segmentation marking group had similar SEM with the whole script marking group. The SEM for the segmentation and whole script marking groups were 11.442 and 11.438. The differential value was 0.004 which is not significant at 95 percent confidence level. Considering the fact that the segmentation marking group, apart from the making scheme, relies more

on the marks in previously marked scripts to assess examinee's performance than the whole script marking group, it is therefore, surprising that both of them have similar SEM.

The results of the study are largely similar with the results of previous studies. The two methods were shown in the work by Ucles (2002) to be similarly consistent in mathematics on screen marking. He also showed that, two examiners in English Literature were similarly consistent in the method of SMM. Ucles (2002) showed in his study that the examiners used in his study, generally were least consistent in the SMM method on screen marking in English Literature.

Fowles, (2005) reported the study carried out by the Assessment and Qualifications Alliance (AQA) in United States, in which markers who used the SMM method on CMI+ e- marking had 98.4% level of agreement, although this agreement was largely in short response questions.

Average Mark Change in the Marks Awarded by the Segmentation and Whole Script Marking Groups. The SMM marking group had a higher AMC, although the difference was not significant at 95 percent confidence level. The percentage difference was 4.36. This disparity

between the groups in the average mark changes (random and systematic errors) in the scores they awarded as compared to the similarity in their SEM (random errors) means that the segmentation marking group was more susceptible to systematic errors. This is not difficult to explain. Once a script has been over - rated or under - rated by the segmentation marking group, this error was exported to subsequent scripts because of their reliance on comparative scoring, whereas the marks awarded by the whole script marking group received systematic errors only from markers with habitual disposition for severity, moderation and leniency in marking. So the very fairness, upon which the comparative scoring of SMM is based, helps to spread any error from one script to another.

Though systematic errors is of little or no consequence among the scripts marked by one individual marker since the error is fairly distributed among the scripts, it is nevertheless a serious problem when comparing grades in national examination as different candidates receive different treatments - severity, leniency, and moderation from different markers.

Influence of Markers' Experience. The review of literature shows that marker's severity (systematic error) is more common with inexperienced markers (Cumming, 1990; Gordon & Kraemer, 1990; Huot, 1998; Ruth &Murphy,1988; Shohmy,1992; Weigle, 1994). The marks awarded by the

segmentation marking group as shown in this study were more negatively affected by markers' inexperience relative to the marks awarded by the whole script marking group, because markers generally are more used to the whole script than the segmented method. This has contributed to the higher systematic errors in the SSM group than the whole script marking group as already explained.

Similarly, the level of SEM in the segmentation marking group in this study is also the result of its higher level of susceptibility to marker's inexperience relative to the whole script marking group. This point is supported by studies which show that inexperienced markers are less consistent, that is more prone to random errors in marking than experienced markers (Pint de Moira, 2003a; Ruth &Murphy, 1988).

Generally the Pearson coefficient of the relationship between the SEM and markers' experience was small because there were markers who had few years of marking with NECO yet had been marking in internal examinations for some good number of years. The presence of such markers in the study therefore tended to equalize the experienced and inexperienced markers since the number of years of marking with NECO was the index of experience used in the study.

Another factor that probably reduced the effect of the number of years of marking with NECO, on SEM generally was senility. An examination of the scatter diagram relating makers' experience with SEM in figure 5 and 6 in chapter four shows that some markers with a good number of years of marking with NECO had higher SEM than the younger markers. This phenomenon was interpreted as senility factor by the researcher. This factor might have contributed in reducing the difference between the experienced and inexperienced markers in marking reliability generally in the study.

Influence of Examinees' Handwriting. The effect of examinees' handwriting on marking was not significant in the two groups. This finding did not agree with previous findings that handwriting significantly affects the reliability of essay marking (Chase, 1968; Markham, 1976; Briggs, 1980). This is most probably because the responses on the scripts were very scanty, which, by the estimation of the researcher, was about an average of two and a half pages, the highest being about four pages on A4 paper. The segmentation marking group was more susceptible to handwriting influence with Pearson coefficient of – 0.187 as against – 0.002 for the whole script marking group. This may be due to the fact that markers while marking whole script become used to the handwriting of

the examinee before they finish marking the script. Thus the influence of handwriting was less in the whole script marking method. The comparative approach in the SMM may also have contributed to the higher influence of handwriting in the SMM group.

Marking Duration. Many of the younger markers generally finished marking before the older ones with more years of marking experience. The segmentation marking group used a higher mean marking duration than whole script marking group. Furthermore, markers' behavior during the marking session was another pointer to the fact that they were not used to SMM method. They were not comfortable with the method as many of them openly complaint about the method as being laborious or demanding.

According to Ucles (2002), markers who used the SMM method complained that it was boring and less rewarding than marking whole scripts because according to them, marking a whole script made them to award a fair mark. Thus Ucles findings in English Literature are similar to the findings of this study both in respect of the results and the behavior of the segmentation marking group.

Conclusion

Since the two groups are equivalent, the difference between their marking reliability can only be attributed to the inherent qualities of the two methods they used. In view of this premise the researcher concluded that segmented marking method was more disposed particularly to higher systematic errors than the whole script method. The performance of the SMM was attributed mainly to two factors namely - the greater acquaintance with the WSMM by markers and the export of errors from one script to another elicited by the direct comparative technique of the method.

Although SMM was susceptible to markers' inexperience and poor examinees' handwriting, it has similar mean SEM (random errors) with WSMM; and as marking experience increases, SEM slightly increases in WSMM and reduces gradually in SMM, making the latter more reliable among experienced markers. Similarly as the clarity of examinees' handwriting increases, AMC is unchanged in WSMM and reduces gradually, becoming less in SMM than in WSMM.

Implication of the Study

In view of the above conclusion and the fact that training can immediately put the inexperienced markers on the same pedestal with the more experienced markers in marking reliability (Shohamy, Gordon & Kramer, 1992), it means that more intensive training in SMM will make it more reliable than WSMM irrespective of years of marking experience.

So the hope that SMM will yield higher marking reliability would be realized if markers are given intensive training in the use of the method. This could be boosted by warning examinees to write more eligibly.

Recommendations

In line with the implication of the study, the researcher suggested, that the various examination bodies should established training units and organize regular conferences and workshops for markers in the use of SMM with certificates of attendance given to participants at the end of such programmes. So that enlistment in subsequent marking of papers in external examinations will be based on attendance to such programmes. And also examination bodies should warn examinees to write more eligibly, with specified penalty for violators.

Suggestion for Further study

The researcher suggests that further research be carried out to assess the level of impact of the training of markers in the use of SMM method and implementation of handwriting rule on marking reliability.

REFERENCES

Akeju, S. A. (1972). The reliability of general certificate of education examination English composition papers in West Africa. *Journal of Educational Measurement, 9* (2), 175-179.

Anyaele, J. U. (2010). *Past questions and answers for SSCE, GCE and NECO economics theory/objectives 1988 – 2009,* Lagos: A. Johnson Publishers.

Bakker, S. & Van Lent, L. G. (2003). *National testing on - line: How far can we go?* Paper presented at the IAEA Conference, Manchester. Retrieved 5 January 2005 from http://www.aqa.org.uk/support/iaea/papers.html.

Berkowitz, D., Wolkowitz, B., Fitch, R. & Kopriva, R. (2000). *The use of tests as part of high-stakes decision-making for students: A resource guide for educators and policy makers.* Washington, DC: US Department for Education.

Black, B. & Curcin, M. (in press). *Marking item by item versus whole script – What difference does it make?* Cambridge Assessment Internal Report.

Branthwaite, A., Trueman, M. & Berrisford, T. (1981). Unreliability of marking: Further evidence and a possible explanation. *Educational Review, 33* (1), 41-46.

Briggs, D. (1970). The influence of handwriting on assessment. *Educational Research, 13,* 50-55.

Briggs, D. (1980). A study of the influence of handwriting upon grades using examination scripts. *Educational Review, 32* (2), 185-193

Brooks, V. (1980). Improving *the reliability of essay marking: a survey of the literature with particular reference to the English language composition*. (CSE Research Project Report 5) Leicester: Leicester University.

Bull, R. & Stevens, J. (1979). The effects of attractiveness of writer and penmanship on essay grades. *Journal of Occupational Psychology*, *52* ,53-59.

Byrne, C. (1979). Tutor-marked assignments at the Open University: A question of reliability. *Teaching at a Distance*, *15* , 34-43.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing. *Research in the Teaching of English*, *18*, 65-81.

Cialdini, R. B., Darby, B. L., & Vincent, J.E. (1973). Transgression and altruism: A case for hedonism. *Journal of Experimental Social Psychology*, *9*, 502-516.

Clark-Carter, D. (1997). *Doing quantitative psychological research: From design to report.* Hove: Psychology Press Ltd.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement*. Washington DC: American Council on Education.

Cohen, Y., Ben-Simon, A. & Hovav, M. (2003). *The effect of specific language features on the complexity of systems for automated essay scoring*. Paper presented at the 29th Annual Conference of the International Association for Educational Assessment, Manchester, October 2003.

Congdon, P. J. & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*, (2), 163-178.

Cooper, P.L. (1984). *The assessment of writing ability: A review of research*. (GRE Board Research Report No. 82-15R, ETS RR-84-12) Princeton, NJ: Educational Testing Service.

Cowie, J. & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, *39*, (1), 80-91.

Cresswell, M. J. (1985). *A review of borderline reviewing*. AEB Research Report, RAC/353.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

Cronbach, L. J., & Shavelson, R.J. (2004). *My current thoughts on coefficient alpha and successors. procedures*. (CSE Report 643) Los Angeles, CA: Centre for the study of evaluation (CSE).

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*, 31-51.

Daly, J. A. & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement*, *19*, (4), 309-315.

Delap, M. R. (1993a). *Marking reliability study in business sudies (665)* AEB Research Report RAC/609.

Duene Deardorff (2000). *Introduction to measurement and error analysis,* University of North Calrolina at Chapel Hill Dept. Of Physics and Astronomy

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

Engvik, H., Kvale, S. & Havik, O.E. (1970). Rater reliability in evaluation of essay and oral examinations. *Scandinavian Journal of Educational Research*, *14*, 195-220.

Fan, X. & Yin, P. (2003). Examinee characteristics and score reliability: An empirical investigation. *Educational and Psychological Measurement*, *63*, (3), 357-368.

Federal Republic of Nigeria, (2004). *National policy on education*, Abuja: Federal Ministry of Education.

Fowles, D. (2002). *Evaluation of an e-marking pilot in GCE Chemistry: Effects on marking and examiners' views.* AQA Research Report, RC/190.

Fowles, D. (2005). *Literature review on effects on assessment of e-marking.* AQA Research Report.

Garson D. T. (2010). *Research design*: Statnotes from North Carolina State University Public Administration Programme.

Greatorex, J. & Bell, J.F. (2002a). *Does the gender of examiners influence their marking?* Paper presented at the Learning communities and

assessment cultures: Connecting research with practice, University of Northumbria.

Greatorex, J. & Bell, J.F. (2002b). *What makes a senior examiner?* Paper presented at the British Educational Research Association, University of Exeter.

Hales, L. W., and Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement*, *12*, 115-117.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing.* Cambridge: Cambridge University Press.

Higher Education Quality Council (1997). *Assessment in higher education and the role of 'Graduateness'*. London: H.E.Q.C., Graduate Standards Programme.

Hill, B. J. (1975). Reliability of marking in BSc examinations in engineering. *International Journal for Mechanical Engineering Education*, *32*, 97-106.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hughes, D. C., Keeling, B., & Tuck, B.F. (1980a). Essay marking and the context problem. *Educational Research*, *22*, (2), 147-148.

Hughes, D. C., Keeling, B., & Tuck, B.F. (1983). The effects of instructions to scorers intended to reduce context effects in essay scoring. *Educational and Psychological Measurement*, *43*, 1047 - 1049

Hughes, D. C. & Keeling., B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement*, *21*, 277-281.

Humphris, G. M. & Kaney, S. (2001). Examiner fatigue in communication skills OSCEs. *Medical Education, 35,* 444-449.

Huot, B. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition & Communication*, *41*, 201-213.

James, C. (1974). The consistency of marking a physics examination. *Physics Education*, (9), 271-274.

Kaczmarek, C. (1980). Scoring and rating essay tasks. In O. A. Perkins (Ed.), *Research in language testing*. Rowley, MA: Newbury House.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgments. *Quarterly Journal of Experimental Psychology*, *42*, 239-254.

Lamprianou, J. (2004). *Marking quality assurance procedures: identifying good practice internationally*. Report prepared for the National Assessment Agency.

Landauer, T. K., & Dumais, S.T. (1997). A solution to Plato's problem: The ltent semantic analysis theory of the acquisition, induction and representation of knowledge. *Psychological Review, 104*, 211-240.

Lave, J. & Wenger, E. (1991). *Situated learning legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lucas, A. M. (1971). Multiple marking of a matriculation biology essay question. *British Journal of Educational Psychology*, *41*, (1), 78-84.

Lumley, T. L., Lynch, B.K. & McNamara, T.F. (1994). A new approach to standard setting in language assessment. *Melbourne Papers in Language Testing*, *3*, (2), 19-40.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professionals*, *13*, (4), 425-444.

Lunz, M. E., Stahl, J. A., & Wright, B.D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement*, *54*, 13-925.

Markham, L. R. (1976). Influence of handwriting quality on teacher evaluation of written work. *American Educational Research Journal, 13,* 277-283.

Mamta, A (2004). Curriculum reform in Schools: the importance of evaluation. *Journal in Curriculum Studies, 36,* (3), 361- 376

McColly, W. (1970). What does educational research say about the judging of writing ability? *Journal of Educational Research, 64,* 147-156.

McVey, P. J. (1976). The 'paper error' of two examinations in electronic engineering. *Physics Education, 11,* (1), 58-60.

Meadows, M & Billington, L (2005). *A review of the literature on marking reliability.* Retrieved from W:\Michelle\Michelle Meadows\Quality of Marking\Lit Review\Review Sections\A Review of the Literature on Marking Reliability.doc

Michael, W. B., Cooper, T., Shaffer, P. & Wallis, E. (1980). A comparison of the reliability and validity of ratings of student performance on essay examinations by professors of English and professors of other disciplines. *Educational and Psychological Measurement, 40,* 183-195.

Morrissy, M. (2000). *Do examiners go off? - Accuracy of examiners' marking.* AQA Research Report, RC76.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation, 7,* (3), 1-7.

Murphy, R. J. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology, 48,* (2), 196-200.

Murphy, R. J. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology, 52,* (1), 58-63.

Myers, M. (1980). *A procedure for writing assessment and holistic scoring*. Urbana, IL: National Council of Teachers of English, and Educational Resources Information Centre.

Myford, C.M., & Mislevy R. J. (1994). *Monitoring and improving a portfolio assessment system*. Princeton, NJ: Educational Testing Service.

Newstead, S. E. & Dennis, I. (1994). Examiners examined: The reliability of exam marking in psychology. *The Psychologist: Bulletin of the British Psychological Society*, *7*, 216-219.

Newton, P. (1996). The reliability of marking general certificate of secondary education scripts: Mathematics and English. *British Journal of Educational Research*, *22*, (4), 405 - 420.

Ofqual (2014). *Quality of marking, review of literature on item – level marking research*. Retrieved from www.ofqual.gov.uk

Okpala, P. N., Onocha, C. O., & Oyedeji, O. A. (1993). *Measurement and evaluation in education.* Jattu – Uzairue: Stirling – Horden Publishers

Pal, S. K. (1986). Examiners' efficiency and the personality correlates. *Indian Educational Review*, *21*, (1), 158-163.

Park, T.(n.d.). *Scoring procedures for assessing writing*. Retrieved from http://www.tc.columbia.edu/tesolalwebjournal/Park_Forum.pdf#search='holistic%20analytic%20scoring.

Partington, J. (1994). Double marking students' work. *Assessment & Evaluation in Higher Education*, *19*, (1), 57-60.

Pinot de Moira, A. (2003). *Examiner background and the effect on marking reliability*. AQA Research Report, RC218.

Pinot de Moira, A., Massey, C., Baird, J., & Morrissy, M. (2001). *Marking consistency over time.* AQA Research Report, RC/129.

Pollitt, A. (2004). *Let's stop marking exams*. Paper presented at the IAEA Conference, Philadelphia, June 2004.

Pollitt, A., & Crisp, V. (2004). *Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions?* Paper presented at the BERA Annual Conference, UMIST Manchester, September 2004.

Price, M. & Rust. C. (1999). The experience of introducing a common criteria assessment grid across an academic department. *Quality in Higher Education, 5,* 133-144.

Powers, D., & Kubota, M. (1998a). *Qualifying essay readers for an online scoring network (OSN)*. (RR-98-22) Princeton, NJ: Educational Testing Service.

Powers, D., & Kubota, M. (1998b). *Qualifying readers for the online scoring network: Scoring argument essays*. (RR-98-28) Princeton, NJ: Educational Testing Service.

Qualifications and Curriculum Authority (QCA) (2002). *Maintaining GCE A Level standards: The findings of an independent panel of experts*. London: QCA.

Raikes, N. (2002). *On screen marking of scanned paper scripts*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).

Ridgway, J. & McCusker, S. (2004). *Literature review of e-assessment*. (Report 10) NESTA Futurelab Series.

Royal-Dawson, L. (2004). *Is teaching experience a necessary condition for markers of key stage 3 English?* AQA Research Report, RC261.

Rudner, L. M., & Schafer, W.D. (2001). Reliability. *ERIC Digest*.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Review of Education*, *13*, (2), 191-209.

Satterly, D. (1994). Quality in external assessment. In W. Harlen (Ed.), *Enhancing quality in assessment,* London: Paul Chapman.

Shohamy, E., Gordon, C., & Kramer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76,* (10) ,27-33.

Smith, B., Sinclair, H., Simpson, J., van Teijlingen, E., Bond, C., & Tyalor, R. (2002). What is the role of double-marking? Evidence from an undergraduate medical course. *Education for Primary Care*, *1*, 497-503.

Sparks, R. & Ballantyne, R. (1997). Quality control methods in large-scale assessment procedures using 'double-marking' or 'partial double-marking'. *Quality Control & Applied Statistics*, *42*, (1), 45-48.

Spear, M. (1996). The influence of halo effects upon teachers' assessments of written work. *Research in Education*, *56*, 85-87.

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research, 39,* (2), 229-233.

Sturman, L. & Kispal, A. (2003). *To e or not to e? A comparison of electronic marking and paper-based marking*. Paper presented at the 29th International Association of Educational Assessment Conference, Manchester, UK, October 2003.

Sukkarieh, J. Z., Pulman, S.G. & Raikes, N. (2003). *Auto-marking: using computational linguistics to score short, free text responses.* Paper presented at the 9th Annual

Sygie (2010). *Standard error* (statistics). Retrieved from www.http://en.wikipedia.org/wiki/standard_error(statistics).

Townsend, M. A. R., Yong Kek, L.Y. & Tuck, B.F. (1989). The effect of mood on the reliability of essay assessment. *British Journal of Educational Psychology*, *59*, 232-240.

Twing, J. S. & Harrison, I. (2003). *The comparability of paper-based and imaged-based marking of a large-scale high-stakes writing assessment in the United States*. Paper presented at the 29th Annual IAEA conference, Manchester, UK.

Uebersax, J. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, *101*, (1), 140-146.

University of Cambridge Local Examinations Syndicate (2000). *Key stage 3 English - A study of marking reliability which investigates three different methods of maintaining consistency between markers*. Report produced for the Qualifications and Curriculum Authority.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. H.-. Lyons (Ed.), *Assessing second language writing in academic contexts.* Norwood, N.J.: Ablex Publishing Corporation.

Wegner, E. (1998). *Communities of practice learning, meaning and identity*. Cambridge: Cambridge University Press.

Weigle, S. (1994). *Effects of training on raters of ESL compositions: Quantitative and qualitative approaches*. Unpublished PhD dissertation, University of California, Los Angeles.

Weigle, S. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative & qualitative approaches. *Assessing Writing*, *6*, (2), 145-178.

Wheadon C. & Pinot de Moira (2012) Gains in marking reliability from item – level marking: Is the sum of the parts better than the whole? *Educational Research and Evaluation: An International Journal on Theory and Practice* 19, (8), 2013

Whetton, C. & Newton, P. (2002). *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment Conference, Hong Kong, September 2002.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. H.-. Lyons (Ed.), *Assessing second language writing in academic contexts*. Norwood, N.J.: Ablex Publishing Corporation.

Wiliam, D. (1993). Validity, dependability and reliability in national curriculum assessment. *The Curriculum Journal, 4*, (3),335-350.

Wiliam, D. (2000). Reliability, validity, and all that jazz. *Education*, *29,* (3), 9-13.

William M.K. Trochim (2006). *Research methods knowledge base, Second edition.* Retrieved from Internet www page at URL:http://www.socialresearchmethodsnet/kb/

Williams, H. G. & van Lent, L. G. (2002). *Project 2F.1: Impact of e-marking on test design*. Utrecht: ETS Europe.

Wolf, A. (1995). *Competence based assessment*. Buckingham: Open University Press.

Wood, R. & Quinn, B. (1976). Double impression marking of English language essay and summary questions. *Educational Review*, *28* (3), 229-246.

Zia Abbasi (2010). *Reducing measurement error in informal sector surveys*. Retrieved from www.Index Overview Help (1997) Error (C), Index Overview Help (1997) Error (C).

APPENDIX I

Table below shows the various computations that were made using the 32 scripts marked by 24 markers in each group. The actual values for those items on table 3are presented in appendices 2 and 3 for the two groups respectively.

Table 3 – ***Description of the Table Used for Computing and Organizing Data in the Study.***

Scripts	**24 Markers**	**Mean Mark**	**AMC**	**MREH**
Serial numbers in the 32 Scripts	30 Marks for each of the scripts.	Mean Mark for each of the scripts	AMCs for each of the scripts	MREHs for each of the scripts
P_r	P_r for each of the 24 Markers			
SD	Standard deviation for each of the 24 Markers			
S_E	SEM for each of the 24 Markers			
Marking Time	Marking Time for each of the 24 Markers			
Marker's Experience	Marker's Experience for each of the 24 Markers			

$$S_E = S_X\sqrt{1 - r_{xx}}$$

S_E = Standard error of measurement

S_X = Standard deviation of the test scores

r_{xx} = reliability of the test scores (Pearson correlation between the first and the second halves of the scores).

APPENDIX 1I

The Marks Awarded By Thirty Markers to the Same Scripts Using the SMM

Candidates S/N	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Marker 7	Marker 8	Marker 9
1	40	69	75	53	63	57	51	57	46
2	42	48	62	45	41	44	47	24	26
3	40	59	80	61	63	64	49	60	49
4	39	78	83	59	58	56	44	56	53
5	39	62	82	55	55	54	43	60	48
6	39	64	76	42	51	43	43	46	44
7	12	31	38	23	36	38	18	45	30
8	30	62	60	41	40	42	38	42	33
9	42	59	74	44	42	51	39	50	40
10	54	70	66	68	69	61	53	71	61
11	45	78	87	60	58	57	50	68	48
12	42	72	70	50	55	46	41	49	45
13	43	75	73	63	63	62	52	63	42
14	43	84	81	65	43	54	53	61	50
15	40	67	81	59	56	59	46	66	52
16	34	58	69	51	38	47	36	40	44
17	53	72	89	71	60	63	52	67	53
18	47	66	81	63	64	59	45	62	49
19	54	78	73	55	68	45	25	46	46
20	32	57	67	40	33	51	36	43	38
21	33	55	70	46	43	54	39	51	46
22	40	54	69	45	48	40	43	52	26
23	21	22	32	31	35	28	37	42	18
24	42	60	77	45	48	55	48	50	42
25	24	60	69	47	50	52	41	60	38
26	38	57	71	49	42	56	37	47	42
27	5	35	40	12	27	46	22	27	19
28	5	37	38	10	25	37	32	21	16
29	30	55	69	42	48	51	58	52	36
30	45	58	82	46	56	59	48	56	56
31	23	54	61	46	52	47	51	53	44
32	20	54	88	45	55	54	49	53	50
r	0.1075	0.2011	0.2399	0.0811	-	0.3609	0.0798	-0.232	0.1571
S	12.626	14.047	14.705	14.214	11.623	8.478	9.261	12.091	11.144
SEM	12	12.6	12.8	13.6	11.6	6.8	8.9	12.1	10.2
Time	5.1	4.5	2.8	4.8	4.2	4.9	4.5	3.8	5.6
EXP	12	4	8	7	12	12	6	11	11

Appendix 1I Contd

Candidates S/N	Marker 10	Marker 11	Marker 12	Marker 13	Marker 14	Marker 15	Marker 16	Marker 17	Marker 18
1	60	62	50	77	72	60	60	57	54
2	36	49	40	69	71	49	52	30	48
3	68	60	48	76	77	65	63	56	63
4	56	67	51	70	81	65	68	60	60
5	52	59	47	69	63	56	47	57	54
6	48	49	44	52	45	48	54	44	47
7	26	32	34	27	46	30	32	25	35
8	46	56	39	47	55	50	36	44	47
9	50	59	44	56	57	55	43	61	49
10	67	71	65	80	72	64	65	75	47
11	64	68	57	76	72	67	64	68	53
12	53	55	49	59	50	42	51	56	41
13	63	64	59	79	71	61	59	57	64
14	65	44	61	80	72	57	58	59	65
15	61	68	55	87	69	66	70	58	65
16	46	62	35	62	70	58	47	36	55
17	69	68	73	88	70	61	62	73	64
18	63	65	58	87	64	64	65	62	64
19	61	62	57	76	61	59	64	64	52
20	41	48	38	46	72	52	42	46	47
21	49	49	46	49	50	46	47	46	57
22	49	58	52	57	60	51	51	45	50
23	20	31	37	26	46	42	29	17	30
24	49	46	47	68	67	54	56	52	53
25	37	46	54	57	65	61	51	47	52
26	45	51	21	55	69	59	49	52	46
27	25	32	33	31	58	42	32	20	34
28	24	24	31	29	48	11	13	18	32
29	37	52	55	61	70	42	49	41	44
30	56	64	66	78	76	64	59	49	54
31	37	50	50	67	69	43	49	42	48
32	42	52	52	73	66	55	46	49	52
R	0.218	0.053	-0.148	0.427	0.631	0.329	0.171	0.047	0.365
S	13.691	11.96	11.333	17.702	9.943	11.871	12.754	14.913	9.539
SEM	12.1	11.6	11.3	13.4	6	9.7	11.6	13.3	7.6
Time	4.8	5.3	3.8	3	2	4.3	2.5	4.3	4.5
EXP	11	9	12	0	4	4	10	12	10

Appendix 1I Contd

Candidates S/N	Marker 19	Marker 20	Marker 21	Marker 22	Marker 23	Marker 24	Mean	AMC	MREH
1	56	62	52	26	85	53	58.21	8.7256	16.5
2	54	57	34	62	78	42	47.92	10.333	20.5
3	63	72	61	77	75	52	62.54	7.7465	12.3
4	71	68	62	70	95	58	63.67	9.6388	7.8
5	59	81	53	67	85	41	57.83	9.0555	17.7
6	49	64	47	60	62	51	50.5	6.625	20.7
7	39	40	31	52	28	42	32.92	6.9097	25.3
8	46	34	41	62	73	40	46	8.1666	21.6
9	50	35	45	72	73	51	51.71	8.3854	18.1
10	66	63	72	81	85	73	67.46	6.4583	15.9
11	74	65	66	77	95	63	65.83	8.8333	18.2
12	52	74	54	65	87	52	54.58	8.4791	14
13	66	67	63	76	90	56	63.79	7.2569	8.1
14	66	64	58	80	90	47	62.5	10.375	11.8
15	78	73	63	80	85	60	65.17	9	8.5
16	62	45	46	59	74	50	51	9.9166	5.8
17	73	87	63	90	100	62	70.13	9.5625	23.6
18	59	77	58	75	93	51	64.21	7.9444	17.9
19	59	88	59	72	60	63	60.29	8.875	9.2
20	59	68	45	57	56	47	48.37	8.6354	11.6
21	57	38	48	55	63	48	49.38	5.6354	16.3
22	59	68	48	22	98	51	51.5	9.3333	16.9
23	35	33	28	57	36	36	32.04	6.9618	17.1
24	70	58	49	69	69	51	55.21	8.0937	16.5
25	41	51	46	72	40	60	50.875	8.5520	21
26	64	77	38	82	85	47	53.29	11.84028	14.2
27	19	48	31	47	30	34	31.21	8.875	13.8
28	15	18	19	47	39	29	25.75	9.6458	19.3
29	48	47	37	62	77	47	50.42	8.9513	20.8
30	68	89	52	67	90	57	62.29	10.090	9.7
31	55	61	40	71	72	41	51.08	8.6805	27.7
32	50	73	40	72	74	49	54.71	9.8611	20.4
R	0.049	0.007	-0.123	0.1438	0.1544	-0.181			
S	14.441	17.647	12.43	14.933	20.268	9.342			
SEM	14.1	17.5	12.4	14.9	9.2	9.3			
Time	4	3.7	4.4	1.8	3.8	2.3			
EXP	5	0	8	11	11	3			

APPENDIX III

Marks Awarded to the Same Scripts by Thirty Markers Using the WSMM

Candidates S/N	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Marker 7	Marker 8	Marker 9	Marker 10
1	51	60	60	59	62	51	56	67	57	58
2	42	59	33	42	53	59	40	70	36	42
3	59	70	65	56	66	58	56	82	57	59
4	54	65	62	57	68	54	53	73	52	48
5	48	54	57	53	63	56	46	65	54	54
6	37	58	54	54	47	44	46	45	42	48
7	28	26	27	35	41	46	24	33	26	30
8	36	48	39	40	46	47	35	48	40	34
9	45	45	51	41	51	61	39	52	46	41
10	51	67	70	66	58	84	66	78	58	58
11	60	65	69	59	57	65	61	77	51	53
12	56	59	60	59	41	53	53	57	48	46
13	55	64	57	54	61	60	55	74	56	54
14	59	56	64	63	67	66	64	78	49	56
15	67	65	62	59	63	55	55	81	55	55
16	58	57	49	47	42	41	46	57	38	43
17	60	56	73	70	69	67	64	40	64	61
18	53	65	61	65	63	59	62	87	56	54
19	60	64	68	70	70	59	59	76	56	56
20	58	43	47	39	48	40	38	52	44	41
21	46	50	51	51	48	65	54	41	36	41
22	50	52	54	63	52	44	58	62	44	47
23	31	27	28	29	28	39	22	47	30	14
24	58	55	58	47	45	55	51	73	43	46
25	27	43	44	46	64	63	40	75	39	45
26	50	54	46	53	41	58	45	60	45	46
27	28	48	15	30	31	44	31	49	27	24
28	16	35	15	20	27	30	18	26	24	18
29	52	57	29	46	57	56	35	63	35	34
30	60	55	58	53	63	69	52	73	49	50
31	60	32	36	52	57	65	40	78	42	44
32	53	38	40	51	61	63	43	77	44	47
R	0.19	0.465	-0.061	0.11	0.456	0.218	0.119	0.2	0.199	0.402
S	12.313	11.678	15.633	11.979	12.013	10.898	12.612	15.799	10.218	11.57
SEM	11.1	8.5	15.6	11.3	8.9	9.6	11.8	14.1	9.1	8.9
Time	3.7	4.5	3.5	4.1	3.4	4	5.8	5.8	3.2	3.4
EXP	12	9	11	11	10	6	11	10	7	8

APPENDIX III CONT.

Candidates S/N	Marker 11	Marker 12	Marker 13	Marker 14	Marker 15	Marker 16	Marker 17	Marker 18	Marker 19	Marker 20
1	71	55	54	63	53	80	69	31	58	74
2	39	41	49	42	42	52	40	8	42	56
3	60	66	54	57	52	76	65	28	55	70
4	53	52	42	55	53	77	62	28	67	76
5	67	54	58	56	59	77	65	28	43	63
6	57	46	34	49	51	60	68	23	45	49
7	39	31	31	30	24	22	18	3	17	35
8	44	39	47	52	40	67	61	12	47	52
9	44	41	53	45	34	50	60	14	46	62
10	71	50	64	57	53	74	69	21	69	76
11	66	60	64	64	49	80	73	27	69	80
12	56	47	42	47	43	67	75	22	54	56
13	68	61	60	51	46	77	76	20	61	69
14	51	61	69	55	55	75	68	23	65	75
15	71	59	58	67	55	53	61	32	60	88
16	42	39	56	65	46	57	53	25	44	52
17	59	56	72	67	46	73	77	28	60	49
18	59	52	60	57	43	64	65	18	52	47
19	60	59	58	65	43	75	73	16	46	31
20	37	37	61	55	43	46	38	23	44	46
21	49	51	51	55	43	69	61	24	46	57
22	46	49	62	54	36	71	69	21	36	54
23	29	28	34	28	12	27	22	3	15	21
24	57	43	71	57	45	58	56	23	49	69
25	41	43	51	42	32	55	36	3	28	56
26	53	44	68	52	44	62	52	16	34	52
27	34	29	46	34	11	29	11	1	18	41
28	10	26	26	23	19	21	16	1	9	33
29	47	50	63	41	49	40	45	10	36	40
30	55	56	79	62	66	69	58	19	48	69
31	60	48	59	55	37	55	56	12	33	60
32	64	48	67	47	40	55	51	11	34	57
R	0.082	0.415	0.467	0.432	0.536	0.281	0.099	0.298	0.178	0.263
S	13.76	10.302	12.56	11.311	12.502	17.086	18.24	9.21	15.79	15.56
SEM	13.2	7.9	9.2	8.5	8.5	14.5	17.3	7.7	14.3	13.4
Time	4.1	5	4.2	1.8	5	1.8	1.8	3.1	3.4	2.9
EXP	12	12	11	0	4	4	11	5	4	0

APPENDIX III CONT.

Candidates S/N	Marker 21	Marker 22	Marker 23	Marker 24	Mean	AMC	MREH
1	55	57	75	60	59.83333	6.902778	16.5
2	37	35	46	40	43.54167	7.972222	20.5
3	53	55	66	59	60.16667	7.041667	12.3
4	56	60	76	66	58.70833	8.850694	7.8
5	45	52	69	46	55.5	7.416667	17.7
6	44	50	62	51	48.5	6.75	20.7
7	32	29	17	27	27.95833	6.378472	25.3
8	47	51	60	41	44.70833	7.649306	21.6
9	42	48	55	66	47.16667	7.513889	18.1
10	53	60	88	71	63.83333	9.944444	15.9
11	55	75	82	65	63.58333	8.652778	18.2
12	44	60	69	42	52.33333	8.444444	14
13	50	59	78	64	59.58333	8.201389	8.1
14	69	71	79	69	62.79167	8.34375	11.8
15	57	82	89	57	62.75	9.020833	8.5
16	38	67	52	42	48.16667	7.763889	5.8
17	58	66	73	72	61.66667	8.777778	23.6
18	55	66	68	53	57.66667	7.861111	17.9
19	46	76	75	58	59.125	10.20833	9.2
20	42	66	53	47	45.33333	6.694444	11.6
21	45	63	59	48	50.16667	7.083333	16.3
22	47	71	65	64	52.95833	9.291667	16.9
23	31	36	16	34	26.29167	7.111111	17.1
24	44	63	61	60	53.625	8.354167	16.5
25	25	48	50	55	43.79167	10.54167	21
26	40	67	69	59	50.41667	8.631944	14.2
27	24	28	21	39	28.875	9.135417	13.8
28	6	25	12	26	20.08333	6.75	19.3
29	28	63	47	53	44.83333	10.02778	20.8
30	55	72	71	67	59.5	8.833333	9.7
31	46	54	52	50	49.29167	10.24306	27.7
32	42	57	53	52	49.79167	9.475694	20.4
R	0.288	-0.046	0.275074	0.431			
S	12.35	14.33	20.06883	12.33			
SEM	10.4	14.3	17.1	9.3			
Time	2.8	3.2	3.3	5.1			
EXP	12	12	3	8			

APPENDIX IV

The Standard Error of Measurement (SEM) in Marks Awarded by 24 Markers Using SMM and WSM and MEXP

S/N	SMM	WSMM	$MEXP_{SMM}$	$MEXP_{WSMM}$
1	12	11.1	12	12
2	12.6	8.5	4	9
3	12.8	15.6	8	11
4	13.6	11.3	7	11
5	11.6	8.9	12	10
6	6.8	9.6	12	6
7	8.9	11.8	6	11
8	12.1	14.1	11	10
9	10.2	9.1	11	7
10	12.1	8.9	11	8
11	11.6	13.2	9	12
12	11.3	7.9	12	12
13	13.4	9.2	0	11
14	6	8.5	4	0
15	9.7	8.5	4	4
16	11.6	14.5	10	4
17	13.3	17.3	12	11
18	7.6	7.7	10	5
18	14.1	14.3	5	4
20	17.5	13.4	0	0
21	12.4	10.4	8	12
22	14.9	14.3	11	12
23	9.2	17.1	11	3
24	9.3	9.3	3	8

APPENDIX V

t-test for Significant Difference between the Means of the SEM in SMM.

and WSMM (Excel Output).

t-Test: Two-Sample Assuming Equal Variances

SEM	*SMM*	*WSMM*
Mean	11.4416666666667	11.4375
Variance	6.93992753623192	9.0685326086956
Observations	24	24
Pooled Variance	8.00423007246376	
Hypothesized Mean Difference	0	
Df	46	
t Stat	0.00510175500945763	
P(T<=t) one-tail	0.497975733769242	
t Critical one-tail	1.67866041403406	
P(T<=t) two-tail	0.995951467538483	
t Critical two-tail	2.0128955673215	

APPENDIX VI
Average Mark Changes (AMC) in Marks Awarded by 24 Markers to the Same Scripts Using SMM and WSMM and MREH

S/N	WSMM	SMM	MREH
1	6.902777778	8.725694	16.5
2	7.972222222	10.33333	20.5
3	7.041666667	7.746528	12.3
4	8.850694444	9.638889	7.8
5	7.416666667	9.055556	17.7
6	6.75	6.625	20.7
7	6.378472222	6.909722	25.3
8	7.649305556	8.166667	21.6
9	7.513888889	8.385417	18.1
10	9.944444444	6.458333	15.9
11	8.652777778	8.833333	18.2
12	8.444444444	8.479167	14
13	8.201388889	7.256944	8.1
14	8.34375	10.375	11.8
15	9.020833333	9	8.5
16	7.763888889	9.916667	5.8
17	8.777777778	9.5625	23.6
18	7.861111111	7.944444	17.9
19	10.20833333	8.875	9.2
20	6.694444444	8.635417	11.6
21	7.083333333	5.635417	16.3
22	9.291666667	9.333333	16.9
23	7.111111111	6.961806	17.1
24	8.354166667	8.09375	16.5
25	10.54166667	8.552083	21
26	8.631944444	11.84028	14.2
27	9.135416667	8.875	13.8
28	6.75	9.645833	19.3
29	10.02777778	8.951389	20.8
30	8.833333333	10.09028	9.7
31	10.24305556	8.680556	27.7
32	9.475694444	9.861111	20.4

APPENDIX VII

t-test of Significant Difference between the Means of the AMC in SMM and WSMM. (Excel Output)

t-Test: Two-Sample Assuming Unequal Variances

	WSMM -AMC	*SMM-AMC*
Mean	8.308376736	8.670138889
Variance	1.37781452	1.699525214
Observations	32	32
Hypothesized Mean Difference	0	
Df	61	
t Stat	-1.166568906	
P(T<=t) one-tail	0.123962511	
t Critical one-tail	1.670219484	
P(T<=t) two-tail	0.247925021	
t Critical two-tail	1.999623567	

APPENDIX VIII

Correlation between AMC and the Mean Rating of Examinees' Handwriting (MREH) in SMM and WSMM

Correlations

		WSMM	MREH
WSMM	Pearson Correlation	1	-.007
	Sig. (2-tailed)		.971
	Sum of Squares and Cross-products	42.285	-1.285
	Covariance	1.364	-.041
	N	32	32
MREH	Pearson Correlation	-.007	1
	Sig. (2-tailed)	.971	
	Sum of Squares and Cross-products	-1.285	888.695
	Covariance	-.041	28.668
	N	32	32

		SMM	MREH
SMM	Pearson Correlation	1	-.191
	Sig. (2-tailed)		.295
	Sum of Squares and Cross-products	46.220	-38.680
	Covariance	1.491	-1.248
	N	32	32
MREH	Pearson Correlation	-.191	1
	Sig. (2-tailed)	.295	
	Sum of Squares and Cross-products	-38.680	888.695
	Covariance	-1.248	28.668
	N	32	32

APPENDIX IX

Correlation between Markers Experience (MEXP) and SEM in SMM and WSMM

Correlations

SMM		SEM	MEXP
SEM	Pearson Correlation	1	-.197
	Sig. (2-tailed)		.356
	Sum of Squares and Cross-products	348.958	-46.542
	Covariance	15.172	-2.024
	N	24	24
MEXP	Pearson Correlation	-.197	1
	Sig. (2-tailed)	.356	
	Sum of Squares and Cross-products	-46.542	159.618
	Covariance	-2.024	6.940
	N	24	24

WSMM		MEXP	SEM
MEXP	Pearson Correlation	1	.018
	Sig. (2-tailed)		.932
	Sum of Squares and Cross-products	348.958	4.963
	Covariance	15.172	.216
	N	24	24
SEM	Pearson Correlation	.018	1
	Sig. (2-tailed)	.932	
	Sum of Squares and Cross-products	4.963	208.576
	Covariance	.216	9.069

APPENDIX X

Marking Durations of 30 Markers Using SMM and WSMM to Mark 32 Scripts

SMM		WSMM	
Markers S/N	Marking Duration	Markers	Marking Duration
1	5.1	1	3.7
2	4.5	2	4.5
3	2.8	3	3.5
4	4.8	4	4.1
5	4.2	5	3.4
6	4.9	6	4
7	4.5	7	5.8
8	3.8	8	5.8
9	5.6	9	3.2
10	4.8	10	3.4
11	5.3	11	4.1
12	3.8	12	5
13	3	13	4.2
14	2	14	1.8
15	4.3	15	5
16	2.5	16	1.8
17	4.3	17	1.8
18	4.5	18	3.1
19	4	19	3.4
20	3.7	20	2.9
21	4.4	21	2.8
22	1.8	22	3.2
23	3.8	23	3.3
24	2.3	24	3.7

APPENDIX XI

t- test Analysis of the Difference of the Mean Marking Durations between Markers Using SMM and WSMM (SPSS Output).

t-Test: Two-Sample Assuming Equal Variances

	SMM	*WSMM*
Mean	3.945833333	3.704166667
Variance	1.097373188	1.263894928
Observations	24	24
Pooled Variance	1.180634058	
Hypothesized Mean Difference	0	
df	46	
t Stat	0.770459335	
P(T<=t) one-tail	0.222483317	
t Critical one-tail	1.678660414	
P(T<=t) two-tail	0.444966634	
t Critical two-tail	2.012895567	