ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2023; 2; 20-37

Automated Essay Grading Software Sustainability in Assessment: A Critical Review for Quality Feedback and Stakeholders Involvement

[a, b]Damilola D. Olaoye* , **[b]Henry Henry Owolabi** , & **[a, c]Seun Tayo Olaoye**

[a]GIRD Business and Environmental Consult Limited, Ilorin, Kwara State, Nigeria
[b]Department of Social Sciences Education, University of Ilorin, Nigeria
[c]School of Education, Kwara State College of Education (T') Lafiagi. Kwara State, Nigeria

Abstract

This paper explores critical review of literatures on automated essay grading software and system development procedure through the nomenclature of technology in assessment. Various techniques and methodology used in essay grading software were identified as well as different software that are valid and reliable in scoring both short and extended essay test items which various stakeholders can leverage on for cost effectiveness, scoring consistency, objectivity, timely result delivery, and quick feedback. Software development stages that are required in the developing automated scoring system are discussed. The state of heart as regard the AES software that require training of manually marked essays and those that does not require training are embedded in this review with various advantages automated essay scoring exhibits over human scoring and its criticism. The evaluation matrices for validating automated essay grading system with human raters were also identified. This reviewed study conclude that with the development in artificial intelligence a reliable and valid assessment in scoring of short-answer and extended essays is viable and realisable with prompt feedback, reduced cost and time wastage and thereby promote objectivity and fairness in scoring to learners that human expert scoring may not achieve. Finally, it was recommended from this review that more automated essay grading software that does not require training with manually marked essay and able to marked different subjects need to be developed and explored.

Keywords: automated essay grading, assessment, software development, artificial intelligence, stakeholders' involvement.

* Corresponding author.

Department of Social Sciences Education, University of Ilorin, Nigeria, e-mail: olaoyedamilola2020@gmail.com

To cite this article:

Olaoye, D. D., Owolabi, H. O. & Olaoye, S. T. (2023). Automated Essay Grading Software Sustainability in Assessment: A Critical Review for Quality Feedback and Stakeholders Involvement. *Journal of Computerized Adaptive Testing in Africa, 2*, 20-37.

Received: 3 November 2023
Accepted: 30 November 2023
Published: 22 December 2023

INTRODUCTION

Education involves teaching and learning in our society. Teaching and learning are what take place in the classroom and to know whether learning take place comes assessment. Assessment is carried out in the form of multiple choice, short-answer essay and extended essay test to affirm the level of what the learners know as regard the subject matter content. Assessment is a measurement process that support learning, accountability, and certification to determine the progress of the students (Fiseha et al., 2020). Before the use of technology in teaching and learning, most especially in assessment of learning, pen-on-paper assessment has been in use which resulted in different issues of cost, wastage, time, lack of timely feedback and also the scoring of the assessment through human raters raised a lot of problems like biasness, inconsistency in scoring the same essay responses. These aforementioned issues in pen-on-paper led credence to the use of computer technology in assessment and scoring.

Technology therefore makes significant advancement in educational assessment in different dimensions. It improves the precision and efficiency of: detecting the actual worth of the observed variables, collecting and processing information; it allows the sophisticated analysis of the available data; supports decision-making and provides quick feedback for all stakeholders; it helps to detect and record all the characteristics expressed by the students as regard three domain in education as well as the social contexts of teaching and learning processes (Csapó, et al., 2012).

Conversely, Hughes et al., (2018) pointed out that human assessors/examiners' experiences influence scores that is given to students in assessment. A cumulative of both academic and impact of real-life events experiences informed human marker. Also, background knowledge about a learner influence grades. Another factor is the benefit of doubt an assessor may displayed by waving error at some lower level of examination thinking that such error could be adjusted as learner progress in the cause of study (Hughes et al., 2018). Inferably, there is paramount need to develop software through development in artificial intelligence and technology that can reduce all the error pose by human scoring.

Automated Essay Grading Software

Essay tests are regarded as the most useful instrument for assessing learning outcomes. Those outcomes assessed in essay test entails: ability to recall, structure, incorporate ideas, and express oneself in writing; ability to identify, interpret and apply data than the ability to merely supply of answers in multiple choice tests. It is on these measurement outcomes that essay tests serve their purposes at

higher order thinking levels or dimensions of the Bloom's taxonomy (Huitt, 2011). Scoring of these essay tests through technology have been studied severally which are generally refers to automated essay scoring (AES) or automated essay grading (AEG). To Dikli, (2006) automated essay scoring systems is the application of natural language processing (NLP) and deep machine learning techniques in awarding scores to essays response for a target prompt automatically.

In 1966, AEG system was first proposed by Page called PEG but today there have been several automated essay grading software. Some of the software requires training with manually marked essay before it can mark essay appropriately while some does not require training because of development in artificial intelligence AI. Artificial intelligence (AI) is a priority technology in the world today, which is stimulated by the accessibility and the emergence of sophisticated methods and framework (Davenport & Ronanki, 2018). To explore AI, there is need to understand the complementary resources needed to be developed and implementing them in the pursuit of realizing performance gain (Mikalef & Gupta, 2021). Therefore, AI is the increasing capability of machines to execute specific roles and tasks currently performed by humans within the workplace and society (Dwivedi, et al., 2019)

Notable AEG system are C-rater (Siddiqi & Harrison, 2006), e-rater (Attali & Burstein, 2006; Cahill, Chodorow, and Flor, 2018), Essay Grading and Analysis Logic (EGAL), Apex Assessor, MARKIT, Computerized Risk Analysis System for Evaluating Student Essays (CRASE) (Valenti et al., 2017), Software for Evaluating and Assessing Responses (SEAR) (Christie, 1999), BETSY (Rudner & Liang, 2002), ES4ES (Ade-ibijola, et al., 2012), IntelliMetric (Vantage Learning, 2002), Automark (Mitchell, et al., 2002), Intelligent Essay Assessor (Landauer, et al., 2003, Pearson, 2012), PEG (Page, 2003), Automated Essay Assessor (AEA) (Kakonen et al., 2008), Automated Text Marker (ATM), Jess, PS-ME (Ramalingam et al., 2018), and Intelligent Essay Marking System (IEMS) (Anher, 2013), all of which required training of manually marked essays but only few AEG System can mark essay written responses without training with manually marked essays which are Jess (Ramalingam et al., 2018), SA-Grader, Essay Test Assessor (ETA) (Sowunmi, 2021), Automated Extended Essay Grading Software (AEEGS) (Olaoye, 2024).

Despite the fact that AEG has been proven to grade essays, AEG is still growing and being improved upon continuously (Ade-Ibijola, et al., 2012). Sowunmi, (2021) carried out correlational research design for the development and validation of Essay Test Assessor (ETA) to mark short essay

with a sampled of 1200 senior secondary II students offering Economics in South-west public schools Nigeria. The result of his findings revealed a relationship coefficient of 0.79 between ETA and 30 human raters using linear weighted kappa measure of agreement which is substantial agreement based on consistency and validity of scores awarded.

Various Automated Essay Grading systems (AEGs) used different techniques which are regression techniques, classification model, natural language processing (NLP), neural networks, machine learning (ML), ontology-based approach (Ramesh & Sanmpudi, 2021). The system that use regression technique used supervised methods to predict essay scores (Darwish & Mohamed, 2020); classification technique classify essays to either low, medium, or high as regard the topic of the question (Salim et al., 2019); neural network model is trained on Bag of Words features (BOW) and it is more accurate than others (Zhu & Sun, 2020); but ontology-based task use information extraction (Ramachandran et al., 2015). The various AEG software are systematically analysed on table 1.

Table 1: Systematic Analysis of Some AEG Software

Software	Agreement Level	Manual Training	Essay Type	Source
PEG	Kappa (0.77)	Required	Short & Extended	Blood, 2017
ETA	Kappa (0.79)	Not Required	Short	Sowunmi, 2021
AEEG	Pearson r (0.97)	Not Required	Extended	Olaoye, 2024
C-rater	Pearson r (0.83)	Required	Short	Leacock & Chodorov, 2004
SEAR	Kappa (0.45)	Required	Short	Ade-Ibijola et al., (2012)
Automark	Pearson r (0.88)	Required	Short	Sukkarieh et al., 2003
IntelliMetric	Pearson r (0.83)	Required	Short & Extended	Duwairi, 2006
BETSY	Pearson r (0.80)	Required	Short & Extended	Rudner & Liang (2002)
IEA	Pearson r (0.86)	Required	Short & Extended	Steedle & Elliot, 2012
AEA	Kappa (0.73)	Required	Short & Extended	Mahana et al., 2012
IEMS	Pearson r (0.85)	Required	Short & Extended	Ade-Ibijola et al., (2012)
Markit	Pearson r (0.79)	Required	Short	Valenti, et al., 2017
E-rater	Pearson r (097)	Required	Short & Extended	Kumaran & Sankar, 2015
ES4ES	Kappa (0.71)	Required	Short	Ade-Ibijola et al., (2012)
CRASE	Kappa (0.77)	Required	Short	ACARA, 2015

The table 1 show that most of the available automated essay scoring software required manual training to establish scoring criteria and training of their algorithms on essays that have been manually

scored by human expert. This manual training will teach the software to score essays while only three software which are Jess, ETA, and AEEGS are able to score essays without training. Therefore, it become necessary to delve into more software that could score essay without manual training as numbers of students are in millions sitting for examination or test in a single subject yearly.

METHODOLOGY

In this review, four categories of the existing methodology in AEG are discussed which are: hybrid methods; Latent Semantic Analysis (LSA); Text Categorisation techniques (TCT); and miscellaneous technique based methods (Anher, 2013). Firstly, hybrid methods which consist of the following software PEG, SEAR, E-rater and so on are better in performance generally than the AEG systems in the other categories. The AEG systems that use combination of natural language programme (NLP) techniques and statistical techniques are classified as Hybrid methods. Secondly, LSA which is was originally proposed by a psychologist named Landaueur and his colleagues in indexing of document (Nikitas, 2010). Due to excellent success of LSA in document indexing, it has been deployed with slight modification to perform the task of the automated scoring of essays. AEG systems that are based on the LSA technique are IEA, AEA, Jess, MarkIT. Thus, LSA is a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information (Anher, 2013).

LSA-based scoring is commonly used for large-scale scoring of essays responses on low and high-stakes exams and short answer scoring (Shermis, 2014). Automatic mathematical technique is used to extract and deduce meaning from the contextual usage of words in large collections of natural discourse in LSA approach (Nikitas, 2010). It is based on the statistics of how words are used in ordinary language than a simple frequency, keyword, or co-occurrence, keywords counting and matching techniques. LSA technique is used for AEG software that score high-stakes assessments like SAT, GRE, and GMAT (Noelle, et al., 2020) and low-stakes writing assessments such as military leadership and medical diagnostic reasoning (LaVoie et al., 2015). Scores generated from LSA-based approach often have high relationship with subject matter experts (SMEs) (Shermis, 2014).

The third method which is Text Categorization Techniques (TCT) is another method of an AEG system that uses several text classification techniques to perform automated essay grading system. These techniques, with little modification, is used by other AEG systems to perform the task

of automated grading of essays. TCT was proposed by Larkey at the University of Massachusetts in USA and the brain behind TCT was 'to train binary classifiers to differentiate 'good' from 'bad' essays, and use the classifiers scores to rank essays and award marks to them' (Anhar et al., 2013).

The AEG systems like PEG, E-rater, Schema Extract Analyse and Report (SEAR), Intellimetric, My! Access, Intelligent Essay Marking System (IEMS), Paperless School free text Marking Engine (PS-ME), and an AES system for CET4 can be classified as hybrid methods. The Intelligent Essay Assessor (IEA), Automatic Essay Assessor (AEA), Jess, MarkIT and an AEG system using generalized LSA can be classified as LSA-based methods. Bayesian Essay Testing Scoring system (BETSY), CarmelTC and two AEG systems using k-Nearest Neighbour (kNN) are Text Categorization Techniques (TCT) while an AEG method using connections between paragraphs, literary sameness, unsupervised learning and modified BLEU algorithm can be classified as miscellaneous technique-based methods (Anher, 2013).

RESULT

Software Development Procedure

Developing any automated essay scoring system required process or stages to follow. This stages are referred to as sets of activities, techniques and guidelines for effective and solution driven software development. There are quite a number of software development stages, some of which are identified by Helingo et al., (2017) namely: Agile, Spiral, Rapid, Incremental, Prototyping, Waterfall, Extreme Programming, Rational Unified Process, V-model, Wheel-and-spoke and Scrum. These are software engineering model that contains steps to follow in establishing standardized software.

Waterfall model: - This is documented in 1956 by Benington which entails operational analysis; operational specification; design and codding specification; development; testing; deployment; and evaluation (Bhuvaneswari & Prabaharan, 2013). This was later reform in 1970 with feedback loop to ensure each stage can be revisited.

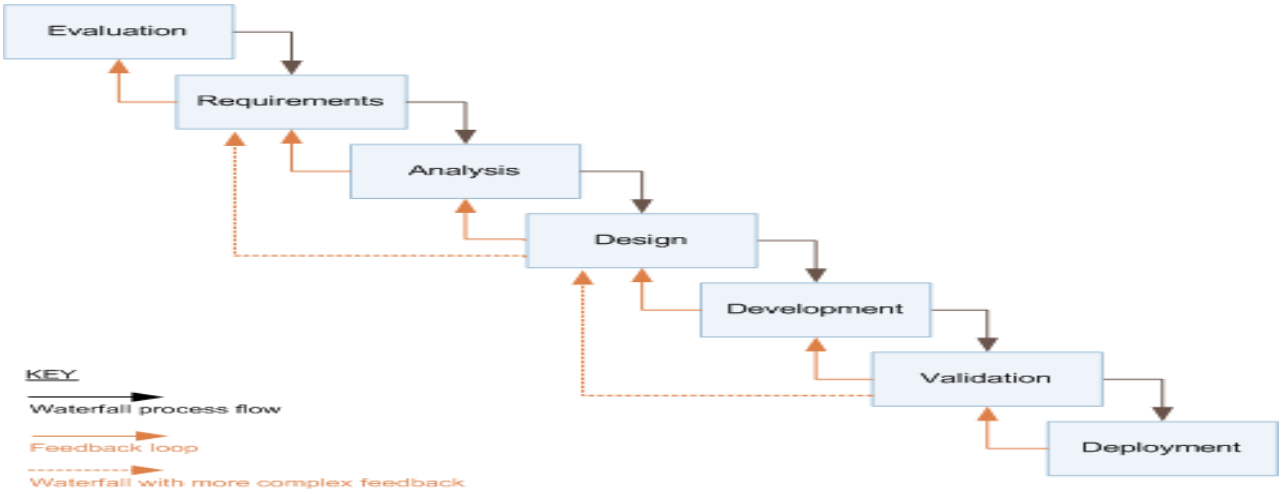

Figure 1: Waterfall model Ruparelia (2010) SDLC Software Engineering Notes

V-Model: - v-model (vee model) in 1991was developed by NASA (Ruparelia, 2010). The left side of the V shape represents the user requirements which comprises of definition and decomposition while the right side deals with integration and verification of the system design with sequential levels of building and implementation. The vertical axis indicates the level of decomposition from top to bottom, which is from the system level to the detail at component level. Also, each phase must be completed before the next phase begins and emphasis is placed more at the testing phase in this model than the waterfall model (Bhuvaneswari & Prabaharan, 2013).

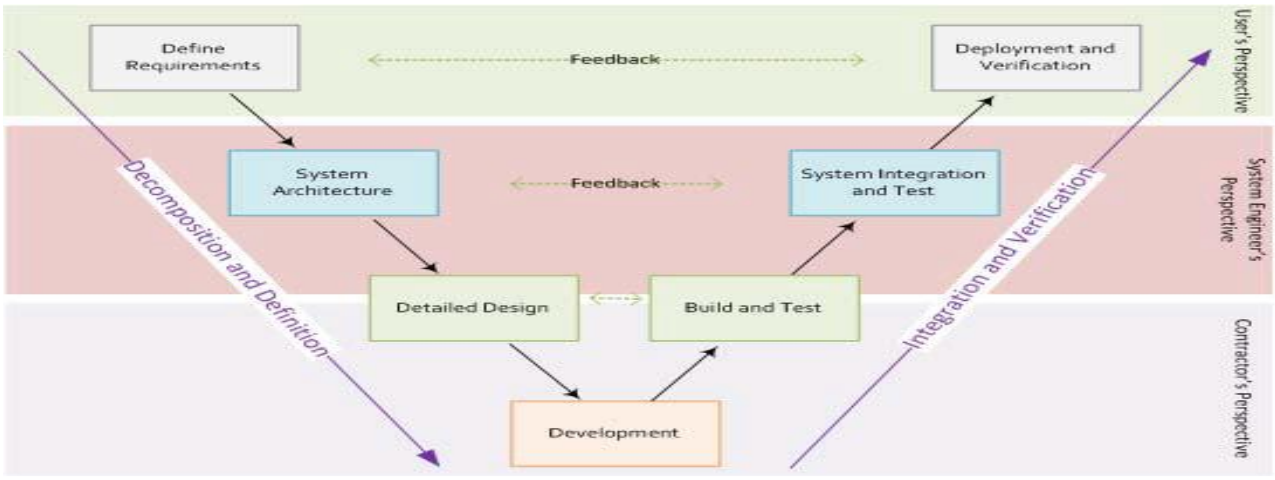

Figure 2: V- Model Source: Ruparelia (2010) SDLC Software Engineering Notes

Spiral Model: -This model was introduced by Boehm in 1986 as a result of modification of the waterfall model by introducing several iterations that came out from small beginnings which backed-

up with the idea of 'start small, think big'. This model focus on risk analysis and use one standard SD model to build the software.

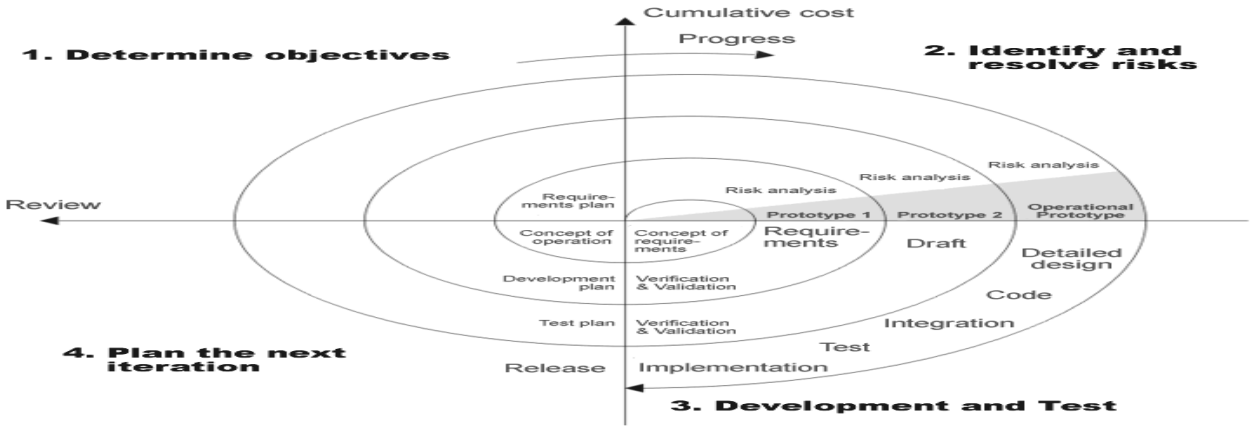

Figure 3: Spiral Life cycle of Boehm Source: Ruparelia (2010) SDLC Software Engineering Notes

Wheel-and-spoke Model: - is initially designed to work with smaller teams but was later scale up to build value faster. It is an approach of bottom-to-up. At first the preliminary design is develop before the requirements for the system is created, later prototype is designed at the implementation stage which it is verified against the requirements to establish the first spoke (Ruparelia, 2010). Also, at the development cycle feedback is added and value is inserted to the next stage to create a refined prototype. Then the second spoke is formed through evaluation of the requirements so as to propagate feedback back to the development cycle. At each consecutive stage a spoke is form by going through a similar verification process. Through iteration of the development cycle a wheel is created to be more adaptable to end-user applications and services.

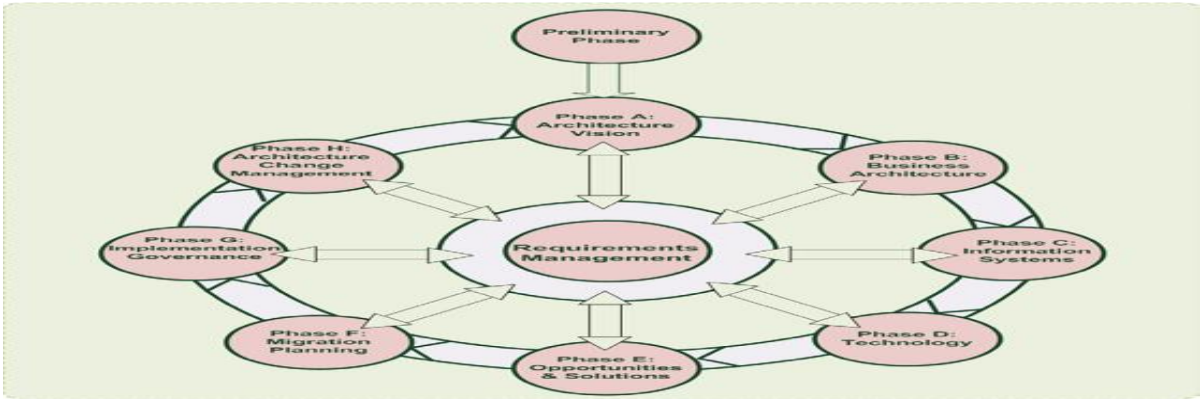

Figure 4: Wheel-and-Spoke Model. Source: Ruparelia (2010) SDLC Software Engineering Notes

Rapid Application Development: -RAD was developed by James Martin in 1991(Ruparelia, 2010). It use prototyping mechanism methodology to promote a collaborative atmosphere and active participation for business stakeholders by creating test cases and performing unit tests in prototyping (Ruparelia, 2010). With RAD, decision making structure is decentralized to make functional teams of projects manager and developers. RAD contain variety of techniques that speed up software development as depicted in the figure 5.

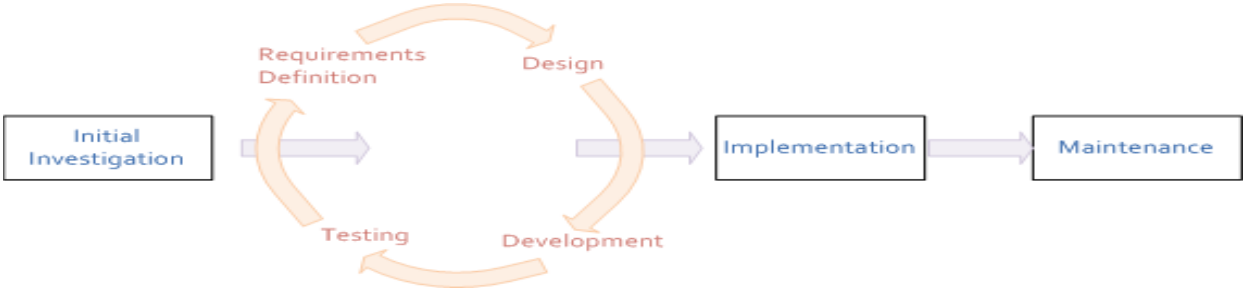

Figure 5: Rapid Application Development RAD. Ruparelia 2010.

Other models includes: - Incremental model which is viewed as a three-dimensional representation of the waterfall model (Ruparelia, 2010). This model is developed through repeated iterative and smaller incremental portions par time, which allows software experts to take advantage of what was learned during development of earlier parts of the system (Bhuvaneswari & Prabaharan, 2013). Iterative model- This model start by specifying and implementing a part of software to be reviewed with no full requirement specification. This to identify further requirements because each steps are repeated to produce new part of the software in each cycle of the model (Larman & Basili, 2003).

Agile model- is another model that is based on scope changes with different features coming up. It was created to brake procedures into smaller sub-procedure. Therefore, within short interval development occurs to capture small incremental changes. Applying agile model large projects become problematic as a result of emphases on face-to-face, real-time communication and personal preference. Also, little documentations are produced with agile model during development, this is why agile model is suited with visual interface. This is a conceptual framework for software engineering that begins with planning phase and ended with deployment phase of iterative and incremental interactions through the software development life cycle (Al-Saqqa et al., 2020).

Unified Process Model- UPM is architecture based and case driven (Ruparelia, 2010). This model is also iterative in nature which was design to tackle specific requirements development of object-oriented system. UPM uses model called Unified Modeling Language (UML), which entails the collection of semi-formal graphical notations. This UML facilitates the construction of several software system that supports both static and dynamic modeling (Al-Saqqa et al., 2020). Also, Extreme Programming (XP) - popularly known XP was established to overcome the limitation of the conventional software development process as a result of emerging constant changes in requirements, and to have a method that is suit object-oriented project in form of multiple programmers in a single location (Choudhary & Rakesh, 2016). Thereby improve software quality through the concept of extreme level of software practices.

Joint Application Development (JAD) is another process that promotes collaboration with the end user. This is done by involving end-user through workshops known as sessions during the design and development phases. The problem with this model is that if the user's requirements are not handled well will create a loop hole in its scope. While Scrum model is more appropriate for small projects where development takes place over a chain of short iterations to measured daily progress. This scrum suits category of visual interface for the end users as it focuses on process independent approaches and less formal (Ruparelia, 2010). Perkusich et al., (2015) used Scrum procedure in his empirical study based on Bayesian networks to assist in detecting the process problems in software development projects and through simulation scenarios the Bayesian network was validated. The most widespread agile method is scrum method (Balle et al., 2018).

Software development procedure are structural framework, guideline, rules and process of all tasks and activities that is implemented in developing software. Each model follows a particular lifecycle to achieve success in the development process and procedure or model used will directly affect the quality of software product (Yu, 2018). Similarly, quality software process give birth to quality software product that is why it is very important for developers and every organization to focus on software development procedures as it may require to mix multiple models to have a reliable and valid software.

Advantages of AEG to Human Scoring

Tests based on computer benefit the user in a number of ways when compared with tests based on papers. It includes the distribution of paperless test and subsequent collection of data, a better standard to administer the test, and getting responses to write and speak and score using

machine, and provision of quality tools (e.g., calculators and dictionaries). It also provide a chance for questions interactive style (Bridgeman, 2009). Also, Hearst (2000), submitted that using computers to grade essay provides benefits of effective instructional materials for improving reading, writing and other communication abilities. Having effective scoring systems will provide universal access to electronic information to our assessment and education systems.

Automated scoring is an amazing introduction of the technology in the field of education that claims the reduction in time and cost when it comes to the question about assessment of higher levels of skills, for instance, written expression, but it require validation regarding acceptance by those who are going to use it (Weigle, 2010). Also, through development of artificial intelligence, there are automated essay scoring software that capable of awarding scores to both short and extended essay type without training with multiple manually marked essay. The strength of AEG in scoring to human scoring are summarily identify from literature in the table 2.

Table 2: Strength of AEG to Human Scoring

S/N	Human Scoring	Automated Scoring
1	Not Consistent and subjective	Consistency and objective
2	To reproduce is very difficult	Easily reproduced
3	Score errors are difficult to track	Tracking of score errors are made easy
4	Not prompt in scoring	Very prompt in scoring
5	Waste time	Time saving
6	No instant feedback	Instant feedback
7	Expensive and not cost effective	Reduced cost with cost effectiveness
8	Rely majorly on human	Does not rely on human
9	Highly direct labour intensive	single trained operator is needed
10	Required Recruiting and monitoring	Recruitment and monitoring are not required
11	Training is needed	Training is not needed

Evaluation Metrics and the Need for Assessment Stakeholders Involvement in AEG

Generally as it stand in automated essay grading, various evaluation metrics have been used to evaluate AEG systems with human raters, such as Pearson's Correlation Coefficient (PCC) and Spearman's Correlation Coefficient (SCC), agreement metrics like quadratic weighted Kappa and error metrics such as Mean Absolute Error (MAE) through Mean Square Error (MSE) (Øistein et al., 2021).

Also, the most frequently used metrics to evaluate AES systems is quadratic weighted kappa (QWK) and PCC (Ramnarain-Seetohul et al., 2022). Methodology used consist of correlational, deep learning, confusion metrics, hybrid method, latent semantic similarity, text categorization technique, regression techniques, classification, neural networks, ontology-based approach, and semantic contextual similarity.

DISCUSSION

An increasing number of schools, examination bodies and higher institutions internationally and nationally are adopting Automated Essay Scoring (AES) to assess students' writing for the purpose of placement, promotion, certification and accountability. Education testing agencies and AES developers have published numerous research results that generally show high agreement rates and strong correlations between AES scores and human raters' scores, yet the predictability rates still being research. Because writing assessment is intimately related to teaching, learning, and thinking. Therefore, the use of AES tools has caused much concern from scholars. However, the realm of AES research has so far been occupied by commercial testing companies. It is important that potential users of AES in secondary and higher education begin to direct their attention to investigating how AES works and to what extent AES can replace human raters, since both writing instruction and students' learning are at stake. Also, various submission of different researchers and experts in the field of automated grading software for assessment affirm the possibilities of grading free response essay/long response essay even without training with manually marked essay sample (Sowunmi, 2021; Olaoye, 2024).

CONCLUSION

To score the essay whether short-answer type or extended type, essay grading system is valid, reliable and applicable as it had be proofed in different researches that automated essay grading software have high level of agreement with human markers with coefficient correlation from 0.70 to 0.97 in term of scoring. It was also concluded that there are automated essay scoring software that are trained with manually marked essay and those that does not require training. Therefore, developing AEG/AES tools that could score extended essay and any other form of test in all subjects without training manually marked scripts would be a landmark achievement in sustaining cost reduction, fairness and quality feedback in assessment.

RECOMMENDATION

It is recommended in this study that any researcher or software developer interested in developing Automated Essay Grading for any form of essay type items should consider the methodology, techniques, and stages of software development and all the areas where researches and invention have reach in the field of AEG. It was further recommended that researchers, institutions and other stakeholders should leverage on this artificial intelligence development especially those that do not require training with multiple manually marked essay in assessment of students or candidates to promote constituency, scores fairness, quick feedback and sustainable cost reduction in assessment practices. Stakeholders in educational assessment should also shift to the use of technology in the assessment process right from registration of candidates to test administration and scoring, down to score or grade dissemination and reporting to results to ensure fairness. This should necessitate the government and curriculum developers as well as teachers to develop a framework in which students right form their early childhood education familiarized themselves with computer usage so as to answer extended items on computers appropriately.

REFERENCES

Abdeljaber, H. A. (2021). Automatic Arabic short answers scoring using longest common subsequence and Arabic WordNet. *IEEE Access, 9,* 76433-76445.

Ade-Ibijola, A. O., Wakama, I., & Amadi, J. C. (2012). An Expert System for Automated Essay Scoring (AWS) in Computing using Shallow NLP Techniques for Inferencing. *International Journal of Computer Applications* 51(10), 37-45.

Al-Saqqa, S., Sawalha, S., & AbdelNabi, H. (2020). Agile software development: Methodologies and trends. *International Journal of Interactive Mobile Technologies, 14(11*) 246-269.

Anhar, F., Farookh, K. H., & Tharam, D. (2013) "An intelligent approach for automatically grading spelling in essays using rubric-based scoring", *Journal of Computer and System Sciences*, https://orcid.org/10.1016/j.jcss.2013.01.021.

Anher, F. (2013). *A Robust Methodology for Automated Essay Grading* (Doctoral dissertation, Curtin University).

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater [R] V. 2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-22.

Australian Curriculum Assessment and Reporting Authority- ACARA (2015). *An Evaluation of Automated Scoring of NAPLAN Persuasive writing.* 1-16. Retrieved from http://www.acara.edu.au/assessment/reseach.html

Balle, A. R., Oliveira, M., Curado, C., & Nodari, F. (2018). How do knowledge cycles happenin software development methodologies? *Industrial and Commercial Training, 50*(7/8), 380-392.

Bhuvaneswari, T., & Prabaharan, S. (2013). A survey on software development life cycle modes. *International Journal of Computer Science and Mobile Computing, 2*(5), 262-267.

Blood, I. (2017). Automated Essay Scoring: A Literature Review. *Working Papers in Applied Linguistics & TESOL,* 17(2), 40-64. Retrieved July 2022, from http://www.tc.columbia.edu/tesolalwbjournal

Bridgeman, B., Trapani, C., & Attali, Y. (2009). *Considering fairness and validity in evaluating automated scoring.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Choudhary, B., & Rakesh, S. K. (2016). An approach using agile method for softwaredevelopment. *2016 International Conference on Innovation and Challenges in Cyber Security* (ICICCS-INBUSH) (pp. 155-158). IEEE. https://doi.org/10.1109/iciccs.2016.7542304

Christie, J. R. (1999). Automated essay marking-for both style and content. In M. Danson (Ed.), *Proceedings of the Third Annual Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T. and Law, N. (2012). Technological Issues for Computer-Based Assessment. *Assessment and Teaching of 21st century skills*, 143-230.

Darwish, S. M., & Mohamed, S. K. (2020). Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. *In: Hassanien A, Azar A, Gaber T, Bhatnagar RF, Tolba M (eds)* The International Conference on Advanced Machine Learning Technologi and Applications.

Davenport, T. H. & Ronanki, R. (2018). Artificial intelligence for the real world, *Harvard Business Review 96* (1) 108–116.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment 5*(1):1-36.

Duwairi, R. M. (2006). A framework for the computerized assessment of university student essays. *Computer in Human Behaviour*, 381-388. Doi: 10.1016/j.chb.2004.09.006

Dwivedi, Y., Hughes, K. L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy, *International Journal of Information Management.* 101994.

Fiseha, M. G., Adeel, H. S., Muhammad I. K., & Basim, A. K. (2020). Challenges of remote assessment in higher education in the context of COVID-19: a case study of Middle East College. *Educational Assessment, Evaluation and Accountability* 32:519–535 https://doi.org/10.1007/s11092-020-09340-w

Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems, 15*(5), 22 37, IEEE CS Press.

Helingo, M., Purwandari, B., Satria, R., & Solichah, I. (2017). The Use of Analytic Hierarchy Process for Software Development Method Selection: A Perspective of e-Government in Indonesia *4th Information Systems International Conference 2017, ISICO 2017, Bali, Indonesia.* Procedia Computer Science 124 (405–414).

Huitt, W. (2011). Bloom et al.'s taxonomy of the cognitive domain. *Educational psychology interactive*, *22*.

Hughes, L. J., Johnston, A. N., & Mitchell, M. L. (2018). Human influences impacting assessors' experiences of marginal student performances in clinical courses. *Collegian, 25*(5), 541 547.

Kakkonen, T., Myller, N., Sutinen, E., & Timonen, J. (2008). Comparison of Dimension Reduction Methods for Automated Essay Grading. *Educational Technology & Society*, *11*(3), 2 288.

Kumaran, V. S., & Sankar, A. (2015). Towards an automated system for short-answer assessment using ontology mapping. *International Arab Journal of e-Technology,* 4(1), 17-24. Retrieved July 2017 from http://www.iajet.org/iajet_files/vol.4/no.1/3.pdf

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. *Automated Essay Scoring: A Cross-Disciplinary Perspective*, 87–112.

Larman, C., & Basili, V. R. (2003). Iterative and incremental developments. A brief history. *Computer,* 36(6), 47-56.

LaVoie, N., Cianciolo, A., Martin, J. (2015). *Automated assessment of diagnostic skill.* Poster presented at the CGEA CGSA COSR conference, Columbus, OH.

Leacock, C., & Chodorov, M. (2004). Scoring free-responses automatically: A case study of a large-scale assessment.

Mahana, M., John, M., & Apte, A. (2012). *Automated Essay Grading Using Machine Learning.* California: Stanford University. Retrieved fromhttp://cs229.stanford.edu/proj2012/AhanajohnsApteAutomatedEssayGradingUsin MachneLearning.pdf

Mikalef, P & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Elsevier Journal of Information and Management.* https://doi.org/10.1016/j.im.2021.103434

Mitchell, T., Russel, T., Broomhead, P., & Aldridge, N. (2002). Toward Robust Computerized Marking of Free-Text Responses. *Proceedings of the 6th CAA Conference,* pp. 231-249. Loughborough University. Retrieved from https://dspace.lboro.ac.uk/2134/1884

Nikitas, N. K. (2010) "Computer Assisted Assessment (CAA) of Free-Text: Literature Review and the Specification of an Alternative CAA System," pp. 116-118.

Noelle, L., James, P., Peter, J. L., Sharon, A., & Robert, N. K. (2020). Using Latent Semantic Analysis to Score Short Answer Constructed Responses: Automated Scoring of the Consequences Test. *Educational and Psychological Measurement, Vol. 80(2) 399–414.*

Olaoye, D. D. (2024). *Automated extended essay scoring of senior school certificate examination economics items in south-west nigeria, using semantic contextual similarity.* An Unpublished PhD Thesis University of Ilorin

Øistein, E. A., Watson, R., Zheng, Y., & Cheung, F. K. Y. (2021). "*Benefits of alternative evaluation methods for Automated Essay Scoring".* In: Proceedings of the 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 856-864. https://educationaldatamining.org/edm2021/ EDM '21 June 29 - July 02 2021, Paris, France.

Page, E. B. (2003). Project Essay Grade: In M. D. Shermis and J. C. Burstein (Eds). *Automated Essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Lawrence Erlbaum Associates.

Pearson Education, Inc. (2012). *Intelligent Essay Assessor™ (IEA) fact sheet.* Retrieved from http://kt.pearsonassessments.com/download/IEA-FactSheet-20100401.pdf

Perkusich, M., Soares, G., Almeida, H., & Perkusich, A. (2015). A procedure to detect problems of processes in software development projects using Bayesian networks. *Expert Systems with Applications, 42(1)*, 437-450.

Ramachandran, L., Cheng, J., & Foltz, P. (2015). Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. *In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 97–106).

Ramalingam, V. V., Pandian, A., Chatry, P., & Nigam, H. (2018). Automated Essay Grading Using Machine Learning Algorithm. *Journal of Physics*: Conference Series. Doi/10.1088/1742 6596/1000/1/012030

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring system: a systematic literature Review. *Artificial Intelligence Review* 55 (pp. 2495–2527) https://doi.org/10.1007/s10462-021-10068-2.

Ramnarain-Seetohul, V., Bassoo, V. & Rosunally, Y. (2022). Similarity measures in automated essay scoring systems: A ten-year review. *Education Information Technologies* https://doi.org/10.1007/s10639-021-10838-z

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' Theorem. *The Journal of Technology, Learning and Assessment, 1*(2), 3-21.

Ruparelia, N. B. (2010). Software development lifecycle models. ACM SIGSOFT Software Engineering Notes. *Hewlett-Packard Enterprise Services, 35*(3), 8-13. *https://orcid.org/10.1145/1764810.1764814.*

Salim, Y., Stevanus, V., Barlian, E., Sari, A. C., & Suhartono, D. (2019). Automated English Digital Essay Grader Using Machine Learning. *In 2019 IEEE International Conference on Engineering, Technology and Education* (TALE) (pp. 1–6). IEEE.

Shermis, D. M. (2014). State-of-the-art Automated essay Scoring: Competition, Results and Future Directions from a United State demonstration. *Science Direct,* 20, pp 53-76. Retrieved November, 2021 from https://assets.documentcloud.org/documents/1094637/shermis-aw fina/.pdf

Siddiqi, R., & Harrison, J. (2006). On the Automated Assessment: Short-Free Responses. 1-11. Retrieved November 3, 2021 from http://www.iea.info/docuntes/paper_2b711df83.pdf

Sowunmi, E. T. (2021). Development and Validation of Essay Test Assessor for Senior School Certificate Examination in Nigeria. *Unpublished PhD Thesis*, University of Ilorin.

Steedle, J. T., & Elliot, S. (2012). The efficacy of automated essay scoring for evaluating student responses to complex critical thinking performance tasks. *New York, NY: Council for Aid to Education*

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2003). Auto-Marking: using computational linguistics to score, free text responses. *29th Annual Conference of the international Association for Educational Assessment,* (pp. 1-5). Manchester, UK.

Valenti, S., Neri, F., & Cucchiarelli, A. (2017). An overview of current research on automated essay grading. *Journal of Information Technology Education:* Research 2: (pp. 319-330) https://orcid.org/10.28945/331.

Vantage Learning, (2002). *A study of expert scoring, standard human scoring and InelliMetric scoring accuracy for statewide eight grade writing responses* (R-726). Newtown, PA: Vantage Learning.

Wang, Q. (2022). The use of semantic similarity tools in automated content scoring of fact-based essays written by EFL learners. *Education and Information Technologies*, *27*(9), 13021 13049.

Weigle, S. C. (2010). Validation of automated scores of TOEFL IBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335-353.

Yu, J. (2018). Research process on software development model. *In IOP Conference Serie: Materials Science and Engineering, 394*(3) pp.032-045. IOP Publishing.

Zhu, W., & Sun, Y. (2020). Automated essay scoring system using multi-model Machine Learning, *david c. wyld et al. (eds):* mlnlp, bdiot, itccma, csity, dtmn, aifz, sigpro.