This is the Pre-Published Version.

The following paper W.E. Yeung, C. Qi, J.L. Xiao, F.R. Wong (2023) Evaluating the effectiveness of AI-based essay grading tools in the summative assessment of higher education, ICERI2023 Proceedings, pp. 8069-8073 is available at https://library.iated.org/view/YEUNG2023EVA.

EVALUATING THE EFFECTIVENESS OF AI-BASED ESSAY GRADING TOOLS IN THE SUMMATIVE ASSESSMENT OF HIGHER EDUCATION

W.E. Yeung, C. Qi, J. Xiao, F.R. Wong

The Hong Kong Polytechnic University (HONG KONG)

Abstract

Technology-enhanced assessment has been widely discussed in pedagogy literature. However, the study on AI assessment, especially AI assessment for summative essay grading in higher education is rare. This research-in-progress paper provides a solid theoretical foundation and study plan to explore how to assess and compare the effectiveness of various existing AI-based essay grading tools in higher education. Specifically, in our research design, reflective essay assignments from one freshmen course that involve thousands of students in university will be used as a summative assessment to realize the above research purposes. 800 student assignments with human-generated grades will be used as training data for the selected AI grading platforms. Grades of another 200 student assignments generated by the selected AI graders will be used as testing data to examine the predictive accuracy and effectiveness of these AI algorithms. The best AI platforms will be selected at the end as the ideal solution for this type of automated essay scoring in the future. Results of quantitative comparisons will further help to draw conclusions on the accuracy and reliability of the selected AI grading platforms.

Keywords: Artificial Intelligence, essay grading, summative assessment, higher education

1 INTRODUCTION

Recently, new developments in Artificial Intelligence (AI)-related educational assessment are attracting increasing interest from educators. AI-based marking or grading tools are able to greatly ease teachers' workload, especially when marking individual essays that evolve thousands of university students. Compared with other assessment methods, essays are notably time-consuming when they are marked manually. Automated grading of essays by the right AI algorithms will not only reduce the time for assessment itself, but also provide an opportunity to test the robustness of human grading itself [1].

In terms of assessment types, there are three categories of assessment: diagnostic assessment, formative assessment, and summative assessment. Diagnostic evaluations are typically short tests given at the beginning and/or end of a course that allows a teacher to gauge students' initial knowledge or summative feedback. Formative assessment is when the teacher carries out small evaluations frequently during the course to collect evidence of progress or areas of difficulty for each student. The types of assessment used here are typically low-stakes items of work such as quizzes, short writing assignments, or group work. Summative assessment is typically carried out at the end of a teaching and learning process and is thus usually seen as the means to measure "how much" a student has learned on the course or module. AI tools are not particularly useful for diagnostic assessment, but are commonly seen in formative and summative assessments.

The essence of AI in both summative and formative contexts is the concept of machine 'learning' – where the computer is 'taught' to interpret the grading patterns in past data and 'trained' to undertake predetermined actions according to those interpretations [2]. In addition, automated essay scoring (AES) is a compelling topic in learning analytics that largely uses deep learning and natural language processing (NLP) as its core technologies. Today, we find the pedagogical value of AES in higher education, especially in terms of formative and summative assessment [3].

In this paper, we aim to test the effectiveness of the existing famous AI-based essay grading tools (e.g., Intelligent Essay Assessor, Intellimetric, e-Rater, Copyleaks, progressay, and ASC) in summative assessment. In the following, we will introduce the literature on technology-enhanced assessment, AI in assessment, and algorithms of AI-based essay grading tools first. We will then elaborate on our research design and method. Lastly, the expected conclusion and directions for future research will be delivered.

2 LITERATURE REVIEW

2.1 Technology-enhanced Assessment

Technology-enhanced assessment refers to innovative assessment practices and systems that use technology to support the management and delivery of assessment [4]. Technology should enhance the design, scheduling, and delivery of assessments and provide opportunities to enhance feedback and support. There are numerous learning theories and contexts in pedagogy to discuss the necessity of using technology-enhanced assessment. The specific contexts include: online learning, mobile device usage for learning interactivity, AES for providing prompt feedback and score reporting; and has covered formative or summative assessment. Similar to many other traditional types of new technologies, AI can help generate assessment tasks, find appropriate peers to grade work, and automatically score student assignments. These techniques offload tasks from humans to AI and help to greatly improve the teachers' working efficiency and enhance students' motivation to learn [5]. The purpose of our research is to further test the effectiveness of the technology-enhanced assessment, especially that of AI in summative assessment.

2.2 AI in Assessment

2.2.1 AI in General Assessment

Assessing students with the help of AI brings benefits to not only teachers but also students. For teachers, AI assessment tools can evaluate students' performance and provide instant feedback to students. AI can also track class attendance, assignment submission status, and performance on specific tasks to help identify learning gaps or flag up worrying behavior [6]. These can definitely save teachers' time and effort to handle other mission-critical tasks. For students, learning experiences can be optimized because assessment can be done on a continuous basis and feedback can be provided immediately.

AI can vastly improve different types of assessments. There are mainly three types of assessment: summative, formative and diagnostic assessment. Summative assessment mainly focuses on measuring how much a student has learned after completing a course or a part of a course [7]. On the other hand, formative assessment is usually used during the learning process in order to monitor students' progress and provide feedback to them. It sees learning as a process and conducting such assessment allows teachers to evaluate areas of the class that need to be improved [8]. Diagnostic assessment aims to identify what students know or do not know, and it is more often done before a course of study begins. Though AI in assessments has gained popularity, its usage is still not widespread and most of the usage focuses on formative assessment [9]. The most commonly seen question types on AI-assisted assessments are multiple-choice questions and short-text responses, as these are easy for the system to grade and provide feedback to [8]. Long-text responses and essays are also found but the effectiveness and reliability of AES are still under research [10]. AI tools in assessments can also be found in summative assessment, though much less frequently than formative assessment. The most well-known AI-assisted summative assessments are GMAT and GRE general test [2]. Lastly, diagnostic assessment is not commonly discussed in association with AI, as it is sometimes seen as a part of formative assessment [11].

Lastly, the applications of AI in assessments can be found in elementary, secondary and higher education and across many different fields of study. For example, in the university context, Hooshyar et al. [12] discussed how AI is adopted in assessing the competence of students from computer programme. Aluthman [13] examined the effectiveness of automated essay scoring in an undergraduate English course. Samarakou et al. [14] proved the usefulness of AI in the continuous monitoring of engineering students' learning progress.

2.2.2 AI in Essay Assessment

With AI being the grading assistant of essay assessment, AES is possible. It saves time in grading thousands of essays [1, 15] and increases the accountability of the marking [3, 10]. The former is especially helpful for higher education as foundation courses usually have thousands of students every semester and hence many essays need to be graded by the end of the semester. The latter stems from a lot of human-related factors such as expertise level, energy level, Halo effect, leniency and inconsistency [1, 3]. Elimination of these issues would greatly benefit students as they would be able to receive more reliable scores and feedback immediately after submission.

However, AES is still undergoing improvements throughout time [10, 16]. Experts are still testing and evaluating which form of modelling would generate better performance [17]. Although encouraging results have been found recently, the technology behind AES is not transparent enough to gain human trust on AI. Opposing voices criticizes on how computers can only search for patterns in texts [18] and can never comprehend inherent meaning as they are not responsive to feelings by definition [1, 19]. Even if the scores produced by human rater and AI rater are statistically similar, the underlying meaning behind the scores are different. While evaluating what role AI is playing in educational assessment, Gardner et al. [2] conclude that the time when AES system will be able to operate on a par with human judges remains a long way off. On top of that, it is generally agreed that AES cannot assess writing that involves creativity and higher-order thinking [2] and hence limiting its application to the higher education sector.

AI or AES has been effective in both summative and formative types of essay assessment. For example, Gardner et al. [2] believed that AES is becoming a very sophisticated tool for grading essay writing in large-scale summative testing programmes, as well as formative assessments where timely feedback from AES is continuously generated via the whole learning process. In summative grading, AES will usually generate an auto score to mimic human's logic, though some researcher would find it inconsistent from the original intention of human marker. In general, compared with those of AES in formative assessment, the literature on AES in long-text summative essay scoring is rare.

What is more, limited research is done on using AES in higher education institution (HEI) [20] and the ones done in the HEI context are for English learning purposes (e.g., [13, 21]). Comprehensive tasks aim to enhance students' skills such as rhetorical knowledge, critical thinking, openness and creativity, which are exactly what writings should possess in higher education. Gardner et al [2] believe that thought there is a significant gap between grading for grammaticality and for more complex and comprehensive tasks, little research has explored the possibility of using AES on summative essay for higher education courses other than language related subjects.

2.3 Algorithms of AI-based Essay Grading Tools

AES can be actualized mainly relying upon two technological advancements in AI - Natural Language Processing (NLP) and Machine Learning (ML). Today, there are three popular commercial AES platforms: Intelligent Essay Assessor (IEA), Intellimetric, and e-Rater [2]. Based on NLP and ML, these three platforms have their own ways of achieving the aim of grading essays. IEA uses latent semantic analysis as an indexing programming tool to look for features such as word variety, grammar, and text complexity in the essays. Intellimetric's system is neurosynthetic which uses deep learning to mimic how human's neuro system works. Besides, NLP, CogniSearch, and other statistical methods such as linear regression and Bayesian analysis are all algorithms used by Intellimetric. Educational Testing Service's e-rater uses multiple regression and NLP. Many research papers have confirmed the reliability of these platforms, and the correlation with human raters and the three platforms are above 0.83 in general [17]. Besides NLP and ML, deep learning and random forest regression are also newer machine learning tools in AES [3].

The purpose of this study is to evaluate the effectiveness of the existing AI-based essay grading tools in the summative assessment of higher education. As summarized above, most of the prior studies have discussed about technology-enhanced assessment in general, AI in formative or diagnostic essay assessment in primary or secondary education. In terms of summative assessment, the literature is lacking a systematic testing and comparison of the effectiveness of the existing commercial AES tools for summative tasks that involve self-reflection and critical thinking. This research therefore aims to fill this by emphasizing our pioneer role in evaluating the effectiveness of several AES platforms on summative essay assessment in higher education.

3 RESEARCH METHODS

3.1 Design

To test the effectiveness of the existing AI-based essay grading tools, 1000 essays with human-graded score will be used for the selected AI grading platforms. The first 800 data points will be used as training data, and another 200 will be used as testing data. Grades generated by various selected AI platforms will be compared with the grades generated by the human graders to indicate the accuracy and reliability of each tested AI grading platforms. The common AI-based essay grading tools currently available include IEA, IntelliMetric, e-Rater, Progressay, Copyleaks, and ASC. Four out of six platforms will be included in this project. The selection is based on the relevancy of AI algorithms, functions, and accessibility.

3.2 Sample and Procedural

Student assignments from a freshman course on innovation and entrepreneurship will be our target sample. The assignment is a summative assessment, as the grade given to the students is the final grade for this single item of assessment, and the assessment happens at the end of the semester. The assignment requires students to write a 600–800-word essay to reflect on what they have learned in the subject and their future plans. Reflective content, citation and reference, and writing language are major assessment criteria for this assignment. Clear rubrics have been established beforehand, and human graders will grade based on the rubrics.

Students are invited to participate in the study voluntarily and their joining of the study will not affect their assignment grade, which will be assessed by human graders. By the end of the semester, 1000 students are estimated to join the study, and their assignments will be used to train and test the selected AI grading platforms. 800 assignments together with the human-generated grade will be randomly selected and used as the training data. The remaining 200 assignments will be used as testing data to assess the accuracy and reliability of the AI graders. Based on platform requirements/functions, the assessment rubrics and marking rules will be entered into each AI grader platform. Finally, we are going to compare the accuracy and reliability of the grades generated by different platforms, and select the one with highest prediction rate for future practice or further research. Statistics such as correlation and T-test will also be used as facilitating statistics to test the differences between platforms.

4 CONCLUSIONS AND FUTURE RESEARCH

AI, as a new type of technology-enhanced assessment tool is gaining popularity in higher education. This study aims to test the grading abilities of various AI grading platforms in higher education. We are among the pioneer studies to explore the effectiveness of using the existing commercial platforms in summative AES, especially in higher education. In our study context, summative essay assignments of university students will be used as the data set. By comparing the grades generated by the AI platforms with those given by the subject instructors, we will draw conclusions about the reliability and accuracy of each AI grading platform and choose the most accurate one for further pedagogical practice.

REFERENCES

[1] V. V. Ramalingam, A. Pandian, P. Chetry, & H. Nigam, “Automated essay grading using machine learning algorithm.” In *Journal of Physics*: Conference Series, vol. 1000, no. 1, pp. 012030, IOP Publishing, 2018.

[2] J. Gardner, M. O'Leary, & L. Yuan, “Artificial intelligence in educational assessment:‘Breakthrough? Or buncombe and ballyhoo?’” *Journal of Computer Assisted Learning*, vol. 37, no. 5, pp. 1207-1216, 2021.

[3] V. Kumar, & D. Boulanger, “Explainable automated essay scoring: Deep learning really has pedagogical value”, *Frontiers in education,* vol. 5, pp. 572367, Frontiers Media SA, 2020.

[4] M. O'Leary, D. Scully, A. Karakolidis, & V. Pitsia, “The state-of-the-art in digital technology-based assessment.” *European Journal of Education*, vol. 53, no. 2, pp.160-175, 2018.

[5] Z. Swiecki, H. Khosravi, G. Chen, et al., “Assessment in the age of artificial intelligence, *Computers and Education: Artificial Intelligence*, vol. 3, pp.100075, 2022.

[6] H. Hooper, “Benefits of AI in Education, with Examples.” 2023. https://virtualspeech.com/blog/benefits-ai-education

[7] D. D. Dixson & F. C. Worrell ”Formative and summative assessment in the classroom.” *Theory into practice*, vol. 55, no. 2, pp. 153-159, 2016.

[8] C. N. Blundell, “Teacher use of digital technologies for school-based assessment: a scoping review.” *Assessment in Education: Principles, Policy & Practice*, vol. 28, no. 3, pp. 279-300, 2021.

[9] V. González-Calatayud, P. Prendes-Espinosa, & R. Roig-Vila, “Artificial intelligence for student assessment: A systematic review.” *Applied Sciences*, vol. 11, no. 12, pp. 5467. 2021.

[10] M. Z. A. B. Azahar, & K. I. B. Ghauth, “A Hybrid Automated Essay Scoring Using NLP and Random Forest Regression.” *International Conference on Computer, Information Technology and Intelligent Computing (CITIC 2022)*, pp. 448-457, Atlantis Press, 2022.

[11] B. Csapó, & G. Molnár, “Online diagnostic assessment in support of personalized teaching and learning: The eDia system.” *Frontiers in Psychology*, vol. 10, pp. 1522, 2019.

[12] D. Hooshyar, R. B. Ahmad, M. Yousefi, M. Fathi, S. J. Horng, & H. Lim “Applying an online game-based formative assessment in a flowchart-based intelligent tutoring system for improving problem-solving skills.” *Computers & Education*, vol. 94, pp. 18-36, 2016

[13] E. S. Aluthman, “The effect of using automated essay evaluation on ESL undergraduate students’ writing skill.” *International Journal of English Linguistics*, vol. 6, no. 5, pp. 54-67, 2016.

[14] M. Samarakou, E. D. Fylladitakis, D. Karolidis, W. G. Früh, A. Hatziapostolou, S. S. Athinaios, & M. Grigoriadou, “Evaluation of an intelligent open learning system for engineering education.” *Knowledge Management & E-Learning*, vol. 8, no. 3, pp. 496, 2016.

[15] A. Vista, E. Care, & P. Griffin, “A new approach towards marking large-scale complex assessments: Developing a distributed marking system that uses an automatically scaffolding and rubric-targeted interface for guided peer-review.” *Assessing Writing*, vol. 24, pp. 1-15, 2015.

[16] R. Yang, J. Cao, Z. Wen, Y. Wu, & X. He, “Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking.” *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1560-1569, 2020.

[17] M. A. Hussein, H. Hassan, & M. Nassef, “Automated language essay scoring systems: A literature review.” *PeerJ Computer Science*, vol. 5, e208, 2019.

[18] P. Deane, “On the relation between automated essay scoring and modern views of the writing construct.” *Assessing Writing*, vol. 18, no. 1, pp. 7-24, 2013.

[19] L. Perelman, “When “the state of the art” is counting words.” *Assessing Writing*, vol. 21, pp. 104-111, 2014.

[20] F. Ouyang, L. Zheng & P. Jiao, “Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020.” *Education and Information Technologies*, vol. 27, no. 6, pp. 7893-7925, 2022.

[21] S. Somasundaran, C. Lee, M. Chodorow & X. Wang, “Automated scoring of picture-based story narration.” *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pp. 42-48, 2015.