

CHAPTER ONE

INTRODUCTION

Background to the Study

The academic world has experienced several stages which span from the advent of writing on slates, scrolls, wooden surfaces, rocks or even cemented surfaces, all with the purpose of passing information or knowledge to others and to make teaching and learning possible. Learning can take place through observation, training and teaching (instructions). The need to measure what has been learnt has led to giving of scores. In order to score learners, education have developed media such as interview, paper-pencil and computer methods, tools and devised techniques such as observation, questionnaire, checklist, interview, rating scale, anecdotal records and tests to make scoring possible.

Among alternative measurement techniques, test is considerably the most commonly used at different levels of education. These tools and techniques elicit responses from learners, users or respondents in different forms with the aim of either collecting data from the respondents or grading them. These measurement techniques are used specifically at appropriate situations and demand for assessment. The word “assessment” takes its root from the Latin word *assidere* meaning ‘to sit beside’. It implies that a teacher is expected to help students to gain understanding or learn by sitting beside them to give instruction and determine its effect on them.

Kumaran and Sankar (2015) described assessment as the most central and significant activity in any educational institution. Assessment of learners in education touches the three main domains of learning which are the cognitive, psychomotor and the affective domains. These domains can be measured at any level of education especially when the teacher and decision makers are interested in doing so. Educational measurement is the assigning of

numbers to constructs in the classroom, which is meant for several purposes such as certification, diagnosis, remediation and promotion among others.

In educational assessment, techniques which are commonly used to examine learners are test (examination), observation and interview. Test can be administered via paper-pencil, oral interview and/or computer based depending on the teacher, population of students, purpose of the test and other factors. The use of Information and Communication Technologies (ICT) has brought certain developments to teaching and learning processes (Mubashrah, Tariq & Shami, 2012).

There is a gradual movement from paper-pencil based test to computer-based test. The use of ICT has been covered with such designations as Online Assessment, Computerized Assessment, Computer Assisted Testing (CAT), Computer Based Assessment (CBA), Web-based Assessment, Computer Based Testing (CBT), Computer Aided Assessment (CAA), Computerized Adaptive Testing (CAT) and E-Assessment. An assessment where computers become an essential part of the process of test delivery, storage of responses/answers, grading of responses as well as reporting results (feedback) is referred to as computer-based assessment (Mubashrah, Tariq & Shami, 2012). Mubashrah, Tariq and Shami (2012) further observed that the traditional paper-pencil method of evaluation is gradually being replaced by automated grading systems. Though the extent of this change is yet to be quantified by academics and educational administrators; computerization of the examination processes has started catching on in Nigeria with e-registration, administration of automated objective test and delivery of result by examination bodies.

In many fields of life, computer has successfully complemented human activities such as stocktaking, sales, file/document transfer through e-mails, fax and short message service, communication, bank tellers, classroom instruction and many others. Computers have been

seen over time to have the competence to do what humans can do successfully especially with little or no supervision required from man (Mitchell, Russell, Broomhead, & Aldridge, 2002).

The field of education has not been left out of the move towards the use of computers and their application as supportive tools for handling school administrations and control. Softwares have been developed for handling educational activities like application, registration, school fee payment, students' attendance, testing and academic reports. The optical mark reader has been developed for scoring of objective tests and this has become popular in education globally (Alber & Debiassi, 2013). There is the aspect of the use of computer to grade essay tests and this concept is commonly called Automated Essay Grading (AEG) or Automated Essay Scoring (AES) or Computerized Assessment of Free responses (Whittington & Hunt, 1999).

AEG can be perceived as a set of network or system that uses computer techniques to correct and score human written compositions automatically. The Australian Curriculum Assessment and Reporting Authority (ACARA, 2015) stated that AEG is the use of computer algorithms to emulate human scoring. A computer may deliver a mark as overall feedback and worth of an essay and as well offer feedback concerning the mistakes/blunders the students have made and means to develop the essay properly through revision. As a matter of designation, such systems are neither referred to as AEG nor AES but Automatic Writing Evaluation (AWE) systems (McNamara, Crossley, Roscoe, Allen, & Dai, 2014).

Automated marking of objective test is common, yet it has been criticized because it assesses the lower order of the cognitive skill (Whittington & Hunt, 1999). The multiple-choice type is the most used form of objective test. Though, there are other forms of objective test like alternative response type, matching type and the completion or supply item type of objective test. The truth is that essay measures other skills that the objective test is unable to. Most times, an additional open-response part is added to tests and then marked manually by

human graders. Balfour (2013) made a broad division of test items into 'select' and 'generate'. Select refers to coded, multiple choice questions (MCQ); it requires a form of readily constructed items; selection from supplied response options. Generate, on the other hand, refers to open-ended responses. To generate responses implies mental processing of information and the necessity to bring in ideas that are not in the question asked.

An essay can be defined as a test item that requires a composed response by the examinee such that no single response pattern can be listed as most correct usually in the form of one or more sentences, and the correctness and value can be judged only by one who is skilled or informed in the subject matter. The skill of writing especially an essay is a cognitive activity as it uses constructed-response task since words are needed to express thoughts and knowledge about an idea, situation or individual. Writing is also a more comprehensive skill as it can take many forms like arguments, persuasion, explanation, speech, storytelling (reporting) and many other forms.

There are two forms of constructed response essay tests, the short and the extended response types. The short essay type entails learners' ability to remember and understand the question item and respond in favour with few words, it also requires not more than three sentences while extended essay type entails learners' ability to demonstrate evaluation and creation of ideas from related texts in not few words, it also does not have limited number of sentences that respondents should write (Callear, Jerrams-Smith & Soh, 2001). An extended response essay type demands that a learner provides details, examples and supporting of answer with reasons while short response essay type does not demand for all of these. Short response type often necessitates that the learner gives just about two facts or pieces of information about a subject matter of interest while extended response usually necessitates that the learner gives about four or more facts or pieces of information over a subject matter (Creswell, 2014).

AEG is used to program the sequence of assessment of either of these two. There are about fourteen products available for the assessment of either short response or extended response type of essay questions. Some of these products are Education Testing Service I (ETS I), e-rater, c-rater, MARKIT, SEAR, ES4ES, IntelliMetric, Automark, Intelligent Essay Assessor, PEG, Intelligent Assessment Technologies (IAT) to mention but a few. Kumaran and Sankar (2015) observed that manual assessment is less preferred to automated assessment because it helps in avoiding bias and human errors; more so, it also helps in conserving teachers' time. They further reported that short response allows a learner to compose their answer to a question in a maximum 3 or 4 sentences. Short answers are typically based on the sentence length but are not required to be grammatically correct (O'Shea & Bandar, 2010). Short answer connotes a reply ranging from one word to a dozen words. Grading of short essay answers is one of the most complicated domains because semantic similarity exists between two words or sentences which a student may use in case the student forgets the target word at the point of question response. Short answers are usually centered on length of the statement but not the grammatical accuracy of the statement (Omran & Aziz, 2013).

The revised Bloom's taxonomy categorized learning outcomes into six levels: Remembering, Understanding, Applying, Analyzing, Evaluating and Creating (Anderson & Krathwohl, 2001). The revised taxonomy explains that multiple-choice questions address mainly the lower levels which is the first two learning outcomes and that the need to measure higher levels of the taxonomy will be catered for by using the free response items. This has necessitated the use of essay items by teachers. Though to mark an essay for a human rater, the need for higher level of examining the response, breaking the information, building inferences, defending opinions and making judgments by combining ideas and patterns in the response to a test item faces the marker of any essay either as short response or not.

Pulman and Sukkarieh (2005) explored computational linguistic techniques in grading short responses automatically. It deals with responses ranging from few words to about 5 lines maximum. These questions are from a known examination called GCSE (General Certificate of Secondary Education) which are stemmed from key words like state, describe, explain, suggest and the like. Pulman and Sukkarieh (2005) used three different machine learning methods which are Decision Tree Learning (DTL), Inductive Logic Programming (ILP) and Naive Bayesian learning (Nbays) where 84.05%, 69.56% and 88.03% agreement respectively was achieved when compared to human marker using annotated approach while the non-annotated approach yielded 63.81%, 66.05% and 62.1% agreement levels respectively.

Attali (2007) reported that automated essay grader has been discovered to have ability to function as a more active cognitive tool. Also, Dikli (2006) believed that computers have the ability to function as a more operational cognitive tool in measuring and evaluating essays. Literature on Automated Essay Scoring (AES) revealed that computers have the capability to play the role of an active thinking and measuring tool (Attali & Burstein, 2004). Dessus, Lemaire and Vernier (2000) stated that computer softwares, when well-structured in a designed model and once it is uploaded in a web-based environment; it has the capacity to assess students' knowledge based on free text writing. They further highlighted that computers do not only have the ability to grade students' responses to a question but it can also engage students in an iterative process of writing and or assessing at a distance. Dessus, Lemaire and Vernier (2000) developed a computer program with the ability of comparing essay writing and the text of a given course of study on a semantic basis discovering the match between the course and the essay which is what the software that is used for plagiarism uses.

The use of technology for scoring students' essay is not new and it was first successfully done in 1973. Then the need for punched cards and mainframe computer made it inaccessible to most instructors and learners (Shermis, 2014). As moderately young as it is, AEG is known

to have recorded about 45-years of history. Historically, the use of computer as an assessment tool has experienced several developmental stages. Its use started in the 1950s with computer games, puzzles, and programmers which developed test compilers to identify errors of syntax and errors of styles. In the 1960s, computers were used as learning machines, also in the 1970s computer programs were developed in automating multiple-choice test. In the 1980s programs were written to support learning though with less attention to assessment; and in the 1990s computer programs were developed with more enhanced assessment tools beyond short item and integrated learning systems. This era also experienced the growth of internet which gave birth to online and on-the-spot grading and correction of answers (Burkhardt & Pead, 2017).

However, Page is most acceptably referred to as the pioneer of AEG (Wang & Brown, 2007). In 1966, Page designed a computer scoring program named Project Essay Grader (PEG). Page considered text qualities that can be certainly extracted by computers and then used multiple linear regression to determine the weighted features that could best predict a teacher's score. Page mentioned "word length, essay length in words, number of commas, number of prepositions and number of uncommon words (which is negatively correlated with essay scores) as features to be reckoned with in essay scoring. This approach to scoring of essays has changed over time.

In fact, what the computer does in an AEG context is to evaluate the written text into noticeable modules. Different AEG systems evaluate different numbers of these modules. Page and Peterson (1995) referred to these modules as "proxes" or *approximations* for underlying "trins" (i.e. *intrinsic* characteristics) of writing. It is these noticeable modules that automated essay scoring engines identify computationally and subsequently use to determine essay scores. AEG statistical models are developed by weighting the various noticeable modules as they relate to intrinsic characteristics of writing (Shermis, Burstein, Higgins, & Zechner, 2010).

Project Essay Grade (PEG) is a model formed by taking five intrinsic characteristics of an essay (*content, creativity, style, mechanics and organization*) and their linking proxies. An example of a simple proxy is essay length (Burkhardt & Pead, 2017). The empirical proof suggests that the lengthier an essay is, the more highly valued it is by a rater (Monjurul & Latiful, 2012). This can be stated to be true because the lengthier an essay is the more likelihood that the writer has released more details in relation to the topic or subject of discussion. However, this relationship is not directly proportional, but logarithmic (Shermis, Burstein, Higgins, & Zechner, 2010). Though more explicitly, it appears as if an essay's length is important up to a point, but beyond a certain edge, it carries little value. Where that "edge" is, is a function of the average essay length.

The length of an essay may not necessarily be a factor in grading the essay depending on the discipline and the purpose of the essay, but it surely reflects fluency of the writer and partly the level of knowledge about the content of discussion (Klobucar, Elliot, Deane, Ramineni, Deess, & Rudniy, 2012). Though Kobrin, Deng and Shaw (2011) noted that a significant and positive relationship exist between essay length and the score assigned to it, they further indicated that it may vary based on the area of discipline. Deductively, the length of an essay is indicative of the ability of a learner to regulate and organize their writing processes effectively.

If most essays have 150 words and a candidate's essay has 200 words, then the writer has probably taken advantage of any contribution the essay length would influence in a grader's decision-making. Further, in the PEG (Project Essay Grade) framework, an example of a complex proxy might be a count of the number of times a word like *because* is used in a particular essay. Though this counting method is admittedly shallow, yet it is a pointer of sentence complexity, and is also tied abstractly to the inherent characteristic of style. It is useful to note that this one proxy is joined with several others to estimate the *trin* (intrinsic characteristics of

an essay) and would not usually be used as a single pointer for the trait. Provided that an AEG tool does not rely on the length of an essay but on relevant tips for assessing writing quality, association of score to length of an essay is inevitable and anticipated.

Balfour (April 5, 2013) in *The New York Times* website broadcasted that Harvard EdX introduced an Automated Essay System (AES) to be used within their Massive Open Online Courses (MOOCs) where students enroll once they have access to course material that is also available through the internet. This MOOCs uses numerous web technologies such as computer-based assessment, video presentation, webinar and online communication forums. MOOCs has registered over 150,000 students, the system has succeeded in reducing the time instructors spend on instructing and assessing work done per learner, computers are used to provide feedback and formulate problems with correct answers. Markoff (2013) stated that two largest MOOCs institutions publicized to make provision of a mechanism that can be used for automated assessment of essay.

Kumaran and Sankar (2015) noted that computer systems have the ability to award scores that can correlate properly to scores given by human scorers automatically, extracting the linguistic features of the essay which have yielded agreement rate of 97% in areas where scoring is more concerned with style of writing. Conole and Warburton (2005) reviewed studies on computers with the substantial likelihood both to ease assessment load and deliver pioneering/groundbreaking methods of assessment.

Computer-based assessment can be subdivided into stand-alone applications that only require a single computer with applications that work on private networks and others are designed to be delivered across public networks such as the web as with online assessment. Burkhardt and Pead (2017) reported that computer-delivered test facilities have proposed less cost in the delivery of tests and test administration since they remove the cost of paper and all, automatic collection of students' responses and marking of such especially without the need

for human interference and other forms of assessment techniques like simulations and incorporating interactive multimedia elements in examinations.

A common challenge to AEG in research is the non-existence of clear cut rules for deciding master texts and an acceptable standard to calibrate human marks. To compare the performance of an AEG system, some set of uniformity should be defined (Omran & Aziz, 2013). Dikli (2006) identified that AEG is weak because computers can not do what human graders will do for the reason that the computer would do only “what it is automated to do” and it wouldn’t “increase the value” of an essay. Another criticism is the construct objections. That is, the computer counts variables that might not be genuinely important in essay grading, focusing on strict aspects rather than structural ones. Yet, an AEG system called Expert System for Essay Scoring (ES4ES) was developed by Ade-Ibijola, Wakama and Amadi (2012) for scoring free response and it was used in a Nigerian University in a computer science course. They reported 0.71 correlation coefficient between ES4ES and the scores allocated by the human experts.

Ade-Ibijola, Wakama and Amadi (2012) reported that the validity and reliability of AEG systems has proven to be very high since studies and research in this field of evaluation is continuous. In addition, Pulman and Sukkarieh (2005) concluded that quite a number of empirical researches has been conducted to find out the accuracy and reliability of AEG systems with scores shows its agreement rate with human expert graders in any field of study (English studies, Arabic, Spanish Language) where it has been used before. Moreso, AEG systems has been proved to be very consistent.

Clipa (2011) discussed the description perceived from students’ point of view about human assessors into the competence criteria which is meant to be assessed, the method of assessment and the assessor subjectivity. Clipa (2011) conducted a study where 909 students were involved. The study concluded that students who have higher performance in grades

perceived both the subject and the human assessor as better compared to those with lower performance, keeping in mind that a teacher's profile largely depends on students' performance. The assessor profile on subjectivity was also perceived positive by students with higher performance. Assessor's competency had five components which are knowledge presentation, understanding the information, applicability of the information, problem solving skill and psychosocial competences of the assessor which was ranked 3.48, 3.45, 2.96, 2.69 and 2.42 respectively.

Human assessors' experiences influences grades that is given in assessment (Hughes, Johnston, & Mitchell, 2018). These experiences are cumulative of both academic and real life impacts of events around the human assessor. An assessor is known as the gatekeeper of any profession. However, there are situations where a student is on the borderline of either to pass or fail, the assessor does not seem to find a reason to pass such student but just appear that finding a cause to fail such a student is needed. Background information about a learner also influences grades.

Other factor of consideration is benefit of doubt, an assessor will at some examination especially at lower levels (100L) may appear to wave errors as acceptable, ones such error can be termed to be a redeemable error, as the assessor will perceive that such errors will be adjusted upon as time goes on in the cause of the study (Hughes, Johnston, & Mitchell, 2018). Considering all these the need to develop software through programming can be seen.

The term programming is a group of codes and instructions given to computer to take certain actions and decisions at the expiration or beginning of another action. Software is a collection of guidelines that is used alongside with hardware mechanisms to build a structure that is meant to perform a given task or an application (Kumar, 2017). The development of software programs as being evolving over decades, changes and improvement from different organizations and developers has been seen.

The term “software engineering” was first mentioned in 1968 at the North Atlantic Treaty Organization (NATO) Software Engineering Conference. Abbas, Jeberson and Klinsega (2012) described software engineering as an approach that involves design, implementation, assessment, maintenance, analysis, testing and reengineering of software. They further stated that software engineering is concerned with the quantifiable and systematic method of development, use and preservation of software. There are five main classifications for software engineering and they are the software development process, software management, software engineering techniques, software re-engineering, and software applications.

There are quite a number of Software Development Methods (SDM) which a developer can consider but more importantly the usability of the application is a pivotal decision to which SDM a software developers should use. The ease with which a software application, website and or a user-operated device can be used by an end user is primarily due to the friendliness of the user interface. The user interface must be informative in simply, clear and concise language. However, the architecture of usability can be established if the following questions can be positively ascertained. At first contact, how easy is it for users to accomplish tasks?; after first contact, how easy can tasks be performed?; after a very long time, how fast can expertise be re-established?; how error prone are users, how severe are these errors and how fast can such errors be recovered from?; finally, a feedback on satisfaction from users is also a good component of efficient usability.

Literatures seem to present the *agile* and the *traditional* approach of Software Development (SD) as two major SDM that are commonly available and applicable. Industrial and commercial software developers have been doubtful about other forms of solutions in SD as they are categorized as difficult to grasp. Though, literatures as majorly being written by SD practitioners and experts while academia reviews has being quite scare (Abrahamsson, Salo,

Ronkainen, & Warsta, 2002). The traditional model has the Waterfall model, V-model and the spiral model are the major types while the agile model several models like Adaptive Software Development, Open Source SD, Dynamic Systems Development Methods, Feature Driven Development, Scrum, Extreme Programming, Crystal family of methodologies and the Rational Unified Process yet more recently, the Agile modeling and pragmatic programming are referred to as “other agile methods”.

This study is interested in using Economics as a subject to validate the software to be developed. Economics as a subject at the senior secondary school level is needed to prepare students with the basic skills and information which could help solve elementary economic problems in any society. The underlying principle is to place Economics as a subject of high importance to everyday life of a student. Examinations at the senior secondary school certification level as existed for about 51 years till date. The format of questions has been two major sections. Paper one is the objective questions part while the paper two is the theory (free response type) aspect. The paper two includes two sections where the first section deals with mathematical, graphical and or computational knowledge about Economics concepts.

Ogundare (2017) reported that the results of WASSCE for the year 2017 recorded a positive progress in performance as candidates obtained credits and above in at least six subjects (including Economics) by 70% as compared to that of year 2016. Adesulu (2018) reported that in the WASSCE result for the year 2018, 50% of the total candidates obtained five credits while for the year 2016 and 2017 the performance was 53% and 59% respectively.

In the WASSCE 2017 Economics examination, the chief examiner for Economics made comments like. Students’ performance was generally unsatisfactory, many performed below par and that only a few brilliant students did well. He further added that some responses of students showed that they were guessing as answers were completely unrelated to the demands

of the question, just a handful of candidates failed to elaborate on points stated and some students could not present answers to questions in meaningful and coherent English.

Though the Chief Examiner highlighted areas of strength as displayed by students, that they were able to use significant instances to illustrate answers, apply correct formulas in data-response questions and also were able to state laws and principles clearly and concisely. Furthermore, he suggested solutions to the weaknesses of students; understanding questions before attempting them is vital, the need to expatiate on points been raised and supplying of relevant illustrations would suffice for good grades and that students should learn to present solutions in clear and coherent English (Chief Examiner, 2017).

The theory part of the Economics examination as being an extended essay response format which expects students to express themselves in quite long essays or limitless sentences. It is anticipated that students are to attempt solutions to only three questions out of six theory questions from the section B of the essay (paper two) while only one question is expected to be answered within the first two mathematically inclined questions which implies therefore that only four questions in all is what a student is expected to answer to gain 100% in a typical economics examination at the WASSCE level. The three questions to be answered by students amounts to a total of 12% of all the topics in the three-year study of a student.

It therefore indicates that one student's reads randomly about seven major topics which is about 27% of the whole Economics syllabus/curriculum an excellent performance can be recorded. This development simply raises a question on the validity and reliability of the examination since just a little fraction and percentage of the curriculum will suffice for an excellent knowledge in the subject. Since there are 26 topics structured into 34 teaching units and distributed to be taught in three-year period of SS1 through to SS3 (Yusuf, 2011). Though,

research as not established the reason for this format of question, yet it is true that the validity and reliability of a test can be increased if more questions can be answered in the test.

The fact that human assessors are also subjective when marking can also be a reason for the unsatisfactory performance that the Chief examiner reported. The AEG software could therefore be a sufficient way of raising more questions to be answered in an Economics examination as a stand point of attempting all the six theory questions could be done in on-shot and yet grading would not be a problem. On the premise of answering six theory questions it implies that at least about 27% of the curriculum content should have been covered to gain an excellent performance compared to the existing 12% requirement. Time allotment could also be a challenge, more reason short-essay response would be a better idea.

Statement of the Problem

Automated Essay Grading (AEG) has been studied by several researchers in other parts of the world and it is of note that the drive to develop and employ AEG systems has been with the aim of using it in large-scale tests or examinations to reduce time and effort spent on marking of responses. Few organizations or institutions and individuals have successfully developed and used AEG systems. This can be attributed to the cost for developing such a system and the fact that it is expected that the system should be able to score students validly and reliably in any field of study. Some developers of AEG systems made the initiated or developed system available for the educational research community or researcher while quite a number of them have restricted or reserved their systems due to proprietary law attached to such system.

The number of essays required for training the Project Essay Grader (PEG) is 100-400, Intelligent Essay Assessor (IEA) requires 100-300, for the E-rater 465 is required, IntelliMetric needs 300 and BESTY must have 1000 essays before it could be used to grade essays (Semire,

2006). There is a need to address the number of essays required to train any Automated Essay Grading system to be able to correctly score with a view to bringing it down drastically.

No existing and reviewed grading system (software) has used only one model answer to be supplied by the teacher or an examiner to score essay responses. A software that will depend on identifying a correct representation of the contents of the model answer and correctly assign scores or marks as the human scorer.

Automated Essay Grading systems are scarce in this part of the World (Africa including Nigeria), this study intends to fill this gap by developing a viable software that can score short-essay responses in Nigeria. Essay responses are of two forms. The extended and short essay format. The short-essay response format is an answer to a test item in not more than 3 or 4 sentences but any response more than 4 sentences is regarded as extended essay. Using computer to grade short essay responses is still very much viable in relation to sentence structure and limited common words unlike extended responses.

In scoring short and extended essays a scorer either human or computer is expected to identify relevant contents or points that is score worthy and award adequate score. Human markers would spend more time in reading either an extended or short response essay unlike a computer that has the capacity to identify points no matter the length of the sentences at equal time spent. Existing softwares have established a correlation average of 0.70 between human expert and their AEG but then there is the need to push the correlation of scores given by AEG and human scorers to a higher level.

Literature has shown that most AEG score writing of English essays (like compositions, letter writing) using the rubrics like content, grammar, mechanical accuracy, expression and organization. Other subjects such as sciences (Mathematics, Biology, Chemistry, Physics, Statistics), Arts (Literature, Government, Religious Studies), Commercial and Social Sciences

(Economics, Business Studies, Office Practices) where students responses can be written in prose form as a response to a test item likewise needs automated essay scoring system.

The commonly available grading systems in Nigeria are the computer systems that scores objective test where options are ticked and scored. The need to move beyond this status-quo into marking essays shows another feasible gap that necessitates the purpose for this study.

Purpose of the Study

This study aimed to design, develop and validate an automated essay grading (AEG) software that can be used to grade short-essay response format items.

Specifically, the study sought to:

- a. develop the Essay Test Assessor for grading Economics essay test items at the Secondary School level,
- b. examine the scores of students' in Economics using the Essay Test Assessor (ETA),
- c. examine the scores of students' in Economics from human raters,
- d. determine the validity of students' scores in Economics using the Essay Test Assessor (ETA) and human expert markers in Economics, and
- e. determine the efficiency and speed at which the Essay Test Assessor (ETA) operates.

Research Questions

The researcher raised the following research questions and proffered possible solutions:

- a. How does Essay Test Assessor (ETA) score students' short response essay test items?
- b. What is the performance of students' in Economics resulting from the Essay Test Assessor?
- c. What is the performance of students in Economics resulting from human raters?

- d. To what degree is the Essay Test Assessor scores in Economics test related to the scores by human raters?
- e. What is the measure of time gain between the two human scorers and Essay Test Assessor based on the average words of response?

Research Hypotheses

The following null hypotheses were formulated and tested in this study:

H₀₁: There is no significant relationship between the scores assigned by the Essay Test Assessor (ETA) and the scores assigned by human scorers; and

H₀₂: There is no significant relationship on the time gained between the two human markers based on the average words of response from each respondent.

Scope of the Study

The scope of this study was to develop and validate an Automated Essay Grading software using Economics test items at the Secondary School level in Nigeria. The population of this study consisted all Senior Secondary School students in the South-west geo-political zone of Nigeria. This study was carried out among senior secondary school II (SS 2) students of public secondary schools in South-west geo-political zone, Nigeria (Ogun, Osun, Oyo, Ekiti, Ondo and Lagos state).

Simple random sampling was used to select three states from the South-west geo-political zone. The proportional sampling technique was used to select five percent (5%) of each of the public schools in each state. Therefore, a total of 55 public senior secondary schools was the sample for this research. A total of 1,200 respondents was involved as sample in this study.

15 short-essay test items related to SS I and II Economics curriculum content in WAEC and NECO for year 2014/2015 through to 2016/2017 sessions was selected as part of the questions to be asked. The marking guide for these few selected test items was used as the model answer for the Essay Test Assessor (ETA). The benchmark for categorizing the high, average and low scorers was the range of scores divided into 3 groups so as to ensure the truth representation of respondents' scores. The test items had a total score of 30 marks. The correlational research method was adopted for this study in order to establish the validity of the ETA. This is aimed at finding the extent of relationship that exist between the scores given by the ETA and human expert marker.

Operational Definition of Terms

The following terms are operationally defined in the context they were used in this study:

Essay Test Assessor (ETA): A software for scoring short-response essay test items.

Validation of Essay Test Assessor: Comparing of scores given by the ETA to those by human raters.

Automated Essay Grader (AEG): General appellation for all automatic program and software meant for grading essay test items.

Grading: Assignment of scores to any test response constructed by examinees.

Senior School Certificate Examination: Economics test items of West African Examination Council (WAEC) and National Examination Council (NECO) standards.

Development: Process of constituting variables, links and factors to producing the grading software.

Significance of the Study

The beneficiaries of this study are the teachers, students, school administrators, examination bodies or organizations and employers of labour. This study focuses on scoring as a skill that computers (tools and software) can also do. In comparing Automated Essay Grading (AEG) software to human scores, the agreement between the computer tool and human scores has been discovered to be quite close such that the AEG tool has been seen to be an effective tool in grading complex critical thinking performance essays.

The findings of this study is of great significance to teachers at the Secondary School level of education; as it would reduce time spent on assessment or grading. Grading of essay usually takes about 30% of teachers' time out of every other activity a teacher must get involved with in relation to teaching and learning process. Grading students' papers can be burdensome for teachers. Particularly, when there are large number of students and if they have been assigned repeated writing assignments; providing individual feedback is also quite time consuming. The AEG is not intended to replace teachers but to complement teachers' efforts in the process of helping learners to acquire knowledge since teachers are still expected to feed the AEG tool with question items, model answers, allot scores and program the expected feedback algorithms.

Scoring with AEG makes students' scores to be more reliable and valid since teachers' biasness and subjectivity that human scoring are susceptible to are eliminated. Score accuracy would increase and reduction in time between administration of a test and delivery of scores as reports to parents or guardians would also be quite fast. Timely feedback in AEG system is also beneficial to students because it can help break the barrier that exist for students seeking further explanation or interpretation on confused area of previously failed skill. This would enable students to quickly review areas of misunderstanding and confusion, it would encourage sustained and continuous student engagement in knowledge acquired. Blood (2017) reported

in a study where students were encouraged to use an AEG tool (IEA) for reviewing their essays repeatedly until they are contented before submitting to their professor for final grading. In this survey, 98% of the students reported that they would certainly use the AEG tool if it were available for their remaining courses.

School administrators (proprietors) would benefit from the use of AEG software since these software performs repeated functions without boredom, tiredness and variations in scores assigned. It further has the ability for adaptability (within preprogrammed pathways), flexibility (as tests can be done at any time of the day by as many students as possible) and the ability to decide without been subjective (sense of bias due to any form of background knowledge about an examinee) and or favouritism which is the opposite of subjectivity.

Examination bodies or organizations would also benefit from an automated essay grading software. It will further help as a money-saver for these educational organizations who are involved in conducting examination since they engage and pay human graders to score free response items. Reporting of grades and scores will be easier and faster also. AEG has been discovered to provide immediate feedback to variety of issues in essay writings like grammatical blunders and discourse related if programmed to do so. Standardized assessment entails an expensive and logistic effort; automated assessment has the potential to drastically decrease time and costs for training and employing human scorers.

Employers of labour both in the private and public sectors would benefit from this study in situations where the need to test more than knowledge and ability to recall information. It would be useful for employers because it would not be a burden to them in relation time and other resources that would be needed to grade an essay test, more so appropriate selection can be made from large number of prospective employees who may be interested in the job. No doubt the findings of this research will be useful and germane to the educational process of teaching, learning and evaluation.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

This chapter presents the review of literatures relevant to this research work. The review focused on the following sub-headings:

- a. Introduction to Assessment, Automated Essay Assessment and Grading Software;
- b. Automated Assessment and Challenges of Automated Essay Grading Systems;
- c. Automated Grading/Marking of Objective Test;
- d. Rationale for Automated Essay Grading;
- e. Validity and Reliability of Automated Essay Grading Systems;
- f. Teaching, Learning and Assessment of Economics at the Secondary School Level;
- g. Agencies/Institutions using Automated Essay Grading System in Nigeria and Africa;
- h. Issues Relating to Software Development and Use of Automated Essay Grading in Nigeria;
- i. Conceptual Framework; and
- j. Appraisal of Reviewed Literature.

Introduction to Assessment, Automated Essay Assessment and Grading Software

Education is expected to prepare students for a life-long and independent learning whereas assessment plays a fundamental role in facilitating this lifelong learning process. Assessment is relevant in the context of teaching and learning, its aim should be to enable meaningful and relevant learning experiences. Assessment influences the direction at which learners learning efforts are channeled as well as the quality and quantity of the effort the learner will put in. Assessment is expected to aim at making learners think critically so as to improve on knowledge and skills learnt. It is also an integral part of the design of a course of study (Heinrich, Milne, Crooks, Granshaw, & Moore, 2006). Assessment evaluates learning and/or

teaching outcomes (Berman & MacGregor, 2001). Assessment varies base on what is to be assessed (personality/individual, character or traits, picture, aesthetic, knowledge and skill) as well as the purpose of the assessment either for diagnostic and predictive. This study focused on assessment as it relates to free-response writing what is commonly called essay.

Free-response writing is a complex ability that a learner would have to develop and cultivate over a period of time in practice. Expertise and proper improvement in free-response writing is a consequence of years of deliberate practice coupled with coaching rather than talent (Pieterse , 2013). The nature or context of any free-response writing varies from work place, to speech, seminar, letter, examination and so on. In free-response writing; development of ideas, thoughts and learnt skills cannot be undermined. In educational free-response writing, assessment plays a significant role in either encouraging the learner or discouraging the learner to put in more effort which may lead to improving future free-response writing and the motivation to learn more effectively. Assessment is a broader concept that comprises of graded and or ungraded events in and outside the classroom environment (Berman & MacGregor, 2001). Assessment could involve evaluation of behaviours or actions that do not explicitly measure learning like attendance, bullying, turancy, peer pressure to mention a few while grading is indicative of a specific learning outcome.

Grading is associated with giving of marks to students in a test. It is also the assessment of the cognitive, psychomotor and affective domains of learners. Human beings can be assessed in all educational domains. Grading has several purposes some of which are for educational placement, selection for a job, remedial need of a learner, to identify the extent which a learner has learnt an instruction content, to predict or foretell learning outcome in the future and to identify or classify learners into groups base on capacity/intelligence level. Grading of free-response writing or essay is a skilled activity of an expert in a field of study. The use of computer software in grading free-response writing is a no mean feat, it has resulted in helping

to reduce time spent by experts on grading and also led to consistent accuracy in awarding grades too (Williams & Dreher, 2006).

Steedle and Elliot (2003) drawn a sample of about 1,500 which was scored by two human experts. A Pearson correlation was done to ascertain the agreement between the two human expert assessors; also between the AEG system and human assessor 1 and also between the AEG system and human assessor 2. It was reported that the AEG system (Intelligent Essay Assessor-IEA) had a 0.87 correlation (a range of 0.79 to 0.93) with the average of the two human experts while between the two human experts 0.77 correlation (a range of 0.67 to 0.83). However, the IEA was trained using over 450 previously administered responses before this grading was done. The primary idea is to arrive at all possible words or phrases which if written by a student can be scored correctly by humans either attracting part or full allotted scores.

The Australian Curriculum Assessment and Reporting Authority (ACARA) (2015) conducted a research to explore the capacity of automated essay scoring systems to imitate human marking successfully. ACARA (2015) employed four automated essay scoring vendors and their engines/systems which are the Project Essay Grader (PEG) system (Measurement Incorporation), the Intelligent Essay Assessor (IEA) system (Pearson), the Constructed-Response Automated Scoring Engine (CRASE) system (Pacific Metrics) and the Lexile Writing Analyser (LWA) system (MetaMetrics). 1,014 essays were collated from the National Assessment and Surveys Online Program (NASOP) out of which 340 was removed to train the tools while the remaining 674 essays were marked by two human raters and the four engines.

ACARA (2015) independently permitted each software vendor to report their findings where they all reported in quadratic weighted kappa. Marker 1 and 2 had 0.79; while 0.74, 0.76, 0.77 and 0.73 weighted kappa between marker 1, PEG, IEA, CRASE and LWA respectively; furthermore 0.72, 0.77, 0.75 and 0.82 weighted kappa between marker 2, PEG, IEA, CRASE and LWA respectively. The set of automated essay grading engines that was used provided an

agreeable level (that is 0.75 as reported by Fleiss, Levin, & Paik, 2003) of reliability and consistency in grading writing at rubric and total score levels.

Wang and Brown (2008) involved 1047 respondents in a research to establish the relationship between AEG system and human scoring. The result showed that the overall holistic scores assigned by the tool has no statistically significant correlation compared with human raters, though correlation was significant the two teams of human raters. There was no significant correlation between the scores assigned to focus, development, organization and mechanics scores by the AEG tool and the scores assigned by human graders except for sentence structure scores assigned by the AEG system and the scores assigned by the human raters that were statistically significant. Researches that used the IntelliMetric noted the need to use about 250 pilot essays to instruct and train the system on how to score future essays that would be given to the system to score.

Wang and Brown (2007) investigated the validity of AEG by comparing group mean scores assigned by AEG tool (IntelliMetric) and human graders. This research sampled 107 respondents and also used One-Way Repeated-Measures ANOVA to examine the difference between the AEG and human raters' mean scores. The result showed that there was a significant difference between the group mean score assigned by the AEG tool (IntelliMetric) and the group mean score assigned by human raters on the standardized writing test where the significant difference was higher in favour of the AEG tool. The Interclass Correlation Coefficient (ICC) statistic which computes the ratio between-group variance and total variance indicated 0.62 which is an acceptable level of inter-rater reliability for the overall holistic scoring, using 0.60 as the cut-off value for acceptable level.

The need for free response items and the advantages attached to it have necessitated the development of AEG systems. Despite critics of the use of computers to grade human free response items, benefits such as time constraints to give instantaneous feedback to students and

the fact that multiple-choice items are not adequate for measuring higher level learning outcomes have accented to the fact that AEG systems have limitless paybacks. Traditional examination boards like GRE (Graduate Record Examination), TOEFL (Test of English as a Foreign Language) (Siddhartha & Sameen, 2010) and GMAT (Graduate Management Admissions Test) have being using it in the conduct and administration of a part of their examination at one point or the other.

Automated Essay Assessment and Grading Software: In recent times, the use of computer software in scoring essays or free text write-ups as being studied and explored from diverse perspectives which necessitates the importance of having effective system for this purpose. Quite a number of models, add-ins, approaches, explanations and applications has been developed by researchers and software merchants either for profitable and academic or hypothetical purposes. Some prominent and highly valued softwares in relation to how often they have been used by organizations or researchers, performance and efficiency have been selected for review in this study, include; the Project Essay Grader (PEG), C-rater, Electronic Essay Rater (E-Rater), Intelligent Essay Assessor (IEA), Educational Testing Service 1, AutoMark, IntelliMetric, SEAR, Markit and ES4ES.

Project Essay Grader (PEG): Monjurul and Latiful (2012) reported that the Project Essay Grader (PEG) was designed by Ellis Batten Page in 1966 while in University of Connecticut which was the first version and later versions were developed and deployed in the 1990s. PEG showed that an automated grading system is not vague to a human rater. Ellis Batten Page made the first attempt at scoring/grading essays using computer.

Page used terms like '*trins*' and '*proxes*' in the process of grading an essay. *Trins* refers to the core variables in an essay or *intrinsic* variables of interest within the essay (what a human grader would look for but the computer can't directly measure) to simulate human rater

grading while *proxes* refers close to core variables like fluency, diction, grammar, punctuation, relative pronouns and other parts of speech, as an indicator of complexity of sentence structure; variation in word length to indicate diction (because less common words are often longer).

The grading approach of PEG involves a training (exercise) phase and a grading phase. The previous phase uses some sample of essays while in the latter phase *proxes* are estimated on each essay and these variables are computed into a standard regression equation $\text{Score} = \alpha + \sum_{i=1}^k \beta_i \rho_i$ where α is a constant and $\beta_1, \beta_2, \dots, \beta_k$ are the weights (i.e. regression coefficients) associated with the *proxes* $P_1, P_2, \dots, P_k$. The predictive characteristics of PEG are average length of word, number of commas, number of prepositions and number of uncommon words, which are stated to be negatively correlated to the essay (Hearst, 2000).

An experimental research carried out by Monjurul and Latiful (2012) showed a high multiple regression correlation of 0.87 with human graders. Page carried out an experiment with a group of students' essays that teachers had already graded where an automatically extractable textual features and applied multiple regression was used to determine an optimal grouping of weighted structures that best predicted a teachers' score (Hearst, 2000). Hearst stated that PEG showed a multiple R to teachers' scores to be 0.78 and as strong as 0.85 correlation between two or more teachers.

In spite of the success PEG recorded, the first version of PEG received lower acceptance within the educational community due to the fact that it uses indirect measures in grading essays. Critics argued that since PEG uses indirect measures, the system is vulnerable to cheating because students can boost their grades remotely using tricks/deceits; students could simply write a longer essay than usual (Hearst, 2000). Also, because it uses the indirect measures, it therefore does not capture important qualities of writing such as organization, style and content as it won't provide instructional feedback to students.

More so, it needs to be trained usually from a set of essays previously marked manually, which in turn enables the marking of new essays. Page's data training needs several hundred essays consisting of 50% - 67% of the total number of the essays to be graded. Another drawback is that it grossly rely on statistical multiple regression techniques to grade essays usually on the basis of writing quality taking no account for the content (Monjurul & Latiful, 2012). Relly, Stafford, Williams and Corliss (2014) reported that PEG was exact and reliable at predicting human scores and had a scoring method that was fairly simple.

Content Rater (C-rater): The C-rater was developed by Educational Testing Service (ETS), it is an automated short-answer grading software that matches syntactic features of learners' responses (Siddiqi & Harrison, 2006). It was developed to measure learners' understanding of specific content of an instruction/course outline without regard for the writing skills of learners (Leacock, 2004). It is aimed at grading factual answers which makes it such that the number of possible correct responses from students is quite numerous. An ideal correct response would be formed by a "content expert". The main duty of the C-rater is to map student's responses to the ideal correct model but before this mapping takes place, student's responses are transformed to a *canonical representation* (a non-ambiguous, mutually exclusive representation of "knowledge"). At the point of generating this representations, variations in student's responses are taking into consideration so that it can be normalized (Siddiqi & Harrison, 2006).

The developers of C-raters divided variations into five classes and they are: *Syntactic variations* (for instance "The banking sector is dominated by males" and "Males dominate the banking sector"); *pronoun reference* (for instance "Folu withdrew the cash and spent **it**" and "She withdrew the cash and spent it all"); *Morphological variations* (for instance sleep, sleeps, slept; or song, songs, sung, sang) ; the use of *synonyms and similar words* (for instance slippery, greasy, oily, slick and many others); and *spelling/typographical errors* (the C-rater usually

corrects the error by itself). After the transformation of students' responses to a normalized canonical representation, the system then matches the ideal response to the canonical representation. At this point it would now award the assigned mark.

The C-rater is a *paraphrase recognizer* because it tries to identify an equivalent correct response of learners. It was designed to identify correct responses no matter the variations or paraphrasing associated with the words of expression. The software uses automated natural language processing. The c-rater model used *Alchemist* (the C-rater model-building interface); this model has over 300 million words current nonfictions, fictions and textbooks where it can select synonyms or similar words can be identified and used by the model to generate possible correct responses from learners' written answers (Leacock, 2004).

The C-rater was used by the National Assessment of Educational Progress (NAEP, 2002) Mathematics Online Project to appraise the written explanations behind solutions to some Mathematical problems. Five mathematical problems were selected, the mean of students' responses was around 15 words or 1.2 sentences long, each response was scored by 2 human markers, 250 to 300 students' responses were used for each of the five mathematical problems and the agreement percentage between the C-rater and the first human marker was 84.4% while with the second human marker was 83.6%. A value that is higher than 0.75 could be taken as a good representation of agreement beyond chance and/or error of occurrence (Fleiss, Levin, & Paik, 2003). This simply imply that C-rater's performance was not by chance or error of occurrence it is an excellent assessment and agreement between grades (Siddiqi & Harrison, 2006).

An empirical research by Leacock (2004) reported that the C-rater models was used for 19 out of the 24 open-ended short-answer questions, there was an agreement of 85% while there were two human graders and the agreement of grades was 93% between these two human graders. The kappa value which connotes the level of agreement base on chance and other

possible errors was 0.77 for C-rater to human graders. In this empirical research, five algebra questions were asked from learners but only four was graded using C-rater and about 170,000 students were marked in two days by C-rater and an agreement of 85.8% and 84.9% was recorded between C-rater and human grader 1 and 2 respectively which was a measurement from about 200 scripts while 91.5% (0.86 kappa value) was recorded between the two human graders.

In another research, the C-rater was used for online scoring and administration of Indiana's English; 11 End of Course Assessment Pilot Study. It was used to evaluate seven reading comprehension questions: Responses were open-ended and about 2.8 sentences or 43 words was the average responses. One hundred (100) students responded and a grade from one human rater as well as the software was compared; 84% was the agreement between the software and the human rater which reconfirm the excellent performance according to the criteria set by Fleiss, Levin and Paik (2003).

The C-rater has two main challenges: *misses* and *false positives*. The misses or miss refers to the failure of the software to identify correct concepts in an answer. Research by Leacock (2004) discovered that the rate of misses was far greater in the Indiana Pilot Study. Another deduction was that as a question get more open-ended, the percentage of misses increases. The false positives happens when the software allots too much credit to a response i.e credits are allotted for points/concepts that are not present in the response. False positives usually occur due to two possible reasons:

- a. The system has been aimed to detect correct ideas/concepts in students' responses but not to identify erroneous ideas/concepts. Therefore, there are instances where students have written correct ideas in the first statement of the answer but then later moved on

to writing entirely wrong concepts that actually nullifies the correct part of the answer.

C-rater will just give marks for the correct part without recognizing the incorrect idea.

- b. There are situations where students' response includes the correct language (words) but does not convey the correct concept/idea properly (Siddiqi & Harrison, 2006).

C-rater produces a fine-grained analysis of the logical relationship between the syntactic components of each sentence appearing in the response of a learner. The C-rater does not require a large collection of graded answers for training. Instead, it uses the single correct answer that has been provided in the instructor's guide as answer key because it is believed unrealistic to require extensive data collection for the purpose of grading low stake quizzes especially when given that a set of short questions is often provided at the end of chapters in a textbook. C-rater accomplished over 80% agreement with the grade assigned by an instructor (Burstein, Leacock, & Swartz, 2001)

Electronic Essay Rater (E-rater): E-rater has been in use by Educational Testing Service (ETS) since February 1999 with different versions which have been consistently updated from version 1.3 to version 2.0 produced with improved qualities and characteristics (Attali & Burstein, 2005). The E-rater was developed by Attali, Burstein, and others. E-rater is a computer program that employs Natural Language Processing (NLP) technology to identify well-defined features of a text/essay and associate them in a regression model to grade the text/essay. E-rater focuses on the well-defined features like grammar, usage and mechanics; style and elements of essay structure (Monjurul & Latiful, 2012).

The most recent version of E-rater uses 10 broad features extracted from the text using NLP techniques, eight of which represent the writing quality and two represent content features. E-rater uses a combination of statistical and NLP techniques to extract linguistic features from the essays to be graded. The operational difference between the operative system of the e-rater v. 1.3 and v. 2.0 lies in the scoring algorithm. The older version is trained on

sample of essays that have been previously marked by human readers, it uses over 50 features and then computed a stepwise linear regression that would produce a predicted value while newer version (v. 2.0) has changed with regard to the features set used in scoring, the model building approach and the final score assignment algorithm.

Essays are assessed against a benchmark set of human graded essays. With e-rater, an essay that stays on the topic of the question has a strong, coherent and well-organized argument structure, and displays a variety of word use and syntactic structure will attract a score at the higher end of a six-point scale. E-rater features include the analysis of the discourse structure of the syntactic structure and of the vocabulary usage (domain analysis). E-rater adopts a corpus-based approach to model building by using actual essay data to analyze the features of a sample of essay responses.

The application is designed to identify features in the text that reflect writing qualities specified in human reader scoring criteria and is currently composed by five main independent modules. Three of the modules identify features that may be used as scoring guide criteria for the syntactic variety, the organization of ideas and the vocabulary usage of an essay. A fourth independent module is used to select and weigh predictive features for essay scoring. Finally, the last module is used to compute the final score.

The Criterion an Online Essay Evaluation Service that is web-based, it was established by ETS to score students' essay writing instantly and give analytical feedback on the written essay (Burstein, Chodorow, & Leacock, 2003). The Criterion contains two softwares. The scoring software uses the E-rater engine while the second software is called *Critique* which consists of programs that score and make available responses for errors in grammar, usage and mechanics; identifies the structure and identifies the unwanted stylistic features. The critique is completely independent from the score generated by E-rater, thus providing additional feedback about qualities of writing related to topic and fluency only. E-rater is trained usually

on a collection of 270 essays that have been manually scored by trained human raters. E-rater is far more complex and requires more training than many other available systems. Furthermore, no on-line demonstration and no downloadable trial version of E-rater have been made available to the scientific community.

The E-rater measures some features of essays which are high-level properties of a text such as grammar, usage (errors), organization or content of specific topic. Each of these high-level features is broken down into a set of ground features; for instance grammar is sub-divided into features which count the number of auxiliary verbs, complement clauses, phrases etc. These feature are extracted from the essay using NLP tools which automatically assign part-of-speech tags to words and phrases, search for specific lexical items and others.

Many of the feature extractor of the e-rater v. 2.0 is written by hand and it's based on essay marking rubrics used as guides for human marking of essays for specific examinations. The resulting counts for each feature are accompanying with cells of a vector which encodes all the grammar features of a text. Similar vectors are constructed for the other high-level features. The feature extraction system outlined above and described in more detail in the references provided, allows any text to be represented as a set of sectors each representing a set of features of a given high-level type.

Each feature in each vector is weighted using a variety of techniques drawn from the fields of Information Retrieval (IR) and Machine Learning (ML). Monjurul and Latiful (2012) claimed that E-rater uses a vector-space model to measure semantic content. Vector-space model is originally developed for use IR, starts with a co-occurrence matrix where the rows and columns represent '*terms*' and '*documents*' respectively. *Terms* refers to any meaningful unit of information commonly used words and short phrases while *documents* refers to any unit of information containing terms such as sentences, articles, paragraphs or books. The cross path of a row to a column is called *cell*, in any cell a value may be represented by a simple

binary 1 or 0, signifying the existence or nonexistence of a term in an essay, or a natural number representing the rate of occurrence of a term in an essay.

Usually, Attali and Burstein (2006) reported that information-theoretical transformation helps to adjust the value in each cell. Such transformation is widely used in IR and weight terms so that they reflect the significance/rank of such terms within the whole essay. For example, one commonly widespread measure known as TF-IDF (term frequency-inverse document frequency) uses the following formula

$W_{ip} = tf_{ip} \log_2 \frac{N}{n}$ where W_{ip} is the weight of term i in document p , tf_{ip} is the frequency of term i in document p , N is the total number of documents (total number of essays in the training set), and n is the number of documents in which i occurs. After the weighting, document vectors are compared with each other using some mathematical measures of vector similarity, such as the cosine coefficient between the documents a and b uses the following formula

$$\text{Cos}(a,b) = \frac{\sum_i (a_i, b_i)}{|a| \cdot |b|}$$

where ‘a’ refers to document a which is the marking guide while ‘b’ refers to document b which is the students’ script, for automated assessment of the content of an unmarked essay, the weighted vector is computed by dropping the condition on p and the result is compared to the aggregated vectors for the marked training essays in each class using cosine similarity.

The unmarked essay is allotted a content score corresponding to the most similar class. This approach transforms an unsupervised weighting technique, which only requires an unannotated collection of essays or documents, into a supervised one which requires a set of manually marked prompt-specific essays. Other vectors are weighted in different ways depending on the type of features extracted. Counts of grammatical usage and style features are smoothed by adding 1 to all counts (avoiding zero counts for any feature), then divided by essay length word count to normalize for different essay lengths, and then changed to logs of

counts to avoid skewing results on the basis of abnormal high counts for a given feature. Rhetorical organization is computed by random indexing (Kanerva et. al., 2000) a modification of latent semantic indexing, which constructs word vectors based on co-occurrence in texts.

Words can be weighed using a wide variety of weight functions (Gorman & Curran, 2006). Burstein, Chodorow and Leacock (2003) describe an approach which calculates mean vectors for words from training essays which have been manually marked and segmented into passages performing different rhetorical functions. Mean vectors for each score point and passage type are normalized to unit length and converted so they lie on the origin of a graph of the converted geometric space. This controls for differing passage lengths and incorporates inverse document frequency into the word weights. The resulting passage vectors can now be used to compare the similarity of passages within and across essays and as above, to score essays for organization through similarity to mean vectors for manually-marked training passages.

The set of high-level feature scores obtained for a given essay are combined to give an overall score. In earlier versions of e-rater this was done by stepwise linear regression to assign optimal weights to the component scores so that the correlation with manually assigned overall scores on the training set was maximized. However, Attali and Burstein (2000) advocate a simpler and more perspicuous approach using the weighted average of standardized feature scores, where the weights can be set by expert examiners based on marking rubrics. Williamson (2009) in his description of e-rater implies a return to regression-based weighting.

There is 87% correctness between the grade that a human rater will grade an essay and the E-rater (Attali & Burstein, 2004). Attali and Burstein also reported that an E-rater does not identify certain words like grammar, humour or spelling errors and that it analyzes through paragraph changes and transitional phrases. It grades content by comparing an essay score to that of other students which implies that if any student writes in such an exceptionally unique

manner, such student may not score an exceptional grade. Wang and Brown (2008) conducted a research using e-rater which shows that there is a significant difference between human expert's grades and that of the e-rater and also no significant correlation among the e-rater and human grader's score. Relly, Stafford, Williams and Corliss (2014) stated that a total of 80% of students in a Chemistry course pilot study believed that the E-rater gave an accurate score of their performance.

The functioning system of E-rater (v.1.3) was trained using a selection of essay writings already scored by human graders. It values more than 50 structures/features in totality and then an analysis was made to pick features of significant impact on the prediction of the essay scores. The training resulted in a regression equation that was applied to the features of new essay written to a particular topic. In this research the E-rater version 2.0 (e-rater v.2.0) would be focused on since it is more recent compared to e-rater v.1.3; the e-rater v.2.0 has improved features with regards to scoring, model building and final score assignment algorithm.

In comparison, the C-rater adopts some the natural language processing tools and techniques developed for e-rater, even as the two systems differ in some important ways. E-rater assigns a score for writing skills rather than on specific content while C-rater is aimed to scoring a response as being either correct or incorrect. This goal is achieved by evaluating whether a response contains information related to specific concepts in the domain; if the response expresses these concepts, it is rated as correct but otherwise it is rated as improper without any regard for the writing skills. Also e-rater provides grades that are partly based on the rhetorical structure of an essay while C-rater needs to identify specific content.

Intelligent Essay Assessor: The Intelligent Essay Assessor (IEA) was developed in 1999 by Foltz, Streeter, Lochbaum and Landauer (Zhang, 2014) and afterwards was procured by Pearson Knowledge Technologies, it uses a method called Latent Semantic Analysis (LSA) to grade essays; the LSA is an approach created/generated by a psychologist Thomas Landauer

with the assistance of Peter Foltz and Darrell Laham (Dikli, 2006). The LSA is a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information. It was originally designed for indexing documents and text retrieval therefore the Latent Semantic Analysis (LSA) needs a smaller corpus of human-scored essays to train in any discipline before it starts to grade essays (Bennett, 2006). IEA being an essay grading and tutoring-feedback system has been used to grade essays and also reply learners in relation to the corrections and errors that is needed to be addressed (Landauer, Laham , & Foltz, 2003).

Landauer, Laham and Foltz (2000) stated that IEA simulate the meaning of words and passages, the vital idea is that the summation of all the situations in which a particular word appears offers a huge system of concurrent equation that defines the connotation of words to each other. The LSA denotes each word and passage as a fact of high dimensional *semantic space*. Using a matrix algebra technique known as Singular Value Decomposition (SVD), new relationships between words and documents are opened and existing relationships are modified to more accurately represent their true significance.

A matrix represents the words and their contexts. Each word being analyzed represents a row in the matrix, while each column signifies the sentences, paragraphs and other subdivisions of the context in which the word occurs. The cells of the matrix contain the frequencies of the words in each context, the initial matrix is then transformed according to an inverse document frequency weighting approach, a very well-known technique of the indexing and information retrieval domain. The SVD is then applied to the matrix to subdivide it into three component matrices that reproduce the original one if multiplied together, owing a reduced dimension of these three matrices in which the word-context associations can be represented, new relationships between words and contexts are induced when reconstructing

close approximations to the original matrix from the reduced dimension component SVD matrices (Whittington & Hunt, 1999).

These new relationships are made visible, whereas prior to the SVD they were hidden or latent. To grade an essay a matrix for the essay document is built, and then transformed by the SVD technique to approximately reproduce the matrix using the reduced dimensional matrices built for the essay topic domain semantic space. The semantic space typically consists of human graded essays. Cosine correlation is used to measure the similarity of the reduced dimensional space constructed from a model answer, such as an instructional text taken from a course text or an essay prepared by the class tutor, against a student essay. Latent Semantic Analysis (LSA) makes no use of word order since the authors asserted that this is not the most important factor for grabbing the sense of a passage. It also requires huge amounts of data to construct a suitable matrix representation of word use or occurrence, and due to the size of the matrices involved computations are very burdensome (Whittington & Hunt, 1999).

This approach used by LSA is usually more correct than mere traditional method that believes in occurrence of same words or word stems in two or more similar passages. IEA determines its grades predominantly from the opinion of human graders about similar write ups and it also constantly preserves the content knowledge while grading essays. The preservation of content knowledge is pretty fair for face validity though graders can hardly separate the mechanical qualities of an essay from the quality of the content (Reiner, Bothell, Sudweeks, & Wood, 2002). IEA practically measures content knowledge, mechanical components and style yet its closely equivalent to grade any essay even if it is not in Latin alphabets which the origin of the IEA.

The operational system of IEA is simply a training of the system with a large collection of relevant documents in the discipline of interest such as a Chemistry textbook when intending to grade Chemistry essays or real life-time test takers scripts when intending to grade a

narrative essay. It determines the overall grade, checks out error in language, identify plagiarism or any effort of the writer to play pranks on the system and further provide tutorial feedback. The difference between IEA and other AEG systems is that; others predominantly locate specific/unique characteristics of an essay as its correlates to human assessors score, then they forecast a formula that can be best of fitness for scoring every other essays but for IEA its LSA-based using direct human evaluation method related to content knowledge which is called *vicarious human scoring*.

The internet-based version of the IEA provides immediate appraisal and remarks which are possibly a teacher's or a grader's comment though it may not certainly be the same but still valued to be a good comment. For instance it can inform the student of missing gap in the content knowledge and refer them to significant points/ideas in their textbooks. It can also recognize unconnected/insignificant points as well as other organizational concepts. Despite the mentioned abilities and capacities of IEA it cannot inform a student of a whether an idea/concept was logically or convincingly presented or that a comma, semi-colon, and other punctuation marks that would have made a statement well presented.

IEA can be calibrated to score essays either base on a section of a textbook that students are expected to have studied or an ideal answer by a human grader to as few as 20 pre-scored scripts. Empirical research on a nerve conduction study was carried where three sets of graders were used; a professor, an undergraduate teaching assistant and a graduate-student teaching assistant, with the same training IEA agreed best with the professor's grade compared next to that of the graduate teaching assistant and least to the undergraduate teaching assistant (Landauer, Laham & Foltz, 2003).

Another study on heart anatomy and function did used IEA for grading short-answer test scores it produced a correlation of 0.76 to grades from two professional readers at 0.72 and

0.81 each. Landauer, Laham and Foltz, 2000 concluded that the total scores from IEA are as reliable to that of a teacher or an expert grader. Another viable use of IEA has being for grading of standardized tests like the General Management Admissions Test (GMAT) from 1999 to 2005 but since January 2006 by IntelliMetric. The standardized tests are graded by two human graders with a third grader been the IEA especially when the first two graders have diverge points or scores. The IEA system has resulted in an agreement between 85% to 91%.

Educational Testing Service I (ETS I): This system was developed by Burstein and Kaplan of the Educational Testing Service in the early 1990s, it works only on a sentence fragment of between 15 and 20 words (Whittington & Hunt, 1999). The technique uses lexical-semantic techniques to build a scoring system, based on small data sets. It uses a domain- specific, concept-based lexicon and a concept grammar, all built-in from the training data. The training data essays are parsed by Microsoft Natural Language Processing (MsNLP) tool, any suffixes are removed by hand, and a list of stop words is also removed. This produces a *lexicon*.

The list of words and terms in the lexicon remains constant while the features associated with each entry are modular, so it can be substituted as necessary. Some manual classification is necessary, such as specification of some words as metonyms (*a word, name, or expression used as a substitute for something else with which it is closely associated*) of each other and so on. Grammar rules are then constructed, again manually, for each category of answer, each category should contain all the paraphrases for that possible answer using syntactic parses of sentences from the training data along with the lexicon.

New essays are then parsed by the phrasal node extraction program that outputs the sentences' noun, verb, adjective, adverb and preposition phrases, the system does not make use of specific parts of speech at this stage so they are collapsed into a generic XP phrase type and the sentence, or what is left of it now is checked for matches against the grammar and the lexicon. The XP phrase type is taken from X-bar syntax which attempts to model common

properties between the different syntactic components of noun, verb, adjective and preposition phrase etc. Instead of building a different grammar rule for each different type of phrase, the X-bar syntax generalizes out to a single rule that applies universally to all. The system involves lots of pre-processing and much of it is manual, although the developers of the software argue that the cost in relation to time spent to manually put in syntax components still worth it in all.

Using ETS I, the developers of the system claimed an accuracy of 80% score when compared/related to grades from human expert, when marking the essay test set and 90% when marking both the training and test set at the same time i.e. using the a part of the set of test to train the system on a set of essays to be marked and then including it in the test set for the marking part as well. The most common source for errors is due to gaps in the lexicon, i.e. to words not manually identified as metonyms. An empirical research was carried out, the developers adopted an augmented lexicon. Constructing this new lexicon involved examining the test set as well as training set to manually place metonyms. This achieved accuracy of 93% when marking on the test set and of 96% when marking both the training and test sets (Burstein, Chodorow & Leacock, 2003).

Automark: Automark was developed by Burstein and Kaplan of the Educational Testing service in the early nineteen nineties. Automark is a software system developed in the quest for a more encompassing electronic marking of free-text answer to open-ended items. It uses the Information Extraction (IE) and Natural Language Programming (NLP) method for grading open-ended responses (Mitchell, Russel, Broomhead & Aldridge, 2002). This software was developed such that it identifies acceptable and improper answers to question responses. The system was constructed to understand the content of a free response without disproportionately penalizing the student for blunders in grammar, spelling or semantics.

Mitchell, et. al. (2002) carried out an empirical research to test the performance of Automark which was done in alliance with the Centre for Research in Primary Science and

Technology for pupils of age 11. A purposive sample of this population was selected to evaluate the effectiveness of the system since spelling, syntax and semantic errors are usual, it was pertinent to carry out the empirical research in order to improve on Automark. This demanding but fact-finding assessment of pupils was chosen to enhance robustness of the software for real-world application. Two experiments were carried out namely the Blind and the Moderation experiments and they resulted in 93% and 96% correlation to the grades awarded by human experts. In Brunel University, London engineering students in their first year are administered test on Java using Automark which is a practice in a number of higher education institutions.

Automark uses about 400 manually marked essays to train its marking algorithm in order to grade more accurately. It has the strength to search for definite content within free-text answers, the content must have being itemized in the form of a number of marking scheme templates. Each template represents one form of a valid or an invalid answer (Valenti, Neri, & Cucchiarelli, 2003). The marking scheme template is robust enough to cater for multiple entry of the same statement or content.

The grading process advances through a number of phases. First, the incoming text is pre-processed to normalize the input in terms of spelling and punctuation. Then, a sentence analyzer identifies the core syntactic elements of the text and how they are connected. The pattern-mulching component searches for connections between the marking scheme templates and the syntactic elements in the student text. Finally, the feedback component processes the result of the pattern match. Feedback is typically provided as a mark, but more specific feedback is claimed to be possible.

IntelliMetric: According to Monjurul and Latiful (2012) IntelliMetric was developed by Shermis and others in the year 2003. It uses a blend of Artificial Intelligence (AI), statistical technologies and Natural Language Processing (NLP). An additional software/setup called CogniSearch was developed for use in conjunction with IntelliMetric to comprehend natural

language as well support essay scoring. Byrne, Tang, Tranduc, & Tang (2008) wrote that IntelliMetric uses Bayesian computer adaptive testing to organize essay structures into three or four point definite scale.

Its utmost invention is MyAccess, an automated writing tool that helps to advance student writing and prepare them for the essay part of an exam, such as the Graduate Management Admission Test (GMAT) for pre-requisite of the MBA programs into the business schools. IntelliMetric regularly demands a set of training and rehearsal with a set of essays that have been scored before especially determined by human expert raters in such field of knowledge/discipline. The IntelliMetric uses a couple of steps to carry out its grading process.

Attali and Burstein (2004) reported firstly that the system adopts the known scores in the training essays. Secondly, the system is made to score a set of smaller essays with known scores from human experts for validation purpose and finally, ones the system scores the model essays as desired, new essays with unknown scores are applied. An average Pearson correlation coefficient of 0.83 has been discovered as the score given by IntelliMetric when compared to expert human rater (Duwairi, 2006).

Schema, Extract, Analyze and Report (SEAR): Christie (1999) developed a software as a result of his ongoing PhD research work. Christie believed that automated assessment of essays requires grading of style and content (where appropriate). According to him the technique for assessing style is based on a set of mutual metrics while content looks into usage and coverage. SEAR for grading content develops a grading content schema which neither requires *training* nor *calibration*. The content schema can be prepared once and then reviewed more easily and quickly. SEAR has not be made available for the

MARKIT: Markit is another AEG system, it was developed by Robert Williams and Heinz Dreher of Curtin University, Australia at about year 2002. Markit uses a combination of NLP

and a pattern matching technique. It has eight subsystems written in Visual Basic for Applications, Java and C++. It was discovered to have a unique advantage over every other commercially available grading softwares which is that it needs only one model answer against which every other answers are to be compared and graded. Markit relies on commonly available technology as compared to specific computing platforms. The computational algorithm that was designed into the Markit such that it can be shrink-wrapped for widespread distribution and local use.

A research was carried out on essays of second year law students from the Curtin Business School in Australia, a total of 66 essays were made available while a randomly sampled 20 essays was finally involved in the research. In the grades awarded by the human experts and by Markit there were cases where it was tangent to each other while there were extreme cases of 16 marks difference either higher or lower to the human grader. Further investigations into these cases reflected that error was due from the human grader rather than the system since the model answer that the system employed to mark was supplied by the same human grader. A correlation of 0.79 was discovered between the human graders scores and Markit. Another comparison of IEA and Markit was done using 9 samples and a Pearson correlation of 0.84 at 1% significant level of a two-tailed test was discovered.

Expert System for Essay Scoring (ES4ES): The ES4ES was developed by Ade-Ibijola Abejide Olu and Wakama Ibiba of Federal University of Technology, Ondo State and Amadi Juliet Chioma of Federal Polytechnic, Idah, Kogi State in the year 2012. It is an expert system for grading open ended items. It uses the Information Extraction techniques. It was designed based on three components which are the Knowledge base, the Inference Engine and the working memory.

Ade-Ibijola, Wakama and Amadi (2012) conducted a research with a sample of 50 items. 10 open-ended items with free responses/answers by 5 students on “Software

Engineering”. A correlation of 0.71 was obtained between ES4ES and a human expert while a comparative analysis of the ES4ES with existing AEG systems like C-rater, E-rater, SEAR, PEG, IntelliMetric and Automark was also conducted and a correlate of 0.83, 0.97, 0.45, 0.87, 0.98, and 0.95 respectively.

Automated Assessment and Challenges of Automated Essay Grading Systems

Automated assessment is the use of computer its components and capabilities to automatically analyze and assess learner's competence. The field of automated assessment can be traced back to the early 1960s and then it was developed as a means to overcome matters arising with standardized assessment. According to Winter, Burkhardt, Freidhoff, Stimson and Leslie (2013) Ellis Batten Page in 1966, who happens to be the father of automated essay grading concluded that AEG is an inevitable development that would have profound educational impact.

The first successful automated essay scoring engine was “Project Essay Grade” (PEG) which was promoted or spread by Ellis Batten Page and colleagues from the University of Connecticut in 1973. This development was first predicted to be developed in an article titled “the Imminence of Grading Essays by Computer” in 1966. Page being an high school English teacher saw the need for this automated grading, claiming that students will gain more as quick feedback of their writings will be easier and also will help reduce the load of papers to be reviewed and marked by a teacher. Though PEG made striking outcomes, the existing technical know-how then was primeval to make it a real wide spread activity because students still have to type using the punch-card and read into the mainframe computer before the evaluation could be done. As a result, this advancement was inactive till 1990 and was awakened with the invention of micro-computer and the internet (Winter, et. al., 2013).

Automated Grading Systems (AGS) have been valuable in several academic endeavour from aiding to grade multiple choice items and short-answer text recently AGS is seen to be moving towards essay/prose writing or theory. Essays are important testing tools for measuring academic achievement because it requires incorporation of ideas and the capacity to recall on the part of the learner but they are time overwhelming to grade manually (Mahana, Johns, & Apte, 2012). They further believed that manual grading of essays takes up a noteworthy volume of teachers' time and therefore it is an expensive procedure. They added that if AGS is verified to equal or surpass the reliability of human graders'; automated grading systems would help restore a level of sanity and reduction in stress that teachers' actually have to pass through in grading essay writings.

Automated essay scoring is a measurement technology in which computers evaluate written responses by examinees (Shermis & Burstein, 2003). According to Shermis, Raymat and Berrera (2003) Automated Essay Scoring (AES) is defined as the computer technology that evaluates and scores written prose. AES is the field of automatically assigning numbers or marks to examinees' essay. According to Fluck, Pullen and Harper (2009), automated essay grader systems should match human graders' reliably since this systyem is used in some high-stake examinations as second grader and as the only grading tool in some low stakes evaluation/examination.

Recently, awareness on development and the use of Computer-based Assessment Systems (CbAS) has grown exponentially in the last few years and it is likely linked to increase in the number of students/learners and the possibilities provided by distance learning. As online teaching and learning become prevalent, the emergence of ways and methods of assessing learners who are involve in online programs turns out to be necessary (Olufisoye, 2012). He further stated that the benefits of automated grading includes students' motivation, immediate feedback and assessing lager classes effectively.

The drive for developing computer supported techniques in assessment and grading of learners is not farfetched from the need for speed, accuracy, reliability and effectiveness (Williams & Heinz, 2006). The concern for development in the use of Computer Based Assessment Systems (CbAS) has also become salient due to large number of learners attending universities and those in the quest for education at different levels, leading to increase in teacher-student ratio, the opportunities provided by e-learning and the fact that computers have the capacity to bring the teacher closer to a students at the convenience of the learner has being made easy with the use of CbAS.

According to Gibbs (2010) what influence students learning majorly is not only teaching method or strategy but the assessment also. Learners in this report submitted that all aspects of their study pattern including what topic they focus on, number of hours they allot to reading, their study habits and study behaviour are influenced by the way they perceive the demands of the assessment system. The assessment procedure of an educational setting speaks volume about the reality of that institution. Assessment is of great importance to what learners do in the learning process. It further displays the measure of how good learners have spent their study time. To discover the reality about any educational process, the assessment techniques should not be undermined (McCulloch, 2007). McCulloch further stated that a learner can escape the challenges imposed by a poor teacher or teaching process, though with difficulty, but cannot escape the effects of poor assessment if they would graduate from school successfully.

The need for assessment differ from participants to participants like students, teachers and the society at large. To a student the need for assessment is to provide feedback, encourage furtherance of knowledge, analyze strengths and weaknesses and stimulate the desire to progress in knowledge. For teachers the need for assessment is to find the level of achievement (summative), permit or deny a student of promotion to the next level/stage, decide the measure

to which a course outline has been attained, for authorization to practice and to envisage/forecast success in the future and employment. For the society, assessment is needed to observe standards, to regulate perhaps it is good enough for a learner to practice what as being learnt over time and mark the development as well as progression of societal demands or expectation.

Assessment in colleges as it relates to learning evolved through four periods; standardized test of learning originated around 1900-1933, general and graduate education assessment of learning between 1934 and 1947, the rise of examining bodies came in from 1947-1978 and the times of external accountability began 1979-till date (Richard, 2007). Assessment makes teaching to have focus or goal-oriented which is why some educators teach to test. Electronic supported assessment or E-assessment is a field of growing importance but it is yet to have a significant influence in the educational sector especially in the higher education sector (Fluck, Pullen, & Harper, 2009).

AEG is a model that focuses on not only assessment of knowledge unlike the multiple choice but a deeper understanding of the learning content. Fluck, Pullen and Harper (2009) used students' personal computers to take eExamination using a customized version of an open source live CD under the supervision of invigilators this was done in their research because students reflected that the noise of keyboard during examinations has been disturbing in their experience. The CD switched/booted the computer using the Ubuntu-Linux operating system, in the examination access to internet or any other resources was not allowed then work done was later saved on the remote computer of each student.

In grading essays, subjectivity or at least the perceived subjectivity of the grading process is a major challenge. The subjective nature of essay assessment leads to differences in grades awarded by different human assessors which is professed by students as great source of

unfairness (Mark, Jill, Derrick, & Klaus, 2011). They also reported that about 30% of teachers' time in Great Britain is devoted to marking, so to free up that 30% which worth about 3 billion UK pounds/year, then an active way that teachers' will trust to mark essays and short text responses must be found. This issue may be solved through the approval of automated assessment tools for essays. A system for automated assessment would at least be consistent in the way it scores essays as huge cost and time can be saved if the AEG can be shown to grade essays within the range of those scores given by human assessors.

According to Hagan, Coleman and O'Connor (2014) using computers will provide a number of paybacks to the educational community. In fact, it may further help to develop more active instructional materials for improving reading, writing and other communication abilities. Although it is fairly easy to construct a model that assesses closed-class types of questions quite accurately, automated assessment faces many challenges when it is meant to focus on grading essay. One of the most crucial considerations is the possibility of building a system that matches human behaviour in reading and making value judgments about someone's writing. This is largely dictated by the ability to assess not only vocabulary, grammar and syntax but also various other aspects; different writing genres, as well as cognitive aspects such as language maturity, intellectual content, the logic behind an argument, discourse structure, clarity and fluency are only a small part of the spectrum that needs to be considered.

Generally, automated assessment is chosen over manual assessment to avoid bias errors, human errors and also foster conservation of teachers' time (Shermis, Burstein, Higgins, & Zechner, 2010). With the advent of new technologies, Automated Essay Scoring (AES) support teachers' classroom assessment and help overcome time, cost, reliability and generalizability issues in written/essay assessment. Research on AES has revealed that computers function as a more effective cognitive tool (Attali & Burstein, 2004). Automated essay scoring uses computer algorithms designed to emulate human scoring. This is achieved

by extracting features from essays and then using machine learning and modeling to establish a correspondence between these features and essays are scored based on a sample of essays that have been scored by human markers. These processes are related to mathematical computations and it further requires replication to realize the most optimal solutions for scoring of essays.

Consequently, automated scoring systems require training primarily through the use of human-marked scripts before they can be used to mark student essays in a test administration. Previous research studies have demonstrated that a high score agreement rate could be achieved between human raters and automated scoring systems (Ade-Ibijola, Wakama, & Amadi, 2012). The automated assessment has been used on closed form items (Multiple Choice Questions) which has further made it possible for a learner with poor content knowledge of a course to pass simply by guessing that is the cognitive knowledge may not be properly understood by a learner yet a fairly good grade can be achieved by the learner just by mere guess of the correct option in a multiple choice item format.

Challenges of Automated Essay Grading: AEG system is not perfect in itself, more so that the scientific community is yet to generally accept the reality and or possibility of using computers to grade learner's essay. Therefore, challenges of AEG (Automated Essay Grader) are not far-fetched from the characteristics of paper-and-pencil test or factors that possibly affect or influence learners and the examiner when involved in a paper-and-pencil test. The existence of any challenge is not enough reason for closing the possibility as well as efficacy of computers in being used to grade essays. The knowledge of the fact that AEG systems still permits development and add-ins to make up for shortcomings of any AEG software till it seemingly becomes a more robust and efficient system.

Strauss (2013) raised some challenges present in existing AEG software as: learners' responses to essays or short answers that may need or demand diagrams or graphs would be

difficult to either grade automatically or even draw for learners even when recognition facilities are available for use to both the examinee and the examiner. Mahana, Johns and Apte (2012) further submitted that some AEG softwares don't allow answers to be altered before submission is been made as well as go back to a previously unanswered and or unsatisfactorily answered to the examinee/learner. AEG softwares should also be made in such a way that it would permit learners'/students' to access past test items in order to aid review and good practice for learners who intend to do so. Though restrictions may be on the level of accessibility of a learner to past tests just as a way of item banking.

Williams and Dreher (2004) also claimed that another limitation to AEG systems is that since model answers contain only a set body of knowledge, the system will only grade other answers regardless of how brilliant and exceptionally good they may be poorly since AEG software would basically be interested in locating the association between the response of the examinee and the model answer that was supplied from the onset.

Expert System for Essay Scoring (ES4ES) was developed by Ade-Ibijola, Wakama and Amadi of the Department of Computer Science, Federal University of Technology, Akure (Ondo State) and Federal Polytechnic, Idah (Kogi State). They reported that ES4ES is not appropriate to grade free-text answers or an essay where the order of arrangement of words are important and this is also a common feature of IEA (Intelligent Essay Assessor). Therefore, these grading softwares are often more operative in short-text answers than bulky texts of over 500 words. AEG systems cannot judge logic, humour, satire and some other aspect of literature or writing that are a bit technical in language. However, none of these shortfalls affects the reliability of the software (Ade-Ibijola, Wakama, & Amadi, 2012).

Strauss (2013) reported a software was developed by Edx (non-profit organization) jointly funded by Harvard University and the Massachusetts Institute of Technology; on the ability of computers in grading essay tests. Strauss (2013) claimed that since the software uses

artificial intelligence in grading learners essay as well as short answers; does that really imply that computers can capture the depth, beauty, structure, relevance, reasoning, convincing arguments, meaningful organization, clarity and creativity of an essay. More so, writing been an highly demanding act should not be compromised on the altar of artificial intelligence marking since creativity is an integral part of essay writing. Research has shown that high-stake writing tests alter the conditions of writing essays by negating students/learners opportunity to think, read, dialogue with others, address tangible listeners, cultivate concepts and review their developing writing skill.

Strauss (2013) further pointed out certain flaws of using automated essay grading as the following: first, computers don't have the ability to identify features that are salient to good writings like reasoning, clearness, precision, concepts connected to a topic, inventive styles and types of argument worth of substantiation and comedy or sarcasm. Second, some of the methods used by computers to grade students' writings are more unpolished methods when compared to human readers. For example, some computers measure the complexity of a terminology by calculating the average length of words and how often the words are used. Third, teachers' become more redundant as creative and inventive ways of grading essays would not be developed. Four, grading by computer favours surface features of writing like grammar, spelling and punctuations; problems in this area depends on the availability of time for the student to reread and edit. Also, students that are less conversant with the use of computer technology may be discriminated against, which implies that learners who are less skillful in computer are disadvantaged.

Some existing systems generally have some short-comings or restrictions some of which are mentioned by Fluck, Pullen and Harper (2009). They reported that AEG software are majorly applicable to a particular discipline/knowledge or domain. For example Pearson VUE for the UK Driving Standards Agency's that is they are developed to solve majorly grade

specific discipline and also attend to secluded knowledge. Second, it demands the examiners know-how on a particular question setting software application or software for instance Articulate *Quizmaker 2.0*, which implies that the examiner needs special skill or knowledge to prepare the question that would be displayed to the examiners. Also, internet connectivity is required during examination which implies that examinees may have access to other materials except the blocking of other software.

Some weaknesses as analysed by Dikli (2006) is that computers can not do what human raters will do for the reason that the computer would do only “what it is automated to do” and it wouldn’t “increase the value” of an essay. Another criticism is the construct objections that is computer counts variables that might not be “truly” important in an essay grading. For instance it may focus on strict aspects rather than structural ones. According to Shermis, Burstein, Higgins and Zechner, (2010) computers do not “appreciate” the written text being appraised. For instance, the computer would not “get” the following joke.

Q-When is a cup not a cup?

A-When it is been turned upside down.

Unlike humans, a computer cannot understand play on words, and deduce that the predicate in the answer.

Automated Grading of Objective Tests

Large-scale testing programs involving millions of students’ at all educational levels are now in use by Universities, corporate organizations (to conduct job test), driving license tests (before issuing the drivers’ license) and examination bodies (Kumaran & Sankar, 2015). Commonly used in our society in automated assessment is the objective type (multiple choice), examination boards like Joint Admissions and Matriculation Board (JAMB) conduct their Unified Tertiary Matriculation Examination-UTME using this method in their test administration. Usually each item consist of one question (stem) and each question has different predefined answers (options) to choose from which has one correct answer (key) and

other options as distracters. Candidates are only expected to check one of the options provided; which will serve as their preferred choice (answer) on the basis of which scores would be awarded. All that is required of the computer is to take a count of the correct options that has been checked by the candidate which is a lot easier when compared to a cloze.

A cloze or short-answer or fill in the gap is a type of test where some words are substituted by blanks. These blanks are to be filled in by the examinee with a word to complete the sentence. Usually, there is a set of words from which the examinee can choose one that they wish to fill into the blank as the key (correct option) where other options are regarded as distracters (incorrect options). Since the correct word to fill into each blank is predefined, it is also perceived to be easier to grade by the computer automatically though a cloze can be done where the examinee is not given predefined options, they are expected to supply the appropriate word or answer. These forms of assessment provide only restricted insight into students' knowledge (Kumaran & Sankar, 2015). The flaw with predefined answers is that it cannot be determined if the student chose the correct answer based on his knowledge or it was just from a random guess. To comprehend the difference between multiple-choice and short answer items is not too difficult but establishing the difference between short answer and essay may not be as straight-forward. According to Burrows, Gurevych and Stein (2015), there are five distinctive measures in an essay:

- a. A recall from external knowledge is required instead of acknowledging or picking the answer from the options provided,
- b. the response must necessarily be in the form of natural language,
- c. the response must necessarily be in one phrase or paragraph length,
- d. to grade response, focus must be on the content not the writing style and
- e. restriction is placed on the level of openness of responses.

Grading multiple-choice items automatically seems more acceptable since there is only one correct answer known as *key* unlike short answer form where correctness and understanding of natural language comes to play. There are likely to be more than one correct phrase or sentence to a question since synonymous words and sentence structures can differ from one learner to another.

In University of Iowa, a computer based assessment tool was developed by Lindquist. This tool is an optical test scoring equipment (Alber & Debiasi, 2013). The program was efficient and effective through the use of computer and scanning technology. This equipment however is only suitable for True-False and multiple choice items, commonly known as Objective tests. Objective tests can measure many learning outcomes. However, there remain some significant instructional outcomes for which no satisfactory objective measurements have been developed. These include such outcomes as the ability to recall, organize, and integrate ideas; the ability to express oneself in writing and the ability to supply rather than merely identify interpretations and applications of data. An objective test or item can be considered easy to measure using computational methods since there is only a single correct response to each item.

In the field of education, summative or formative assessments are often categorized as either objective or subjective assessment. Objective assessment is a form of questioning that has a single correct answer. Subjective assessment is a form of questioning that have more than one correct answer or more than one way of expressing the correct answer. There are various types of objective and subjective questions. Objective questions include true/false, multiple choice, multiple response and matching questions. Objective assessment is well fit for the increasingly popular computerized or online assessment formats. Subjective questions include extended-response questions. However, concern has been expressed that such formats may not

always be assessing what the teacher believes they are, partly because they require the recognition of the answer, rather than the construction of a response (Nicol, 2007).

According to Valenti, Neri and Cucchiarelli (2003), more than forty commercial Computer-based Assessment Systems (CbAS) are available in the market. Most of those tools are based on the use of the objective-type questions: multiple choice, short answer, selection/association, hot spot and visual identification. Researchers agreed that some aspects of educational content and learning outcomes are complex to measure using objective-type questions. Learning outcomes implying the ability to recall, organize and integrate ideas, the ability to express oneself in writing, ability to interpret and apply knowledge all these and require less structuring of response that objective test items are characterized by (Kumaran & Sankar, 2015). It is in the measurement of such outcomes corresponding to the higher levels of the Bloom's taxonomy namely evaluating and creating that the essay question serves its most useful purpose.

Many e-assessment tools for short-answers have significantly spread around over the years, but the e-rater seems to be more commonly known and used (Kumaran & Sankar, 2015), the e-rater engages the holistic grading approach. This system has been able to record over 97% agreement to human reader scoring, though the e-rater reflects the writing style than the content of the essay, then it should rather be used to grade essay compared to grading short-answer items. Siddiqi and Harrison (2006) emphasised that in order to grade short answer items, content is more salient than the writing style. The Latent Semantic Analysis (LSA) is a different method that has been used for essay grading which has also obtained high agreement with human graders. LSA focuses on metrics which relate more to writing style and measures of vocabulary usage.

Mohler and Michalcea (2009) explored a variety of techniques for automatic grading of short answer. Grades were allotted using a measure of text-to-text resemblance between

students' answer and model answer. An approach is to use Machine Learning (ML) algorithm where grading is based on features like the use of synonyms, text length, word ordering and a pair of nouns appearing within five words. Another method searches the effectiveness of relating WordNet-based on word-to-word resemblance measures.

A software developed by Intelligent Assessment Technologies (IAT) was used by the Open University in Nigeria which has the strength of enabling a question author with no knowledge of natural language processing (NLP) to use the software. Contemporary developments in NLP and ML have encouraged researchers to employ these methods in automated grading of short answers. ML method can assess texts without having to understand student answers which implies that there is no need for understanding context and semantics though it can identify complex arrangements in students' answers.

Alotaibi and Mirza (2012) used an hybrid technique that combined Information Extraction (IE), Decision Tree Learning and Machine Learning (ML) to grade short answers and the effectiveness of this technique has not been discussed/researched. AutoMark is another automated grading software which was developed by the Intelligent Assessment Technologies, it uses the IE to grade short answers. AutoMark harmonizes learner's response with a stated pattern that evaluates the responses/answers of the learner. It requires manually produced pattern for each correct response. The ontology mapping technique was used to also grade short answer responses; the ontology mapping concentrates on validation of learning objects, instructional processes and the design of learning (Kumaran & Sankar, 2015). Some other possible techniques of automated grading uses the corpus-based and the knowledge-based techniques, with the best outcomes being attained from the corpus-based measuring technique using the Wikipedia united with a "relevant feedback" technique.

Rationale for Automated Essay Grading

‘Essay’ also known as ‘constructed-response’ or ‘free-response writing’ involves learners writing their own answers. It allows learners to directly provide and express concepts in their own words. By doing this, learners’ exhibit a range of characteristics/skills: explaining personal reactions, describing interpretations, generating conclusions, or supporting critical evaluations. However the grading of these essay is demanding and prone to measurement error due to discrepancies/differences in grading processes.

Warschauer and Grimes (2008) claimed that there is a consensus in educational research that learners need more writing exercises to enable them improve on needed skills in expression. However, teachers have inadequate time to score their scripts. Automated Essay Grading (AEG) also known as computerized essay scoring uses what is known as *artificial intelligence* to mark and reply to essays easily and almost immediately, it can vividly simplify this burden/need on teachers thereby tolerating more writing rehearsal and quicker improvement on the part of the learner.

Shermis (2014) emphasized that AEG holds the potential of playing a very viable role in high-stakes writing assessment. Klobucar, et al., (2012) studied Automated Essay Scoring to determine its validity in writing assessment. They focused on performance of about 1,006 first year students combining a two AEG writings. Firstly at the beginning of the semester and the other at the end of the semester allowing 48-hour completion, their findings showed that AEG is compatible in a digitalized educational environment. They also found that AEG represents of a promising stage of educational development as it applies to teachers and students’ involvement. This hence implies that learners’ no matter their numbers can be gainfully engaged effectively and efficiently in writing essay responses to a learning construct regardless of fewer teacher or teaching time period that is available to learners .

Essay writing requires expression of knowledge and information in such a way that reflects understanding and other factors like mannerism and style of the writer/learner thus the challenge of grading essay to a teacher comes in the ability of the teacher to read and decode facts from illustration, styles and other forms of distractions in the expression. Another difficulty is the grading bias that can be involved especially when human raters do the grading. Many researchers claim that the subjective nature of essay grading leads to differences in grades awarded by varied human raters, which is seeming to students as a great source of inequitableness (Valenti, Neri, & Cucchiarelli, 2003). However, Dikli (2006) concluded that AEG being a developing technology has proven to be needed and relevant as it has been used to overcome time, cost and generalizability problems in written assessment.

Dikli (2006) reported that machine scoring technologies can also help to increase the practicality of administering and evaluating large-scale assessments in essay format. The use of human raters on a large scale assessment of essay writings has been adjudged to be expensive and time consuming. Apart from the fact that AEG helps to save time and cost it has also being discovered to have a variety of feedback on the learner (Shermis, 2014). It has a great benefit over the teachers' manual work, during the process of automatic scoring, computers do not give grades to students by subjective judgment, and it can also do this job more efficiently. No matter how long the time will last, the computer won't get tired and make biased correction results.

Dikli (2006) further stated that the use of AEG technology, facilitates the performance of recurring activities without monotony and variation, adaptability (within preprogrammed pathways), flexibility (analysis/assessment can be carried out at any time, for a range of purposes, and on any number of candidates), and the ability to make decisions without being judgmental (in the sense of being biased). Mitchell, Aldridge and Broomhead (2014) also added that the weight of marking is eliminated from teachers because in previous times using paper-

pencil method, the grading input of teachers for 800 students is summed up to an estimation of 30-man days work load while on a 2.4GHz computer operating a Windows XP, it will take just about $6\frac{1}{2}$ hours (nearly 30 seconds per script of 270 questions). Rigidity is also removed as learners who are unavoidably absent on the day of the examination can still retake the examination without encumbering staff. Responses can be scanned through quickly, so that adjustments can be made to defective items such that it can be used in future examinations which would in turn increase confidence in AEG.

Computerized feedback has been researched in studies as an alternative for enhancing the effectiveness of feedback. Researchers have found difficulties with the quality of feedback given by teachers; because of lack of time and large classes, teachers sometimes fail to give timely and precise feedback, in spite of the ample positive effects of feedback, these issues can critically and seriously limit the benefits of feedback. Understanding this problem, researchers and educators began to pay serious attention to the automated essay scoring system because of its potential as a mechanism for consistent and prompt feedback and essay grading.

Also, Ben-Simon and Bennett (2005) wrote that automated essay scoring has the potential to decrease processing cost, speed up the reporting of results and improve the consistency of grading. As they stated the National Commission has recognized the potential value of this technology on writing in American schools and colleges and have recommended research and development of AEG system for; standardized tests. Automated Essay Scoring (AES) offers increasing as well as scoring consistency, introducing varied high-stakes assessments, reducing processing time and keeping the meaning of standardization by applying the same criteria to all the answers written by learners in a test.

Several studies has been carried out to demonstrate the accuracy and reliability of the AEG systems with respect to the writing assessment and the agreement rate between human

raters and AEG systems have been discovered to be high (Attali & Burstein, 2004; Dikli, 2006). Automated grading is quite reasonable for short answer questions where key words are sought in the response such as register and glossary.

Validity and Reliability: Essay Test and Automated Essay Grading

The use of automated scoring must be validated and confirmed (Klobucar, et al., 2012). Grading is expected to be based on effective model and any unfamiliar answers from a student should be scored correctly. There is the need for an agreement between the automated essay grading system and the key models/structures that are meant for the grading. This assessment benchmarks are meant for building structures that are used in validating assigned grades in examinations like Educational Testing Service (ETS): weight and representation of construct, correlation with experts score, association with other autonomous variables of interest to the content, impartiality in grades across subgroups and results/consequences of engaging automated essay grader in any circumstance. Klobucar, et al., (2012) stated that the e-rater extracts certain structures that uses regression-based features used in many grading models therefore the need to validate these structures and processes of generating the regression line models that has been generated in order to score an essay.

Klobucar, et al., (2012) reported length of *text* and average length of *word* as important aspects of validity and reliability to grading an essay. They justified their claim saying that since a typical essay write up takes place in a classroom setting with time duration allocated (two hours at times), students' ability is directly proportional to the length of the text, though they suggested that the length of an essay may not necessarily be a consideration in scoring an essay, yet it is not out of place since other valid variables/factors correlate with length and indirectly relate to the essay score. The *average length of word* meaning the mean number of letters per word as a correlate to grade but quite faintly correlate with the length of word. Moreover, this correlates strongly with the parts of speech.

Russell, Goldberg and O'Connor (2003) raised several concerns connected to validity of test scores using computer and it was examined. They discussed the evolvement of computer validity, then psychologists perceived computerized assessments as a method for regulating test variability, eradicating assessor's bias and increasing productivity by freeing them from monotonous function of test administration and scoring. Some comparative studies compared the conventional and the automated grading format for a diversity of test including intelligent test, personality assessment tests, vocational interest inventory and cognitive capacity test in this comparative researches test equivalence was established where it was concluded that both groups do not differ significantly between modes. Though, there were studies that their result revealed that the conventional paper-pencil test score was significantly higher than the mean of automated grading score especially in arithmetic reasoning, they emphasized that this was due to 'rough work space' for computation during the examination (Russell, Goldberg, & O'Connor, 2003). They additionally stressed that when students or examinees don't have access to review and revise what has been done; this factor also affects performance during automated grading especially the objective system therefore they recommended that only softwares that allow all the amenities of paper-pencil tests.

Mahana, Johns and Apte (2012) in their research work titled Automated essay grading using machine learning used 8 sets of essays as data, which was written by grade 7 to grade 10 students in an ASCII text format, a total of 12,978 essays. Each essay has one or more human expert grader's score. Each essay is about 150 to 500 words in length. They selected features that may stand as proxies as what human graders are likely to be looking for before awarding grade/mark for an essay. A linear regression model was then predicted based on the parameters selected. The scores generated by the system were then compared to those of human graders to derive an error metric. They concluded on an average Kappa value of 0.73 (Quadratic

Weighted Kappa) as the relationship between the automated essay grader's score and the score by the human graders.

Mahana et. al. (2012) extracted five parameters or proxies in order to train the system as well as help in grading the essays. These parameters are: phrases or clauses otherwise referred to as 'Bag of Words'; total word count per essay and or mean word length per essay; part of speech count e.g count of nouns, adjectives, adverbs and verbs; counting of words wrongly spelt; and counting of sentences that have been correctly punctuated. They added that there is no perfect system quoted and unquoted that can replace the evaluation by human experts as cases of unique students, special writing abilities and imaginative skills of a learner may be quite shady for an automated grading system, yet they believed that performance on content specific and amazing essays can be improved on with the use of advanced NLP features.

Landauer, Laham and Foltz (2000) stated that a standard way to weigh accuracy of a set of grades for essays is to quantify how well two independent/autonomous scorers agree with each other either a score by an AEG system and the other by an expert or the two scores being generated from two human experts in the field of discipline. Yet, they argued that this criterion is not absolute (entirely adequate) since two human experts may vary in opinion. They further claimed that there is no justification why an AEG won't be able to envisage a human grade better than another human grade, once an AEG system is more reliable/constant in assessing same factor(s) in all essays that fall in the same category and topic while stressing the need for the AEG system to actually assess the major factor(s) that human graders would have measured. Another standard for measuring essay grading accuracy is to correlate the grade from the AEG to that of human expert grader. Evaluating studies of IEA's reliability versus human graders

Additionally, it is equally important to be able to identify and automatically extract from texts measures of writing quality that are also a true reflection of the intrinsic qualities

that form the basis of human judgments. The methodology and assessment criteria adopted by such systems should be transparent, logical and meaningful. As the practical utility of automated systems depends strongly on their robustness to subversion, threats to their validity should also be identified and addressed. Several other challenges arise, such as their further development to function as learning tools, giving feedback on someone's writing skills and progress in similar ways and as useful as humans do.

Teaching, Learning and Assessment of Economics at the Secondary School Level

Economics examination was first taken as a subject at the WASSCE level in the year 1967, then it was attempted only by private candidates for the General Certificate Examination (GCE) before it was later incorporated as a secondary school subject. Economics was a two-year course content at start which can be connected to the fact that school certificate was just a two-year course as at then.

The National Policy on Education (2006) as structured subjects at the secondary school level into compulsory (cross-cutting) core subjects, Science and Mathematics, Business Studies and Humanities. Economics has been grouped as part of the humanities and one of the elective courses. The Economics curriculum at the secondary school level has been structured into about 26 topics in all and 34 teaching units for a three-year study programme (Yusuf, 2011).

The nature of Economics as a subject studies human behaviour. It is an aspect referred as social sciences. Efforts has been made to incorporate and link the theoretical principles of the subject to practical applications. The objectives of studying Economics are to raise students' respects for dignity of labour, to facilitate students' skills for practical solution to basic economic problems of the society, to arm students' with the rudimentary philosophies of Economics required for living, and to organize and inspire students' to be judicious in their management of scarce resources out of many other objectives in the curriculum.

Ojo and Nkoyane (2016) reported that quite a lot of factors affect effective teaching and learning processes of Economics in senior secondary schools in Ogbomosho where they claimed that teacher-student ratio, paucity of teaching aid, dearth of library facilities and low availability of textbooks are part of these elements that affect teaching and learning process of Economics in the area. In their study, a total of 5,864 students were recorded against 26 teachers which implies that the teacher-student ratio stood at 1:225; this figure was taken from nine senior secondary schools in Ogbomosho-North Local Government Area. The teaching of economics in our schools is surrounded with many difficulties especially at its implementation stage and these have contributed massively to the drop in performance in public examination. They recommended that to raise the performance of students, teachers at the senior secondary school level should at least have a University certificate in this subject area and a background in education, this will amount to qualified teachers.

A fall in general performance at the senior secondary school level was recorded in the year 2018. Odunsi (2018) analysed the result of the WASSCE 2018 that was released claiming that 1,021 students were withheld out of 11,307 students (that sat for the examination) due to reported cases of malpractices. He further stated that 72%, 56%, 42%, 29% and 18% obtained credit passes in at least two, three, four, five and six subjects respectively while 17% of the students obtained a minimum of five credits including Mathematics and English Language whereas in 2017, it was 26%.

Chief Examiner (2017) described some pitfalls of Economics candidates as diverting from the actual question (answering out of point), weak knowledge of the subject matter and a display of inadequate preparation for the examination while possible remedies were also suggested like students should be enjoined to read recommended textbooks for the subject, thus this reading should be broad enough to expose students to terminologies and vocabularies of the subject and understanding the question been asked is not an optional decision before

providing answers to the questions. All these are part of the demands and challenges attached to assessment in Economics.

In answering Economics questions the use of peculiar terms and terminologies are necessary, a student can not afford to answer questions using common words of expression otherwise it may lead to a different meaning from the actual intention of the student. This premise as made it good enough for the use of AEG softwares to mark Economics essay writing. In addition, Economics demands that a student gives full expression to questions asked as citing of example informs the marker that a student has really learnt as expected. The evaluation, continuous assessment and examinations in Economics is expected to involve students' capacity to solving economic problems and discussions on the implication of increase or decrease of any economic variable.

Agencies/Institutions using Automated Essay Grading System

Many agencies such as test companies (non-governmental), universities, public schools, government and government agencies have been noticed to use AEG tools and still using AEG system. They have most times used it as a second-order tool to support or validate human scores though there are agencies that uses it as the only scorer of the essay written by students/learners. There are several reasons institutions engage AEG systems but notable ones are speed needed in delivery of scores, high population of the examinees, distance between learners and instructors.

Pulman and Sukkariah (2005) also stated that the UK school system uses computational linguistics techniques in automatically marking short free response. About 10 different subjects in General Certificate of Secondary Education (GCSE) usually factual sciences and they further expect few words up to 5 lines as responses to items. Items are usually stated in statements that starts with words like state, mention, describe, distinguish, explain, suggest e.t.c. An AEG

system (c-rater) was also developed by an organization named Educational Testing Service (ETS) (Sukkarieh & Blackmore, 2009). The ETS was founded in the year 1947, it's headquarter is in the United States. This organization conducts the Test of English as a Foreign Language (TOEFL), the Test of English for International Communication (TOEIC) and the Graduate Record Examination (GRE). Blood (2017) reported that the e-rater is presently used by ETS to score both high-stakes and low-stakes examinations like the TOEFL and the GRE. It was also claims that the e-rater has assisted in a quicker delivery and high quality of test scores. The ETS stated a correlation of 0.97 between the e-rater and the average human score.

McNamara, Crossley, Roscoe, Allen and Dai (2014) conducted an empirical study where three levels of class (9th grade, 11th grade and college freshman), 1243 persuasive or argumentative essays using 14 different items and four diverse time limits (untimed, 10, 15 and 25 minutes increase). Three research instruments (AEG tools) were used in this study which are the Linguistic Inquiry and Word Count (LIWC), the Writing Analysis Tool (WAT) and the Coh-Metrix. Largely, the developed AEG tools scored all the essays an exact accuracy of 55% and an adjacent accuracy of 92% between the scores of the essay and the human scores.

The University of Cambridge Local Examinations Syndicate (UCLES) provides assessment through three units which are English for Speakers of Other Languages (Cambridge-ESOL), Cambridge International Examinations (CIE) and Oxford, Cambridge and RSA Examinations (OCR). The UCLES conducted a research to investigate the application of Natural Language Programming (NLP) techniques in marking free-short response essay of Biology in General Certificate of Secondary Education (GCSE). These short-response items are questions that necessitate students to write one or two sentences only and they are worth one or two marks most times. 201 answers were used to train the AEG tool, 88% accuracy to human raters was achieved (Sukkarieh, Pulman, & Raikes, 2003).

The Graduate Management Admission Council (GMAC) is the organizing body for the Graduate Management Admission Test (GMAT) has been in existence for over 61 years since 1953, with the aim of gaining admission management programs of business schools across the World/globe. This test examines candidates on four main areas of knowledge which are Analytical Writing Assessment, Integrated Reasoning, Quantitative Reasoning and Verbal Reasoning. GMAT is a standardized computer-based test and the result from this examination is stated to be valid for 5 years from the date of passing the examination. AEG is widely used in scoring short-written responses and also in assessment tasks in higher education admissions (GMAT). It has been proved that AEG is reliable and valid in grading essay test in GMAT (Steedle & Elliot, 2003).

A research was conducted where essay tests in GMAT were used by Intelligent Essay Assessor (IEA) and the agreement with human graders were between 85% and 91% (Valenti, Neri, & Cucchiarelli, 2003). To assess the accuracy of AEG tool (IEA) in a grading situation, a sample of 600 GMAT essays (argumentative and opinion essays) were scored and a correlation of 0.86 between human raters and the IEA while a correlation of 0.87 and 0.86 respectively between the two human raters for opinion and argumentative essays respectively (Blood, 2017).

The ACT is an examination body that organizes standardized test paper-pencil test. It was established in the year 1959 by Everett Franklin Lindquist (Professor) from the varsity of Iowa. Rudner, Garcia and Welch (2005) reported that AEG is a widely accepted tool to compliment but not replace human scorers. They wrote a paper to provide an overview of GMAT Analytical Writing Assessment (AWA) and ACT test that was done in a twofold evaluation process where 3 to 4 human scorers were used to score same AWA test and then one of the human scorer was replaced by the AEG machine (IntelliMetric) while on the long-run the scoring of the test is left to the IntelliMetric system. 270 essays were used as training

manual and 500 essays were used as validation essay. An average of 0.83 Pearson R correlation agreement was established between the human raters and the AEG system (IntelliMetric).

Klobucar, Deane, Elliot, Ramineni, Deess and Rudniy (2014) stated that SAT writing (SAT-W) (an essay) has been used by New Jersey Institute of Technology (NJIT) as a measure for placement and for prediction of students' outcome in a course. They carried out a research in the fall of a semester on all entering first-year students at the NJIT (N=1006) with 800 word limit. Two essays were written by each students but only 249 students responded in 48 hours. A random sample of 135 essays were scored traditional by human raters. A weighted Kappa (holistic score) of 0.62 was calculated between human scores and the AEG tool (e-rater). The rubrics score of rhetorical knowledge, critical thinking, writing process and conventions; had their weight Kappa to be 0.63, 0.47, 0.70, and 0.63 respectively. They concluded that there is a relatively strong basis for validation and evaluation of the relative scores.

An AEG tool (e-rater) is presently used by the Educational Testing Service (ETS) for both low-and high-stake examinations especially to score issue and argumentative essay items in the internet-based version of the writing section in the Test of English as a Foreign Language (TOEFL), the Graduate Record Exam (GRE) and other online practice tools for learners. The ETS claimed that the AEG tools has proved to be relevant in delivering higher quality scores and faster conveyance of test scores to candidates. Siddhartha and Sameen (2010) reported that TOEFL examination uses AEG systems and human graders where the average of score is been given to candidates.

To the researcher's knowledge or findings, there is no report of any institution in Nigeria that has made use of AEG system or that is making use of it. Though, finding or report exist that they make use of computer base test and or Computer Assisted Testing (CAT) facilities which is majorly the use of computer to administer objective test (Multiple Choice

Test). The Joint Admission Matriculation Board of Nigeria has been found to be a major user of CAT.

Issues relating to development and use of Automated Essay Grading in Nigeria

In Nigeria, quite a few Universities have engaged the use of computer to deliver, take and score examinations or test electronically which is referred to as E-exams or Computer-Based Testing (CBT). These universities includes University of Ilorin Ilorin, Federal University of Technology Minna, Nigerian Open University of Nigeria (NOUN) and Covenant University, Otta. In all these institutions, only NOUN uses internet base for their e-exams while others use intranet setup within the institutions system. A range of 50 to 200 computer units and a server are regularly the setup though UNILORIN CBT center has well above 1,400 computers. It is of note that all these institutions use only the multiple choice automated marking system, they do not grade essays automatically (Adebayo & Abdulhamid, 2014).

However, quite a large number of Nigerian Universities also have widely embraced multiple choice automated marking system in conducting Post Unified Tertiary and Matriculation Examination (Post-UTME) otherwise known as pre-admission screening (Adebayo & Abdulhamid, 2014). Most AEG system has been developed to have certain characteristics and models which can be referred to as architectural designs of the tool. More often, there is a 3-tier design comprising of the presentation tier, the logic tier and the database tier. The database tier handles the pool of items and answers to questions, the presentation tier comprises of the users' interface while the logic tier serves as a system that process user's requests.

This design is commonly used within Nigeria and in other countries with little or no modification but it does pay little attention to security issues and impersonation (Adebayo & Abdulhamid, 2014). Several AEG tools are monetized hence the specifics of the features and model of computing the scores are not disclosed. Nevertheless, for most part, AEG systems

tend to depend on a mixture of regression analysis techniques (linear or step-wise regression) and machine learning algorithms. In some cases benchmarks are set such that to be awarded certain marks, the essay must have certain features. An algorithm is measured to be successful if human rater scores are moderately same as the algorithms scores (McNamara, Crossley, Roscoe, Allen, & Dai, 2014).

Lack of efficient training and skills in use of computer can be a challenge to an e-examination. Students and staff of institutions should be properly informed and acquainted on the outlook or interface of the automating marking system that is of use in that institution. The fact that humans are expected to monitor and invigilate the e-Exam is a possibility of human influence. Therefore the need for activities/event recording facilities would not be out of place in order to ensure that nothing is fishing due to human interference (Adebayo & Abdulhamid, 2014).

According to Zupanc and Bosnić (2015) the concept of validity has been associated to what students has learnt over a period of time. Validation is defined as the gathering of information to support understanding and use of a test score. They further defined reliability as the ability of a test score to remain the same accurate score regardless of the change in situation. Commonly used approach by AEG systems authors to establish validity and reliability is to correlate the score awarded by their AEG tool to that of a human experts. Researcher, has concluded on several occasions that an AEG tool is skillful enough to grade essays.

Though, there is the challenge of thinking that AEG tools are capable of replacing teachers' responsibility of scoring an essay yet the machine does not understand an essay. However, the assumption that human raters are not prone to making mistakes in grading essays is also out of place because in reality, human graders are usually inconsistent, bias and unreliable. Attali and Burstein (2006) proposed three steps to validating an AEG tool. Firstly, establish the features that the AEG tool is measuring (attribute validation). Secondly, what is

the effect of summing the different grading features on the total allotted score to an essay. Finally, that an aggregate of the human and machine score should be a true score of an essay.

A persuasive criticism of AEG is negative washback which is caused by the awareness on the part of students and teachers that the AEG tool will be employed to make their essay writings (Blood, 2017). The argument is that when students are fully notified that their essays will be solely machine marked the tendency that students will be writing to the machine rather than the human audience is very high which implies that they will specifically target features like essay length and use of specific vocabularies that most strongly affect scores. Moreso, when administrators and decision makers are concerned about low performance of students, the pressure to upgrade scores of learners may result in a situation where teachers will further encourage their students to practices washback in order to conquer the machine scoring algorithm (Blood, 2017).

The method of preparing questions for the e-exam is regularly from the lecturer in-charge of the course, submitting the questions to an administrator (departmental exam officer in some cases or an administrator at the exam centre/facility) which will in turn take the questions to the point of pooling the items into the database of the exam facility. The implication here is that when examinations pass through third parties either in hours or less, there is the risk of question leak (Adebayo & Abdulhamid, 2014). AEG tools may be faulted if not considered in the development process of the tool or if the AEG system is not trained in such mannerism, the tendency to mark down or mark wrongly a students' response where negation of a negation is turned affirmative or positive. For example "I don't want any other coloured car except red" and when the response is "I want a red coloured car".

Researches has evidenced that AEG tools have the ability to provide effective feedback to learners but in all the institutions mentioned above. Learners' are not allowed to see the marked tests even if they so wish neither are the AEG tools putting up a feedback to learners'

or students'. However, common errors in AEG systems comes from the lexicon that is to words not manually identified as metonyms (Valenti, Neri, & Cucchiarelli, 2003). In Nigeria, AEG system has not been of popular use though there is an existing or newly developed one called ES4ES (Expert System for Essay Scoring) but it has not been made available for research and academic scholars which is the reason this researcher could not get empirical literature on ES4ES (Ade-Ibijola, Wakama, & Amadi, 2012).

Conceptual Framework

Automated Essay Grading (AEG) system is a mixture all six or any five or less but never one of these techniques, which means that at least two techniques is needed for an AEG system to thrive and work well. Natural Language Processing (NLP), Artificial Intelligence (AI) also called Machine Learning, Statistics, Linguistics, Text Categorization and Web Technologies (Siddhartha & Sameen, 2010). Seven out of ten most commonly used AEG makes use of NLP, a combination of linear regression or algorithm and text categorization was added in some instances which is a part of the variations in AEG software. The theoretical

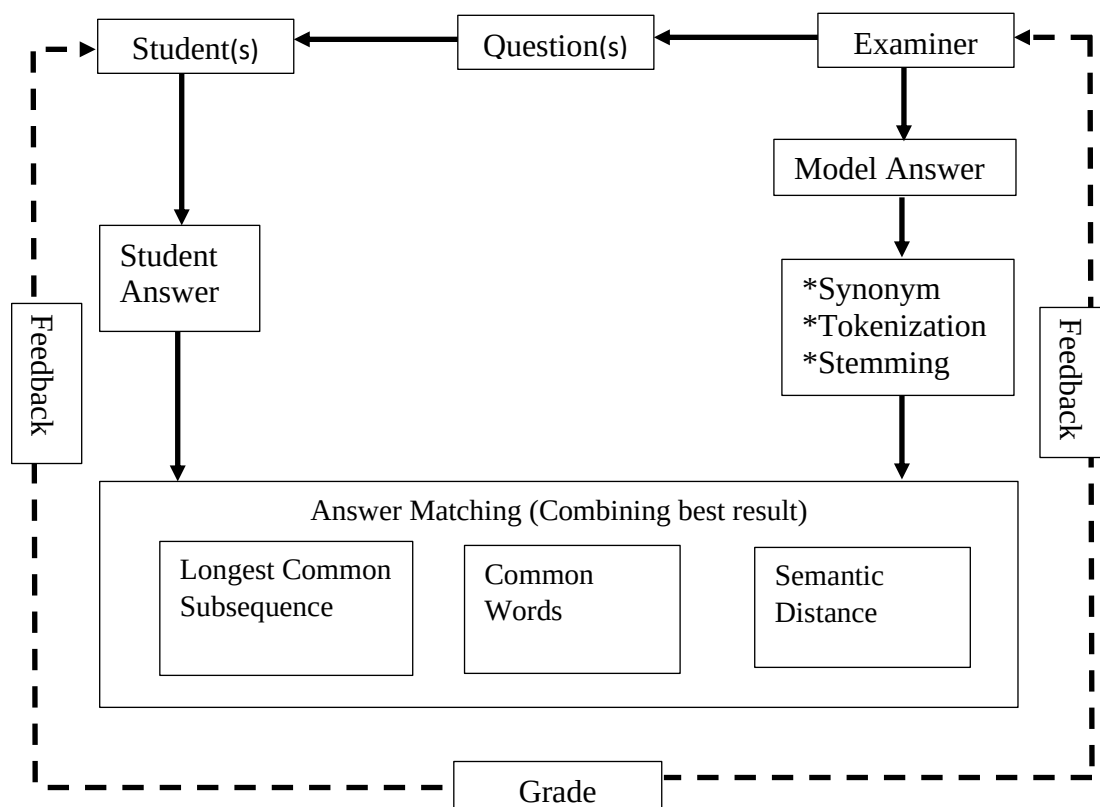


Fig. 1: Architecture of Automated Essay Grader

framework diagram in figure 1 captures the activities and concerns of the two major players in teaching and learning process which are the teacher (the examiner) and the student (learner).

The examiner is expected to supply both the question and the expected model answer. Student responses are collected afterwards and a match of students response is done with the model answer before a grade is been awarded to the response. The examiner is expected to further identify the key words in a model answer to the system, though the system automatically can do this by removing ‘stop words’ (auxiliary verbs and prepositions) such as if, as, the, to, at, an, a, what, where, that, on, of and so on using tokenization. Once the removing of stop words has been done, sentences will have reduced to short key words the next step is the stemming where Porter Stemming removes all prefixes and suffixes to words except it has been identified by the examiner as part of the key words when the model answer was drafted. This is done to get the canonical of a word. Similar words (synonyms) to the key words will also be identified in conjunction with root words. The distance between two or more words (arrangement or placement or position of words) in sequence will also be taken care of by semantic distance and longest common subsequence.

The matching phase of the model, this phase combines the similarity algorithm (answer matching) which are Common Words (COW), Longest Common Subsequence (LCS) and Semantic Distance (SD). The system will run the responds of the students combined with the weight of each algorithm base on their strength and will award grade in relation to the model answer as supplied by the examiner. The dashed lines imply that there is an indirect connection back to the examiner and the students. The feedback does not necessarily have to get back as a link between the examiner to the student nor through the student to the examiner. It can always stop at the end of each of them especially when the connection between the student and the examiner is a distant type.

Appraisal of Reviewed Literature

Artificial intelligence is featured relatively by computer system experts around the world. Considering an educational perspective, the question should be what exactly are we interested in; should human markers become more like machines (Automated Essay Graders- AEG) or should computers be more like humans in grading?. Otherwise, can we consider our perception of strengths and weaknesses of the human grader and the AEG system to foster an enhanced center ground for researchers and other stakeholders to improve the output of educational grading. This feat seems to be more realistic by comparing the scores generated by machines to scores given by human experts in any specific field as well as two human graders to the same script of a learner. Though in the field of AEG, it is particularly important to keep in mind that correlation does not imply causation (Östling, Smolentzov, Hinnerich, & Höglin, 2013).

The discovery of Automated Essay Grading (AEG) system has been characterized with the use of software or codes that are meant to function up to the level of human intelligence. The marking algorithm has experienced encryptions like Natural Language Processing, Information Extraction using tokenization, Statistical Keyword Analysis, Domain-Specific Dictionary; Latent Semantic Analysis, Text Categorization, Fuzzy-Scoring Model, Singular Value Decomposition and others. This idea or move of human intelligence is not new as autonomous (self-drive) cars and the talking engines.

A challenge associated with Automated Essay Grading system is that other languages are still seeking to develop functionality in this field. Although, few languages have their fair share of the AEG system, which are the Swedish, Arabic, Spanish and English Languages. The diversification and uniqueness of each of these languages has further pushed certain unique tools to be created and employed by these developers and towards building an Automated Marking System in different languages. For instance in the Swedish Language, OVIX lexical diversity measure was developed to cater for the diversity in their language.

Another shortfall is that all fields of study that have experienced the creation of AEG system are mostly textual or word and sentence based such as English, Biology, Psychology, Computer Science and Medicine; which does not seem to close the gap for mathematically inclined subjects. Though, it can be perceived that one foreseeable challenge to this can be aligned to the fact that even learners may

likely not find it easier to show mathematical workings/calculations on a screen that needs the keyboard to get/type answers for the AEG system to mark successfully.

AEG system are computer intelligent program designed for problem-solving, in achieving this certain features have generally been used by researchers and developers of AEG systems as basic qualities for awarding grades to an essay. These features are not necessarily strong or correct in their totality to be used as basis for grading an essay. A few of these features are text length, average word length, language lexical diversity measure and parts of speech. Some other features are induced by the collection of previously written essays like vocabulary, spelling error and split compound words.

The challenge of validity of grades/scores given by computers, 'construct objection' though machines do not experience tiredness or boredomness as human do and lack of human relation despite its ability to give feedback spontaneously. Another consideration of limitation of the AEG system is that it may limit learners' ability to be creative since it is meant to compare the ability or qualities of a response to the supplied model answer of the teacher from inception as the correct answer. Another noted weakness is the need to use manually marked essays to train the computer before it can successfully grade new responses. An average of 100 previously marked responses is used.

Despite all these, the reliability and capacity of computers or AEG systems to score essays cannot be undermined since a value higher than 0.75 or more has been achieved when compared. It can be taken as pure representation of agreement beyond chance and/or error of occurrence when correlated to the grades supplied by human raters (Fleiss, Levin & Paik, 2003). More so, it is established that most researchers in the field of AEG believed that computers or machines have the accurate ability of two human expert judge in any field of study and that computers are powerful or better scoring tool for cognitive skills of a learner. Ade-Ibijola, Wakama and Amadi (2012) stated that AEG is a developing technology yet it has been proven that it has the capacity to accurately and reliably grade essays. This study is focused on building automated software that can score written essays in Economics as a field of study. The researcher has not read of any published paper or empirical study in Nigeria that can be used to award grades/scores to written essays in Economics has a subject of study.

CHAPTER THREE

RESEARCH METHODOLOGY

This chapter discusses the on the under listed sub-topics:

- a. Research Design
- b. Population, Sample and Sampling Techniques
- c. Instrumentation
- d. Procedure for Data Collection
- e. Ethical Consideration
- f. Data Analysis Techniques

Research Design

This study adopted the correlational research design. This research was aimed at developing and validating an automated grading system for essay responses in Economics at the Senior Secondary School level in Nigeria. The correlational design is aimed at establishing the extent of relationship between scores from the Essay Test Assessor (ETA) and those from the human raters. This research design was adopted because the variables are not intended to be manipulated. This study is not experimental nor is there a cause and effect relationship but a measure of the strength and direction of relationship. The correlational research design is used because it is impossible, impractical and unethical for the researcher to manipulate any of the variables in this research since none of the test scores can be said to be independent or dependent on the other. Finally, the study is interested in establishing the degree of agreement between the scores awarded by human scorers and the ETA without leaving it to occurrence of chance in its measurement.

Population, Sample and Sampling Techniques

The population for this study consisted of all senior secondary school students offering Economics in the South-west geo-political zone of Nigeria. The South-west of Nigeria comprises Ogun, Osun, Oyo, Ekiti, Ondo and Lagos States. Each State in this zone has three senatorial districts. The target population comprised of all senior secondary school two students of Economics.

Multi-stage sampling procedure was employed in this research. The simple random sampling technique was used at the first stage to select three out of the six states in the South-west geo-political zone, since each state has three senatorial districts and three states were selected, it implied therefore that students were sampled only from nine senatorial districts out of the eighteen. This implies that 50% of the Senatorial Districts of States in the South-west geo-political zone of Nigeria were covered by the sample.

Records of Osun, Ondo, Ekiti, Lagos, Oyo and Ogun State Ministry of Education as at 2016/2017 session showed that there were 456, 569, 311, 255, 484 and 363 public senior secondary schools in each of the States respectively. The proportionate sampling technique was used to select five percent (5%) of each of the public schools in each state (Singh & Masuku, 2014). The purposive sampling technique was used to select 5% of public secondary schools, 55 schools were involved in this study. The public senior secondary schools was purposively selected based on availability of functional computer facility in the schools.

The population of the Senior Secondary School II students in Lagos state was 111,895, while Oyo state had 69,873 and in Ogun State there were 56,882 students making a total of 238,650 students in all. Out of which only 11, 932 students are offering Economics. The sample of Senior Secondary School II students selected to participate in this study was 1,200.

The human raters for this study were selected base on teaching and marking experiences in Senior Secondary Certificate Economics Examination. Human raters selected were teachers

that have a minimum of 5 years marking experiences and their willingness to participate in the marking of respondents' scripts was highly sorted and only 32 human raters participated.

Instrumentation

A software meant to grade responses to short-essay items was developed by the researcher. It was named Essay Test Assessor (ETA). The researcher developed the ETA software using the Natural Language Programming (NLP). The sublime text a code writing application was also used in coding the language to instruct the computer.

Firstly, Hypertext Markup Language (HTML) was used to achieve the lettering, the font size and site layout effects. Secondly, the Cascading Style Sheets (CSS) was used to achieve the colour display, graphic and related designs. Thirdly, MySQL was used as the data base for collection and collating (save keeping) of all data. Fourthly, Personal Home Page (PHP) was used as interpreter and part of the script language. In addition; pattern match, word-match and Google synonymy token as well as Natural Language Programming techniques (Statistical Keyword Analysis and Surface Linguistic) was plugged in as the marking algorithm. Lastly, Laravel was used for the routing, sessions and caching.

This ETA tool was developed to grade free response items in Economics in line with teachers' marking scheme. It presents to testees opportunities and characteristics of a paper-pencil test. These includes examination instructions, stipulated time, flipping through the questions at once if so desired, start and stop the examination at the same time. There is the teachers' back-end, students' front-end and the administrators' end with each of these personnel assigned powers and requisite restrictions. The teacher is free to grade for content accuracy and textual evidence if need be.

Five theory items were selected from both the West African Examination Council (WAEC) and National Examination Council (NECO) Senior School Certificate Examination (SSCE) items within year 2015 through to 2017 in Economics. The test items picked were

within the SS 1 and early SS II curriculum contents because only these areas were expected to have been covered by teachers of the selected schools. These questions were scored over 30 in all. A model answer of SSCE standard was developed by human raters in Economics who are teachers of the subject for the Automated Essay Grading test items used for the study ranged in contents from demanding such skills as defining, stating, listing and identifying to explaining Economics concepts and terms.

Procedure for Data Collection

The researcher collected a letter of introduction from the Head of Department, Social Sciences Education, University of Ilorin and thereafter proceeded to seek for permission from the school authorities to get responses from their students. For ethical purpose, the researcher informed all the respondents that their responses would be treated confidentially. In line with this, no respondent was asked to write his/her name or any means of identification on his/her essay.

The researcher used the help of research assistants and also liaised with teachers of computer science in the selected secondary schools so that respondents can make use of the school computers (computer laboratory) to answer the questions using the notepad (text document) and this was collected via an external device (Hard-Disk, Flash Drive) or mailed to the researcher. The researcher also collected some students' responses using hardcopy (paper and pencil) from schools that are either not having sufficient computers or the challenge of power supply. All responses that were collected using paper-pencil method were all typed back into the computer so that the ETA could score as well.

Respondents were duly informed to answer the essay questions in short-response form that is responses were expected to be in the average of three to four sentences and the time allotted to all respondents was a maximum of one hour. All responses typed by students into

the computer was done using the “notepad” so as to avoid the inherent capacity of Microsoft word to either auto-correct spelling errors or underline in red, words that are wrongly spelt. All typed notepad were then collated and collected in softcopies. They were later printed out for human raters (Economics teachers at the senior secondary school level) who also award scores for each of the questions asked. All paper-pencil answer sheets were also typed appropriately without correcting errors (grammatical or spelling).

All respondents’ essays were graded by the Essay Test Assessor (ETA) and 32 human raters. The human raters scoring was done in two groups. Firstly, the 1200 scripts were given to two human raters who scored all the testees (respondents) independently, recording the awarded scores in different papers (score pro-forma) this is to guide both raters from been influenced by the scores the other rater as awarded. Secondly, the 1200 scripts were also given to 30 teachers to score 40 students each.

The microtime was used to calculate the time spent by ETA when scoring each respondent. The microtime is a program in HTML used to measure speed of an application from the start point to the end point of a specific action. It was used to measure and record the time spent (speed) by Essay Test Assessor (ETA) to score each respondent.

Ethical Considerations

In this study, ethical issues was well catered for. Participants were adequately informed and briefed on the purpose of the study and were allowed to decide either to participate in the study or not. This is to ascertain and fulfil informed consent. Participants were informed that their responses would not be used for any other purpose than this study.

Furthermore, respondents were assured of confidentiality and anonymity as their responses would not be traceable to them. More so, they were neither requested nor required to write any means of identification on their responses. The researcher only used arbitrary

identification for each respondent, in order to ensure that scores for individual respondents is not mixed up with another respondent's script during analysis.

At the point of data gathering, only respondents who were willing to use the note-pad in writing responses were taken. The note-pad was used to avoid auto-correction of spelling errors in words (Microsoft Word). The benefits of these ethical consideration is to minimize the physical, psychological and social risks that this research study may have on participants.

Data Analysis Techniques

The microtime was used to calculate the speed ETA used for scoring each respondent. The scores from the Essay Test Assessor (ETA) and the 32 human raters (experienced Economics teachers at the Senior Secondary School level) were collected and collated. The average score of the first two human raters was done for each respondent and this was analyzed using linear kappa and the Spearman Rho in relation to the score from the ETA.

The data were analyzed using the linear Kappa. It is a metric measure of the degree of agreement between two scores by weighing the general score to the score of each respondent in the data. It is comparable to correlation coefficient but not the same and its value ranges from zero to one. Linear kappa inter-rater reliability analysis is the dominant evaluation measure used for Automated Essay Grading systems and correlation to measure the degree and direction of the relationship between scores. Multiple bar chart, line graph, percentage, mean, median, mode, standard deviation, kurtosis and skewness were the descriptive statistics used to answer the research questions while spearman rho, eta, linear kappa inter-rater reliability and simple regression were the inferential statistics that were used to answer the research hypothesis. The benchmark for categorizing the high, average and low scorers was base on the range of scores divided into 3 groups so as to ensure the truth representation of respondents' scores. The test had a total score of 30 marks. The human raters' scores had a range of 1 - 24

where 1 - 8, 9 - 16 and 17 – 24 were categorized as low, average and high scorers respectively while the ETA had a range of 1 – 18 where 1 – 6, 7 – 12, and 13 – 18 were categorized as low, average and high scorers respectively.

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

This chapter discussed the results and the findings of this study based on the research questions and hypotheses that were generated for the study. The summary of the bio-data of respondents were presented using frequency count and percentage while research question one was answered using qualitative description while research question two and three were answered using multiple bar chart, line graphs, mean, median, mode, standard deviation, kurtosis and skewness. Linear kappa inter-rater reliability analysis, spearman rho correlation and simple linear regression was used to answer research hypothesis one while spearman rho correlation and eta were used to answer research hypothesis two.

Demographic Description of Data

The results presented below are the demographic characteristics of students' gender and age using frequency count and percentage.

Table 1: Percentage analysis of respondents' based on gender

Gender	Frequency	Percentage
Female	552	46
Male	618	51.5
No Response	30	2.5
Total	1200	100

The results on Table 1 indicated that there are 618 (51.5%) male students, 552 (46%) are female while 30(2.5%) did not respond. This implies that there are more male respondents in this study compared to their female counterpart.

Table 2: Percentage analysis of respondents' based on age

Age (Years)	Frequency	Percentage
14-16	324	27
17-19	792	66
20-22	20	1.67
No response	64	5.3
Total	1200	100

The results on Table 2 indicated that 792 (66%) of the respondents fall within the 17-19 years of age, 324 (27%) fall within the age group of 14-16 years, 20 (1.8%) fall within the

20-22 years while 64 (5.3%) of the respondents did not respond appropriately. This implies that majority of the sampled students fall within the age of 17-19 years.

Research Questions

Four research questions were raised for this study, research question one was answered using qualitative description, research question two and three were answered using multiple bar chart, bar chart, line graph, mean, median, kurtosis and skewness; research question four was answered using spearman rho correlation and the two hypotheses were answered using linear kappa inter-rater reliability analysis, spearman rho correlation and eta.

Research Question 1: How does Essay Test Assessor score students' short response essay test items?

There are five main steps preceding ETA scoring short response essays. The first four steps are supplied by the examiner while the student supplies the fifth.

- i. The supply of the items (questions),
- ii. the model answer,
- iii. the key words in the model answer,
- iv. the total score for each item and
- v. the short essay response.

After which the following steps follows, the ETA

- a. then matches the model answer to the students' responses,
- b. now allocate scores in percentage of the total score stipulated by the examiner for each item,
- c. displays score on both the examiner and students platform, and
- d. scores are awarded to each item separately.

Research Question 2: What is the performance of students in Economics resulting from human raters?

To answer research question 2, the scores from the 30 human raters with each scoring 40 scripts, the human rater 1 and human rater 2 were used to draw the multiple bar chart. A total of 1200 scores were collated, computed and used for each group of raters. Participants' scores were grouped into three; those who scored between 0.1-8.0 were regarded as low scorers, those with 8.1-16.0 were categorized as medium scorers while those scoring between 16.1 and 24.0 were referred to as high scorers because the range of scores fell within 0-24. The percentage presentation representing the groups are drawn in the multiple bar chart (figure 2) accordingly.

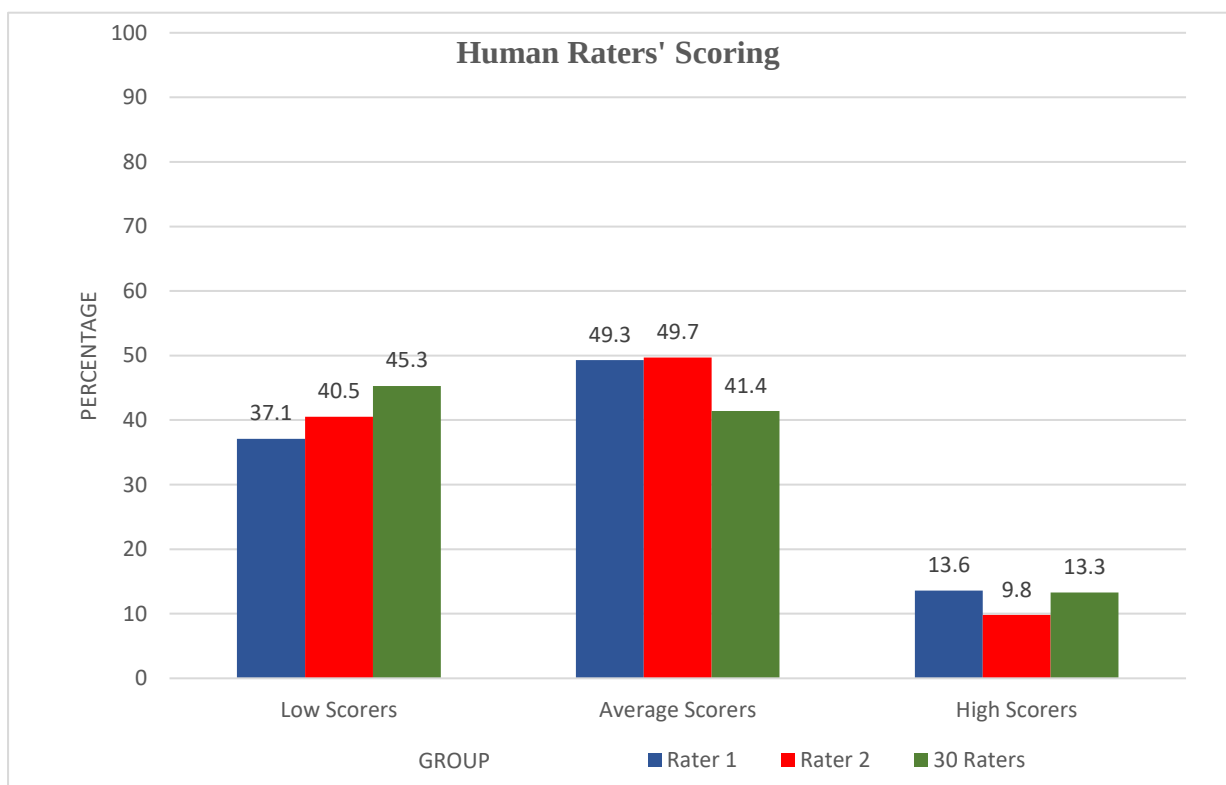


Fig. 2: Performance of Students in Economics (Human Raters)

The result in figure 2 presented a multiple bar chart showing rater 1, rater 2 and the 30 raters for the low scorers group to be 37.1% (441 scores), 40.5% (482 scores) and 45.3% (539 scores) whereas at the average scorers group to be 49.3% (587 scores), 49.7% (591 scores) and

41.4% (493 scores) while at the high scorers group to be 13.6% (162 scores), 9.8% (117 scores) and 13.35 (158 scores) respectively. The general overview of the graph showed that in the low, average and high scores group; the 30 human raters, the human rater 2 and the human rater 1 had the highest percentage in each group respectively this implied that there was no particular human rater group that was very consistent, it was more or less a fair distribution of scores.

In figure 3, 40 scripts each from each of the 30 human raters making a total of 1200 scripts were scored and collated. Human rater 1 and rater 2 scored the 1200 scripts each separately, their scores were at this point divided into 30 groups of 40 scripts for each rater accordingly. The mean score for each of the 30 groups by the human raters were then computed and drawn in the line graph (figure 3) accordingly.

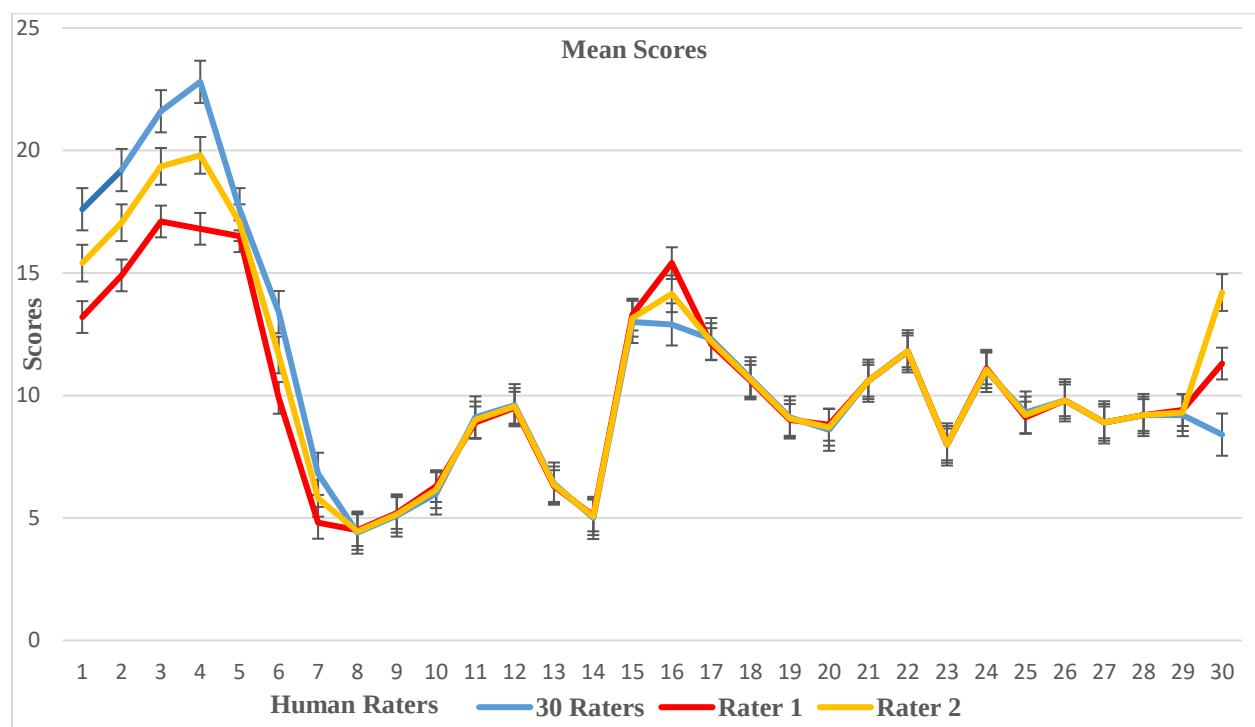


Figure 3: Mean Score from Human Raters

The line graph in figure 3 presented the mean scores for each group. The highest mean was from human rater 3 at 21.6 while the lowest mean was from human rater 8 at 4.4. The line graph displayed a very close line between the human raters 17 through to rater 29, also group 8 through to 15 but there were slight differences in group scores from group 1 through to 4, 16

and 30. This line graph showed a general close range than divergence yet the few differences can be due to subjectivity in human regardless of both teaching and scoring experiences.

In table 3, all human raters in group of human rater 1, rater 2 and the 30 human raters who scored the 1200 scripts separately, their scores were collated. The mean, median, mode, standard deviation, skewness and kurtosis were computed to further explain the performance of students' in Economics resulting from the scoring by human raters. The result is displayed in Table 3 accordingly.

Table 3: Analysis from Human Raters Scores

Human	Mean	Median	Mode	Std. Deviation	Skewness	Kurtosis
Rater 1	9.27	9	5	3.92	0.17	-0.70
Rater 2	9.51	9	5	3.51	0.21	-0.61
30 Raters	9.92	9	5	4.93	0.56	-0.25

The results in Table 3 displayed the test of normality for this study which indicated that the mean score of participants were 10.27, 10.51 and 10.92 for human rater 1, rater 2 and the 30 human raters respectively out of 30 marks which implied that 26.5%, 27.2% and 28.3% was the average score from the data set from each group of the human raters, the median score was 9.0 which is an indication of the middle score of the respondents. The modal score was 5; the values of the mean, median and the mode implied that the scores of the respondents was poor as values are below 10 marks out of 30 marks in all. Skewness indicates the asymmetry distribution of the scores around the mean, which indicated that respondents' scores cluster either to the left (negative) or the right (positive) tail of the distribution; since the values are 0.17, 0.21, and 0.56 it implied that the scores clustered towards the right side and the mean and median scores are higher than the mode.

The standard value of kurtosis is 0 for a perfectly normal distribution; kurtosis is a presentation of the level of peakedness of the scores in a distribution, an indication that the tail part is slightly flatter and longer that is the score distribution is normally/evenly distributed; -0.70, -0.61 and -0.25 kurtosis implied that the score distribution is flatter and widely spread

than a normal curve. The standard deviation 3.92, 3.51 and 4.93 showed how each score set is close to the mean, while the larger the standard deviation value is the farther away from the mean the scores are. The coefficient of variation for each of the human raters which is computed as the ratio of standard deviation to their mean are 0.42, 0.37 and 0.49 respectively, which implied that the scores are not too far from their mean scores.

Research Question 3: What is the performance of students' in Economics resulting from the Essay Test Assessor (ETA)?

To answer research question 3, the Essay Test Assessor (ETA) was used to score all the 1200 scripts collected from respondents. These scores were grouped into three, those who scored between 0-6.0 were considered as low scorers, those who scored between 6.1-12.0 were regarded as average scorers while those that scored between 12.1-18.0 were referred to as high scorers because the range of scores fell within 0-18. The percentage presentation representing the groups are drawn in the bar chart (figure 4) accordingly.

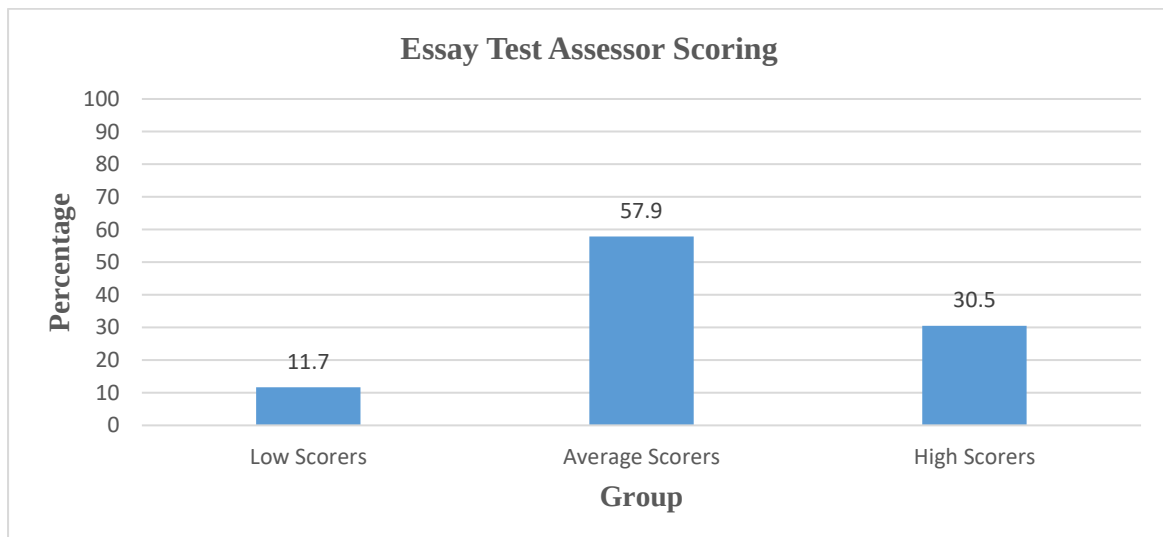


Figure 4: Performance of Students in Economics (Essay Test Assessor)

The result in *figure 4* presented a bar chart showing the highest group as 57.9% (695 scores) for the average scorer group, 30.5% (365 scores) fall in the high scorer group while the lowest group has 11.7% (140 scores) representing the low scorer group; making a total of 1200

scores in all. This implied that most of the respondents scored average scores when compared to all other respondents.

In figure 5, the ETA scores were grouped into 40 each (sequentially) to make a total of 30 groups. The mean score for each group was computed and drawn in the line graph (figure 5) accordingly.

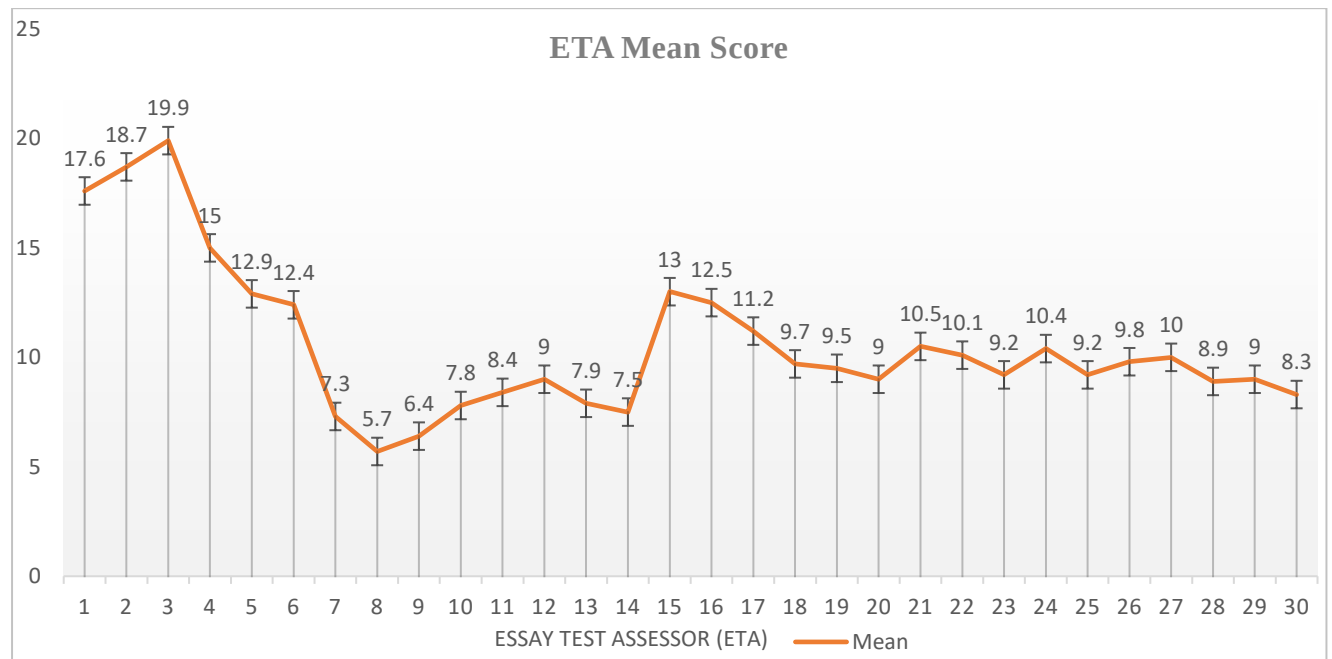


Figure 5: Mean Score from the ETA grouping

The line graph in figure 5 presented all ETA scores in 30 groups, the 3rd ETA group had a mean score of 19.9 as the highest mean, the 8th group had 5.7 as the lowest mean score while the 12th, 18th, 19th, 20th, 23rd, 25th, 26th, and 29th are in the range of 9.0 – 9.8, these means are closer to the grand total mean of 9.39. The line graph displayed two major high points at the 3rd and 15th groups but on the average the graph was fairly flat especially towards the tail end.

In table 4, all Essay Test Assessor (ETA) the 1200 scripts were collated. The mean, median, mode, standard deviation, skewness and kurtosis were computed to further explain the performance of students' in Economics resulting from the scoring by ETA. The result is displayed in Table 4 accordingly.

Table 4: Analysis for ETA Scores

Mean	Median	Mode	Std. Deviation	Variance	Skewness	Kurtosis
9.39	9.00	5	3.091	9.553	0.320	-0.786

The results on Table 4 displayed the test of normality for this study which indicated that the mean score of respondents was 9.39 out of 30 marks which implies 26.8% was the average score from the data, the median score was 9.0 which is an indication of the middle score of the respondents. The modal score was 5; the values of the mean, median and the mode implied that the scores of the respondents was poor as values are below 10 marks out of 30 marks in all. Skewness indicates the asymmetry distribution of the scores around the mean, which indicated that respondents' scores cluster either to the left (negative) or the right (positive) tail of the distribution; since the value is 0.32 it implied that the scores clustered towards the right side and the mean and median scores are higher than the mode.

The standard value of kurtosis is 0 for a perfectly normal distribution; kurtosis is a presentation of the level of peakedness of the scores in a distribution, an indication that the tail part is slightly flatter and longer that is the score distribution is normally/evenly distributed; -0.789 kurtosis implied that the score distribution is flatter and widely spread than a normal curve. The standard deviation 3.09 showed how close the score set is to the mean, while the larger the standard deviation value is the farther away from the mean the scores are. The coefficient of variation is 0.33 or 33% which is computed as the ratio of standard deviation to the mean, this implied that the scores are not too far from the mean score.

Research Hypotheses

The following research hypotheses were raised and answered to achieve the objectives of this study:

H₀₁: There is no significant relationship between the scores assigned by the Essay Test Assessor (ETA) and the scores assigned by human raters.

In Table 5, the result displayed were the relationship between the scores of human rater 1, human rater 2 and the 30 human raters. The Spearman Rho correlation was computed and the result of the analysis are shown in Table 5.

Table 5: Inter-Correlation between Human Raters

Rho	Human Rater 2	30 Human Raters
Human Rater 1	0.657**	0.678**
Human Rater 2		0.681**

**. Correlation is significant at the 0.01 level (2-tailed)

The result in Table 5 showed a positive correlation between all human raters, this implies that all human raters have the possibility of scoring students' responses high if their answers are score worthy according to the marking scheme and vice versa. All sig. value also showed 0.000 which is lower to the critical value at 0.01 level, this implies that there is significant relationship between all human raters regardless of how many they are in number especially when they all have cognate experiences in scoring the subject. The spearman rho figures were all in the range of 0.65 – 0.68 which also implies that all human raters' scores were fairly strong when compared to each other.

In Table 6, the result displayed were the computations from correlating the scores of human rater 1 and human rater 2 that scored the 1200 scripts separately against that of the ETA. The scores from the 30 human raters that later scored 40 scripts each making 1200 scripts was also correlated as a whole against the scores from the ETA, this is because table 6 is focused on computing correlation between the 1200 scores unlike table 7 that focused on the group of 40 scripts each.

Table 6: Inter-Correlation between Human Raters and the Essay Test Assessor (ETA)

Rho	ETA
Human Rater 1	0.629**
Human Rater 2	0.701**
30 Human Raters	0.727**

** . Correlation is significant at the 0.01 level (2-tailed)

The result in Table 6 showed the Spearman Rho correlation co-efficient between the human rater 1, human rater 2, the 30 human raters and the Essay Test Assessor (ETA). It was discovered that between human rater 1 and ETA there was 0.629 which is a fairly strong relationship, between human rater 2 and the ETA there was 0.701 which is a fairly strong correlation while between the 30 human raters and the ETA there was 0.727 which a strong relationship. A general deduction from Table 6 was that the correlation figures between ETA and human raters were moderately high which implied that the ETA seems to be more consistent in awarding scores.

In Table 7, the result shown is focused on correlating the scores of each of the 30 human raters to their corresponding group of scores from the ETA scores. This was done by categorizing the scores from ETA into 40 scripts per group.

Table 7: Inter-Correlation between each of the 30 Human Raters and the Essay Test Assessor (ETA)

Human	ETA
Rater 1	0.661*
Rater 2	0.692**
Rater 3	0.622
Rater 4	0.602
Rater 5	0.631*
Rater 6	0.836**
Rater 7	0.636*
Rater 8	0.589
Rater 9	0.685*
Rater 10	0.687**
Rater 11	0.687**
Rater 12	0.696**
Rater 13	0.606
Rater 14	0.653*
Rater 15	0.682**
Rater 16	0.683**
Rater 17	0.617

Rater 18	0.696**
Rater 19	0.681**
Rater 20	0.634*
Rater 21	0.624
Rater 22	0.697**
Rater 23	0.668*
Rater 24	0.639*
Rater 25	0.624
Rater 26	0.667*
Rater 27	0.687**
Rater 28	0.600
Rater 29	0.633**
Rater 30	0.588

** . Correlation is significant at the 0.01 level (2-tailed)

* . Correlation is significant at the 0.05 level (2-tailed)

In Table 7, the result revealed that the highest correlation was that of rater 6 at 0.836 which a very strong relationship while the lowest correlation coefficient was rater 30 at 0.588 which a moderate relationship. However, it is of note that most of the correlate co-efficient fall in the range of 0.65 – 0.68 which means the variation in the coefficients fall between ± 0.03 . It therefore implied that the ETA scores were consistent to human raters.

In Table 8, the result displayed were the computations from correlating the scores of the 30 human raters that scored the 1200 scripts in group of 40 each and the scores from the Essay Test Assessor (ETA).

Table 8: Measure of Agreement between the ETA and the 30 Human Ratets

N	Kappa	Rho	Beta	Sig.
1200	0.791	0.727	0.776	0.000

The data on Table 8 showed that the linear kappa measure of agreement (inter-rater reliability) between the ETA and 30 human raters was 0.791 or 79.1%, which is a substantial agreement based on consistency and validity of scores awarded. The linear kappa inter-rater reliability analysis is a test between adjacent scores. The kappa measure of agreement is used when differences in the scores awarded by two or more raters are not be left to occurrence of chance in measurement. Kappa is an inter-rater reliability analysis also used to determine the consistency among raters. The sig. value 0.000 was lower than the critical value of 0.05 which

implies that the null hypothesis is rejected; that is there is a significant relationship between the scores assigned by the Essay Test Assessor (ETA) and the human raters.

The Spearman Rho (*non-parametric test*) was done to determine the level of association between scores of the ETA and human raters, 0.727 ($p < 0.01$) a positive association was noticed which implied that the higher the scores awarded by the ETA, the higher the scores that the human raters would award to the same essay and vice versa. The 0.727 also implied that there is a strong relationship between scores of the ETA and the human raters. The effect size (*coefficient of determination, r^2*) is 0.529 which means that 52.9% of the variation among the scores awarded is anticipated to relate while conversely, 47.1% of the variation in scores cannot be explained. This means there are other factors that came into play in determining the scores awarded.

Simple linear regression was used to predict expected the human rater score based on the scores from the Essay Test Assessor (ETA). The ETA scores significantly predicted the human scores as $B = 0.776$ ($e^{0.301} = 0.727$, $p < 0.05$) indicating that for every score given by the ETA, the odds of a one-point gain in the human raters score increases by a factor of 0.727. Similarly, there is a 0.301 probability that the human rater will give a point for each idea (key word) that the ETA assigned a score to.

H₀₂: There is no significant relationship on the time gained between the two human raters based on the average words of response from each respondent.

In Table 9, a record of time spent was made by the two human raters that scored the 1200 scripts separately. This record was made subsequently after every 10 scripts that was scored. The recording led to 120 figures from each human rater as data for the analysis.

Table 9: Time gained in Scoring

N	Rho	Eta	Sig.
120	0.857	0.987	0.000

The results in Table 9 showed that there is a positive relationship between the time gained during scoring of the 1200 scripts between human rater 1 and 2. The spearman rho correlation value of 0.857 implied that there is a very strong relation between the human raters. The sig value 0.000 is lower than the p-value at 0.05, which implied that there is a significant relationship on the time gained between the two human raters base on average words of response from respondents. The eta is 0.987 which is an indication of the measure of time gained between the two human raters in this study. It implied that about 98.7% time was gained over the period of scoring the 1200 scripts each unlike the Essay Test Assessor which was consistent in relation to time spent; where there was really no difference in the scoring of the testees from the first script to the last. The eta is a measure of the non-linearity of the relationship between the two human raters not the causal direction of the scores awarded.

Summary of the Findings

- (a) The Essay Test Assessor (ETA) scored each of the question separately. Scores were awarded in relation to the percentage of comparism as perceived by the ETA to the model answer that was supplied by the examiner.
- (b) The performance of students in Economics resulting from human raters (human rater 1, human rater 2 and the 30 human raters) showed that most scores were in the category of the average scorers. None of the human raters scored the scripts either too high or too low.
- (c) The performance of students' in Economics resulting from Essay Test Assessor showed that fairly above half were average scorers while the low and high scorer group shared the remaining percentage. The line graph showed that scores are relatively close to each other.

- (d) There was a significant relationship between the scores given by all human raters in this study. The significant relationship was also positive and fairly strong.
- (e) There was a significant relationship between the ETA, human rater 1, human rater 2 and the 30 human raters. All the inter-relationship was positive and strong. The kappa value of ETA to the 30 human raters indicated a substantial level of agreement. In grouping the ETA scores into 30 as the 40 human raters, most scores were significant and the range of correlation coefficient had a 0.03 difference. The simple linear regression also showed that the 30 human raters score significantly predicts the ETA score and there is a definite level of probability for every key word scored by a human rater that it would be scored by the ETA.
- (f) The ETA was consistent in time gained compared to human raters during the scoring throughout the 1200 short-essay responses. The spearman rho correlation also showed that there was a significant relationship on the time gained base on average words of response between the two human raters who marked the entire 1200 responses.

CHAPTER FIVE

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

This chapter deals with discussion and conclusion based on the data analyzed and results presented in the preceding chapter. In view of the findings of this study, recommendations and suggestions were made for further studies. The aim of this study is to develop and validate a software that can score short-essay responses. Four research questions and two research hypotheses were formulated and statistically tested; data were collected using word-pad in computers while paper-pencil approach was also used in some schools. A total of 1200 Economics students from senior secondary schools were assessed in the South-western states of Nigeria. The data collected was analyzed using percentages, bar charts, line graphs, mean, median, standard deviation, kurtosis, skewness, eta, simple regression, spearman rho correlation and linear kappa has statistical tools.

Discussion of Findings

This study developed an automated software named Essay Test Assessor for grading short-essay response in Economics at the secondary school level in Nigeria. The findings of this study revealed that there is a significant relationship between the Essay Test Assessor (ETA) and human expert marker in Economics test items. The level of agreement was measured to be 79.1%. Going by Fleiss, Levin and Paik (2003), they stated that 0.75 (75%) is an agreeable level of reliability and consistency in grading written essays both at the rubric and total scoring level.

Australian Curriculum, Assessment and Reporting Authority (2015) independently engaged four software vendors (Project Essay Grader-PEG, the Intelligent Essay Assessor-IEA, the Constructed-Response Automated Scoring Engine-CRASE and the Lexile Writing Analyser-LWA) to report their findings on the same set of essays where they all reported in

quadratic weighted kappa after they have given them scores awarded by two human raters (Marker 1 and 2). Marker 1 and 2 had 0.79; while 0.74, 0.76, 0.77 and 0.73 weighted kappa between marker 1, PEG, IEA, CRASE and LWA respectively; furthermore 0.72, 0.77, 0.75 and 0.82 weighted kappa between marker 2, PEG, IEA, CRASE and LWA respectively.

The Project Essay Grader (PEG) was reported to have 0.71 or 71% coefficient correlation with human expert raters (Blood, 2017). The newer version of PEG was reported to have 0.77 or 77% coefficient correlation which is lower compared to the correlation coefficient of 0.791 for this study. Blood (2017) stated that the E-rater software was compared to a single human rater and the average scores from two human raters gave a correlation coefficient of 0.97 between the average human score and the E-rater. This is above the minimum expected correlation, it is also higher than the result of this study. However, Klobucar et. al. (2014) also reported a weighted kappa (holistic score) of 0.62 between the human raters and the E-rater which is lower than the result of this study.

The Intelligent Essay Assessor (IEA) achieved a 0.86 correlation compared to human raters. Though, using IEA a coefficient of 0.80 was achieved when it was compared to the average score of two human raters of the same essay whereas for this study 0.79 linear kappa measure of agreement was reached. Furthermore, Steedle and Elliot (2003) also reported on IEA a correlation coefficient of 0.77 between the average score of two human raters and the IEA after it has been trained with over 450 manually marked essays. Rudner, Garcia and Welch (2005) indicated 0.73 coefficient between human raters and the IntelliMetric after training the system with 270 essays. Ade-Ibijola, Wakama and Amadi (2012) obtained 0.71 correlation between the Expert System for Essay Scoring (ES4ES) and an expert human rater. They further reported that C-rater gave a 0.83 correlation co-efficient and Schema, Extract, Analyze and Report (SEAR) recorded a 0.45 co-efficient to human raters. Automark also obtained a 0.95

correlation coefficient usually after its marking algorithm has been trained with about 400 manually marked essays.

Most existing software, are focused on English Language essay writing such as argumentative, narrative, fiction and opinion essay which implies that they would rather score elements like content, organization, expression and mechanic accuracy. Mahana, Johns and Apte (2012) indicated 0.72 as the average kappa value from grading 12,978 essays of different forms (persuasive, narrative and contextual) using an Automated Essay Grader (AEG), the essays were sub-divided into 8 groups for easy of scoring and data analysis. They further clarified that the model works better on persuasive essays. Wang and Brown (2008) conducted a correlational study between an Automated Essay Grading (AEG) tool and human raters to determine the validity and usefulness of AEG for large-scale placement tests, a statistical significant relationship was deduced only in sentence structure out of other dimensions like focus, development, organization and mechanics.

Some findings did not correlate with the findings of other studies, a critic of AEG system wrote an essay to the e-rater system composed of actual words and later submitted same in confused and illogical manner, nonetheless he recorded the maximum possible grade (Gregory, 2013). Wang and Brown (2007) conducted a study where two standardized tests (namely Texas Higher Education Assessment-THEA and the WritePlacer Plus) were involved, they stated that there is no significant relationship in the overall holistic scores awarded to essays between the human raters and an AEG software (the IntelliMetric) score; though they resolved that faculty human raters' scores had a significant correlation with THEA writing test, the correlational scores between the AEG and the human raters base on five dimensions (namely Focus, Development, Organization, Sentence Structure and Mechanics) were 0.16, 0.17, 0.06, 0.21 and 0.07 respectively. They further debated that it appears like the IntelliMetric awards higher scores than human raters.

Semire (2006) reported that an average of 800 essays is required to train a few selected softwares before it can substantially score essays. Though few other Automated Essay Grading systems uses a minimum of 100 manually marked essays for training their machines on the average. The ETA was able to successfully establish the possibility of using one model answer (teacher supplied marking scheme) to grade short-answer response. As well as establishing the corrective confidence of the scores been awarded by the software.

Conclusion

The software (Essay Test Assessor- ETA) has the ability to grade short-essay responses with equal predictions to the possible grades a human expert will give. This is possible by the relative close range of scores that the ETA awarded to participants in this study. Scores were awarded reasonably to the related knowledge exhibited in students' responses. The ETA can also be taken to a level of the level of artificial intelligence.

The ETA has established and affirmed the possibility of using one model answer to score short response essays successfully and commendably. The dominant paradigm in the field of automated essay grading is that systems are usually trained with quite large sum of manually scored essays before they can be used to score essays effectively.

The ETA also concretize the fact that scoring using software are faster, accurate and cost-effective; therefore institutions of learning and examination bodies should be open to employing its use in other to reduce workload of examiners. Other factors such as prompt delivery of grades/scores to examinees, reduction in cost of scoring and removal of bias in scoring are also of emphases in the need for system grading of short-response essays.

The automated essay grading system should not be seen as a threat to replace human raters or tutors but in reality it is meant to support and help in several aspects, the human rater is still needed to provide the model answer, the key words and questions to be administered and tested. It can be further suggested that an average score of both the score from the grading

software and human raters can be gotten in high-stake essay tests, which is typical of an intention to supplement and not replace teachers' effort.

The ETA is now set to move from the research sphere to the real World of the educational system. The present model grades content more like human raters but also with minor enhancement it would do improve. The ETA would be of great use in several disciplines where short-response essay is used to examine learners.

Recommendations

This study has shown that automated essay grading is a viable tool in assessment of Economics short-essay responses, teachers can use this tool to assist in administering short-essay test as grading would not be a challenge and it will save 30% of the time teachers' spend on scoring essay so that it can be channeled to other relevant activities.

Teaching and learning processes can be made to be more fun when automated essay grading software are employed. As teachers will find it easier and less burdensome to give written assignments repeated especially if need be. Learners chances to practice writing, reading, thinking and learning as frequently as possible on any subject matter on any subject matter or topic with the aim of improving knowledge acquired is now very possible, knowing fully well that immediate and speedy scoring or feedback is achievable thereby they can improve on their knowledge and skills.

School administrators can also employ the use of essay grading systems to help improve on the writing skills of their students since the machine will also not be bias. Examination bodies can also reduce cost and time of administering their examinations and releasing grades to candidates. Measurement experts can also use automated grading software to develop test items of better psychometric parameters that are relevant for scoring essays either when they are short or extended response forms and further assist in constituting better scoring rubrics in teaching and learning activities in classrooms for all stakeholders.

Suggestions for Further Studies

Base on the focus and findings of this study, the researcher wish to suggest that studies on generalizability of Automated Essay Grading software, comparative studies that examine the mean score differences across Automated Essay Grading software/tools mean scores on a group of same respondent, experimental studies that would explore which feature of an essay impact the Automated Essay Grading software more than others to assign either high or low scores. Development of grading system that will grade extended essay responses can also be considered. Development and validation of an automated system that will score tabular and graphical contents, simulations may assist in this regard. Another study that would probably use other Nigerian languages in short- prose responses should be carried out. Even qualitative researches could be used to explore the divergence in the scores of human raters and that of an Automated Essay Grading tool, the impact of its use on learners.