

Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool

Isaiah T. Awidi^{a, b}

^a University of Southern Queensland, Qld, Australia

^b Adjunct, NMSW, University of Queensland, St. Lucia, Qld, Australia

1. Introduction

Artificial Intelligence (AI) has become a debatable topic within educational communities, regarding its integration in teaching and learning. While academic institutions worldwide engage in discussions about AI's role in assessment, the focus often centres on concerns about academic integrity, rather than its ability to improve the reliability of the assessment process (as compared with human intelligence in performance assessment and evaluation). However, some universities in the UK and Australia, are exploring AI's potential to enhance alternative forms of assessment and reliability through grading and personalised feedback. These institutions are carefully considering the ethical, philosophical, and legal implications of AI integration in learning and teaching. Conversely, others hesitate to adopt AI in assessments, opting to preserve the conventional and tried traditional methods due to concerns about academic integrity.

AI is a simulation of human intelligence processes by a digital computer system to perform human intelligence tasks associated with decision-making, learning, visual perception, speech recognition, adaptation, sensory understanding, and interaction (Bishop & Nasrabadi, 2006; Russell, 2010; Russell & Norvig, 2020). Over the past decades, educational researchers have documented challenges they faced in developing automated essay grading (AES) systems (Ramesh & Sanampudi, 2022; Shermis & Burstein, 2003). Despite rigorous efforts, these systems encountered several limitations. Ramesh and Sanampudi (2022) noted key issues including difficulties in converting sentences into vector form, a crucial step in feature extraction, and machine learning model training. Additionally, current AES systems fail to assess the overall completeness of an essay and do not provide personalised feedback on student responses. They also struggled to identify coherence within essays. Furthermore, these systems were challenged by irrelevant or adversarial student responses, raising questions about their effectiveness. While progress was being made, significant hurdles remained in creating fully effective and reliable automated essay grading systems

(Shermis & Burstein, 2003).

However, with the rapid evolution of technology and the sophistication of machine learning, generative AI is capable of recognizing language and text. It is therefore important to explore ways by which generative AI can improve the quality of assessment, academic integrity, turnaround time for assessment, or save money for schools and faculty, particularly in the areas of marking, grading, and providing formative feedback on students' written assessments. Currently, little is known about AI's use to improve the reliability of the assessment process including grading of students' written assessment. In addition, the impact on written assessment grading, outside of the traditional human marking of written assessment normally has subjective and inconsistent effects in grading. The inherent subjectivity of the human marker process is a significant challenge, where variations in grades assigned to students can occur in large classes, particularly where multiple markers are involved (Haines, 2021; Hounsell, 1995). Marking written forms of assessments can be expensive and time-consuming, especially where examiners have multiple classes to teach. Lack of standardization in grading criteria and difficulty in providing feedback further compound the challenges (Boud, 2007). Hence, the question, "Can AI tools like ChatGPT mark, assign a grade/score, and provide personalised feedback on reflective essays effectively?"

The study's primary objective was to evaluate the effectiveness of generative AI tools, like ChatGPT, in grading reflective essays, focusing on whether they offered faster, more consistent, and objective assessments compared to human grading. It aimed to identify the challenges and limitations of using AI for grading, contributing to the ongoing debate about AI's role in education and its potential to improve the consistency of student assessments. Additionally, the study explored the implications of AI in education, highlighting the need for educators to understand AI's role and adapt to evolving teaching and learning needs while considering the impact of immediate feedback on student learning and examiner focus on personalised feedback. The structure of this article takes a conventional form, which begins with an introduction and

E-mail addresses: isaiah.awidi@unisoq.edu.au, i.awidi@uq.edu.au, atisaiah@yahoo.com.

<https://doi.org/10.1016/j.caeai.2024.100226>

Received 26 September 2023; Received in revised form 27 February 2024; Accepted 20 April 2024

Available online 25 April 2024

2666-920X/© 2024 The Author. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

literature review, followed by the methodology, results and findings, discussion, and conclusion.

2. Literature review

Assessment and evaluation in education are essential for understanding student progress and learning outcomes (Earl & Katz, 2006). It serves as a means of guiding both students learning and instructional effectiveness. Through thoughtful assessment practices, examiners can gain insight into the depth of students' understanding, identify areas of strength and weakness, and structure instructions to meet their diverse needs. Formative assessments, such as class discussions, quizzes, and peer reviews, provide real-time feedback that informs examiners' decisions and supports students' growth. Summative assessments, such as exams and projects, offer a snapshot of students' achievement at the end of learning units or courses. Meanwhile, evaluation extends beyond individual performance, to include teaching strategies, curriculum design, and institutional goals. Thus, by analysing assessment data and reflecting on teaching practices, examiners can refine their approaches, enhance students' engagement through valuable feedback, and develop an active learning environment that encourages continuous improvement of learning and personal growth. Effective constructive feedback that is timely should relate directly to the assignment, appropriate, consistent, and clear in its guidance for improvement (Wiggins, 2012; Brookhart, 2017; Winstone & Carless, 2019). Contemporary assessment practices focus not only on what students have learned but also on the authenticity of their learning. This approach encourages students to demonstrate understanding and competencies in their learning activities (Wiggins, 1990). The assessment process incorporates observing, describing, collecting, recording, scoring, and interpreting student learning. This may take the form of written assessments, and reflective writing among other forms of assessments. Writing essays in higher education provides students with the opportunity to develop various competencies and skills, such as critical thinking. This requires students to analyse and interpret information before presenting their arguments (Fitzgerald, 1994; Warburton, 2020). By doing so, students can enhance their communication skills, as they need to clearly articulate their thoughts and ideas logically and coherently. They can also improve their knowledge and understanding of the topic and develop writing proficiency (Fitzgerald, 1994).

The theoretical perspective on short essay forms of assessment varies depending on the specific context and purpose. Several authors have described and justified why short forms of assessments are relevant for students learning. For example, Sweller (2003) draws on cognitive and social learning theories to support the argument for using shorter writing assignments as a more effective way to promote student learning and engagement. Chamberlain et al. (2004) also argued that short-answer questions can promote higher-order thinking and support student learning of disciplinary concepts and skills by breaking down final summative assessment tasks into smaller, manageable tasks. Thus, depending on the design, short-answer questions with well-designed rubrics can assess higher-order thinking skills, such as the ability to analyse, synthesise, and evaluate. When used effectively, they can also enhance the retention of information and promote deeper learning. Supporting the argument with theory, Tindall-Ford and Sweller (2020) suggest that cognitive load theory can inform the design of short-answer questions, while the schema theory can explain their effectiveness. They opine that short-answer questions can require students to monitor and regulate their learning, leading to improved metacognitive skills. By encouraging students to reflect on their thought processes, short-answer questions can promote self-regulated learning. In addition, short-answer questions can be more efficient than essay questions, requiring less time to grade and providing more detailed feedback to students. The diverse theoretical perspectives that underpin the use of short essay forms of assessment demonstrate the importance of aligning assessment formats with specific learning objectives and theories of learning (Bunch &

Cizek, 2007). These benefits justify the need to assign students short reflective essays, along with rubrics to guide their writing.

The marking of written essays continues to pose significant problems for both students and examiners, especially in large classes. Zak and Weaver (1998) discussed the challenges and possibilities of grading writing, including the difficulties of creating fair and accurate assessments. Grading written assignments is a complex task that requires a range of skills and knowledge. According to Zak & Weaver, it involves not only assessing the content of the assignment but also evaluating the structure, coherence, and clarity of the writing. Effective grading also requires knowledge of relevant standards and criteria, as well as the ability to provide specific, actionable, and timely feedback. Examiners also face the challenge of managing workload, ensuring fairness and consistency, dealing with student plagiarism, and providing effective and personalised feedback. Therefore, in light of AI integration in teaching, learning, and assessment (Krendl & Lieberman, 1988; Luckin et al., 2022; Xia et al., 2022), it is essential to investigate how these technologies can enhance the scoring of short reflective essays. It is expected that the benefits of short reflective essays can be achieved when the scripts submitted by students are effectively marked and graded with consistent and objective formative feedback. With good prompts, AI's capabilities can help address the limitations that may arise in this process.

2.1. The rationale and potential of AI for marking written assessments

Although the examiner challenges as observed can be addressed through professional learning opportunities that focus on improving markers' skills and knowledge in these areas, the emergence of AI tools like ChatGPT and other AI tools has proven that Machine Learning Systems (AI) can be trained or prompted to perform grading tasks efficiently (Kalervo et al., 2022; Swiecki et al., 2022) and can write reflectively (Li et al., 2023), just like Human Markers (Expert Tutors). While professional learning can support teachers in improving their grading practices, developing AI systems can also be effective in supporting examiners to focus more on providing personalised feedback to improve more efficiently on their practice (Maier & Klotz, 2022) where the AI is unable to do that well. Examiners can use the personalised feedback to actively engage with struggling students and provide them with the needed support to make progress in their learning. Santamaría Lancho et al. (2018) explored the use of semantic technologies to provide formative assessment and personalised feedback in online courses. The authors introduced an automatic assessment tool and observed that the tool provided personalised feedback that allowed the students to improve their answers and writing skills, leading to a better understanding of concepts and knowledge building. They proposed that the tool offers enriched and personalised feedback which proved to be entirely satisfactory for the students. Hence, AI grading systems can be useful in providing opportunities for formative assessment (Santamaría Lancho et al., 2018), where students can receive formative feedback to improve their writing skills, and performance and increase their motivation and engagement (Gibbs & Simpson, 2005).

2.2. Impact of AI on higher education delivery

Some contemporary studies have shown that the potential benefits of AI in higher education extend beyond the system's ability to enhance the quality of higher education delivery (Al Darayseh, 2023), automate assessment (Yildirim-Erbaşlı & Bulut, 2023), improve students' learning outcomes, and streamline administrative tasks (Khosravi et al., 2022; Ouyang et al., 2022). Educational research has shown that AI has the potential to personalise learning by providing adaptive feedback in real-time, tracking students' progress, and creating tailored learning experiences for individual students based on their learning pace, preferences, and learning styles (Chen et al., 2020; Fadel et al., 2019; Holmes et al., 2020; Southworth et al., 2023). According to

Zawacki-Richter et al. (2019), AI can automate grading and assessment tasks, reducing faculty workload and improving grading accuracy, despite some ethical and reliability concerns raised (Luckin et al., 2022; Munir et al., 2022). By demonstrating the potential to enhance learning outcomes by providing personalised and adaptive learning experiences that are tailored to individual students' needs and preferences (Chamberlain et al., 2004; Lee et al., 2022; Van der Kleij et al., 2015), examiners can use AI resources to support education delivery and students learning experiences. An example of such an AI tool is ChatGPT.

The OpenAI model ChatGPT, developed using deep learning, is capable of generating, classifying, and summarizing text with high coherence and accuracy, demonstrating broad knowledge and domain expertise. In a 2023 study by Li et al., ChatGPT was employed to generate reflective responses for pharmacy course assignments, using nine different prompting strategies. The study found that ChatGPT's responses were of higher quality in all six assessment criteria than those written by students. Additionally, it explored deep learning classification methods to differentiate between student-written responses and those generated by ChatGPT. A domain-specific BERT-based classifier was able to effectively distinguish between the two, outperforming experienced teaching staff and a general-domain classifier, even when the model was tested with unknown prompts.

2.3. Some early research on AI for marking/scoring written assessments

The effectiveness of human marking of essays compared to AI marking of students' assessments continues to generate debate among educational researchers. According to Shermis (2010, 2022), AI systems are reliable and consistent in scoring essays, with inter-rater agreement levels comparable to those of human graders. They can also score essays more quickly and efficiently than human graders, reducing the time and cost associated with large-scale essay grading. AI systems can provide valuable diagnostic feedback to students, helping them to identify areas of strength and weakness in their writing and improve their skills. Current research and trends for the future suggest that AI systems can assess a range of skills and knowledge, including language proficiency, critical thinking, and problem-solving, and provide personalised feedback to learners (Al Ghatrifi et al., 2023; Joksimovic et al., 2023; Xia et al., 2022; Yildirim-Erbasli & Bulut, 2023). However, Expert Tutors have a greater ability to assess the overall quality of an essay. There is an assumption that, while AI scoring systems can be highly effective at measuring certain aspects of writing, such as grammar and syntax, they may struggle to evaluate the overall coherence, organization, and logic of an essay (Elliot & Klobucar, 2013). These complex aspects of writing require a nuanced understanding of the content and context of the essay, something that humans are generally better able to provide. According to Elliot and Klobucar (2013), Expert Tutors can recognize and evaluate creativity and originality, maintaining that AI scoring systems may struggle to recognize and evaluate subjective aspects of writing, such as creativity and originality. These qualities can be important in certain types of writing, such as poetry or creative writing, and may be more difficult for AI systems to evaluate. Moreover, while Expert Tutors can provide detailed and personalised feedback tailored to the needs of each student, recent AI systems like ChatGPT have proven to interact effectively with feedback. On the other hand, AI scoring systems may be prone to certain types of errors that Expert Tutors are better able to identify and correct (Williamson et al., 2012). AI systems may struggle to recognize sarcasm or irony in an essay, leading to an incorrect score, while Expert Tutors are better able to recognize and correct these types of errors.

These limitations on AI systems for scoring written assignments justify for this study to identify meaningful ways of improving machine learning algorithms to improve AI scoring. Shermis et al. (2016) discuss the potential of automated writing evaluation (AWE) and artificial intelligence (AI) technologies. The authors suggest that AWE has the potential to revolutionize writing assessment and instruction by providing

more timely, objective, and consistent feedback to students. AWE can provide immediate feedback on a range of writing features, such as grammar, mechanics, style, and organization, and it can also assist with higher-order skills such as critical thinking and argumentation. Furthermore, the authors suggest that AWE has the potential to improve writing instruction by providing teachers with insights into students' strengths and weaknesses, allowing them to tailor their instruction to better meet students' needs. Regarding the potential of AI technologies, the authors note that ongoing developments in machine learning, natural language processing, and other areas have the potential to make AWE systems even more effective and reliable. Future AWE systems may be able to provide even more sophisticated feedback, such as identifying patterns of errors across a student's writing and providing targeted practice exercises to address those errors. The authors were of the view that AWE and AI have significant potential to improve writing instruction and assessment, and ongoing research and development in these areas will continue to yield new and innovative tools for educators and students alike. However, as noted by Foltz (2020) practical considerations must be taken into account when using AI models for automated scoring of writing. This would require collaboration and dialogue among experts from various institutions and fields (AI developers and Educators) to ensure transparency, high-quality data training, and examiner preparation, as well as ethical considerations, in order to foster interdisciplinary harmony acceptable by educators.

The ongoing research on AI scoring of written assessments continues to yield results that are comparable to human grading in assessing the quality of essays (Mizumoto & Eguchi, 2023; Richardson & Clesham, 2021; Zawacki-Richter et al., 2019). These studies have demonstrated that, in addition to their efficiency in grading and time-saving benefits, especially in large-scale assessments, automated systems also possess the potential to offer more consistent and objective evaluations of essays. They can help avoid potential biases that may arise from human marking and provide immediate feedback to students to allow them to improve their learning, writing skills, and knowledge of the subject matter (Shermis, 2022; Stephen et al., 2021; Zhang, 2021). Achieving these goals can also reduce the cost of assessment. However, it's essential to further research these aspects to enhance the reliability of these tools' use. When proven to provide quick, accurate, and objective assessment feedback through research, examiners can focus more on course design and effective ways of writing assessments to make learning authentic and more meaningful.

2.4. Issues and challenges of AI marking/scoring assessments

The benefits of using AI for marking notwithstanding, there are several challenges associated with this approach, including technical issues, ethical considerations, and the need for transparency and fairness in the automated assessment process (Celik et al., 2022). One of the main challenges is that some AI systems have limited ability to understand the meaning and nuance of text, which can restrict their ability to evaluate more complex aspects of writing and result in biased evaluations based on the data they are trained on (Burstein et al., 2004). Additionally, there is a risk that educators and students may become overly reliant on automated evaluation systems, which can lead to a devaluation of human expertise in writing assessments and ethical concerns around issues of privacy and data security (Burstein et al., 2004). Several studies have highlighted the challenges and limitations of using AI for marking essays, indicating that the effectiveness of AI-based assessment tools depends on various factors, such as the quality of data and assessment rubrics. Technical issues, including the need for high-quality data and models, and ethical considerations, such as privacy, transparency, and fairness, are also highlighted. The subjectivity of writing and the potential for biases and errors in AES systems are additional concerns. Improving on the systems would require developing more sophisticated models, incorporating human marking, using a wider range of features, and focusing on developing more nuanced features.

3. Methodology and methods

The research approach for evaluating ChatGPT marking compared to Human Marking (Expert Tutor) was determined based on the specific context, research questions, and study goals. A predominantly objective (quantitative) experimental design approach was adopted (Bruscia, 2016). This approach allows researchers and educators to assess AI’s ability to evaluate reflective assessments objectively, without biases, ensuring consistent and fair evaluation. Particularly relevant for assignments involving subjective opinions (reflections), this study explores AI’s efficiency, consistency, scalability, and its capacity to provide personalised feedback to students compared to human marking. The following hypotheses were therefore used to guide the research methods and analysis to address the research question “Can Artificial Intelligent tools like ChatGPT effectively mark/score and provide personalised feedback on reflective essays comparable to Human Marking (Expert Tutors)?”:

- HQ1. There is a consistent pattern in scores and feedback provided by Expert Tutors (ET) and ChatGPT across all scripts.
- HQ2. There is no significant difference between scores assigned by Expert Tutors and ChatGPT.
- HQ3. ChatGPT demonstrates consistent and effective scoring comparable to Expert Tutors.

3.1. Context

This study is part of a Teaching and Learning grant project titled “Artificial Intelligence for Reflective Marking. How effective is it?”, It aimed to compare current AI marking systems with Expert Tutors in assessing written reflective assessments in First Year Engineering courses at an Australian University. These courses, including Engineering Design, Engineering Modelling & Problem Solving, and Engineering Design, Modelling & Problem Solving, are large-scale, project-based STEM courses with diverse student cohorts. The chosen assessment item was a short (400 words) reflection piece, typically marked reliably by Tutors without needing expert engineering knowledge. Reflections served as a suitable “entry-level” test for AI marking in STEM written assessments. Archived reflections, along with Tutor marking rubrics and comments, provided a comprehensive dataset for both training and testing AI programs. These programs were trained using reflections and associated rubrics as input and then utilised for marking reflections independently. Detailed comparison of Tutor and AI marks aimed to

assess the AI program’s performance and highlight potential human limitations. This research was expected to yield insights into AI marking and Tutor marking methodologies. The dataset from this completed project was used for this study.

3.2. Adopting ChatGPT for reflective essay marking (scoring) process

The study employed several key steps to evaluate ChatGPT’s effectiveness and accuracy in grading the reflective essays (details in Fig. 1). They include:

1. Inputting Developed Assessment Instructions and Rubrics: ChatGPT was provided with specific guidelines (task prompts) and evaluation criteria (rubrics) to learn and memorize the assessment parameters.
2. Testing ChatGPT’s Understanding: The AI’s comprehension and ability to link instructions to assessment standards were verified by summarizing the instructions and their connection to the rubric criteria.
3. Creating and Using a Graded Essay Dataset: Reflective essays previously graded by Expert Tutors were used to pre-test ChatGPT’s grading abilities, assessing alignment with human judgment.
4. Feature Extraction and Verification: Key essay features like “Depth,” “Analysis,” and “Clarity/Logic of Writing) were extracted to verify ChatGPT’s assessment accuracy, focusing on the rubric criteria.
5. Ensuring Prompt Consistency: Consistent prompts for grading were maintained to evaluate the reliability and uniformity of ChatGPT’s scoring.
6. Guidance and Optimizing ChatGPT: The ChatGPT was guided with a small dataset of scripts, adjusting prompt parameters to improve grade differentiation (not machine learning).
7. Independent Scoring and Comparison: ChatGPT and expert tutors independently scored the same set of essay scripts in a randomized order to minimize bias, directly comparing grading performance.

The study’s focus on using separate, unlearned scripts for the final evaluation tested ChatGPT’s generalization capabilities and allowed for a fair comparison with human marking, particularly in a STEM context. This approach highlighted AI’s potential in reflective assessment while addressing the limitations of human marking.

3.3. Data source

The study used a sample of 108 written reflection assessment data from a 2018 archive, consisting of 117 reflective assessments selected

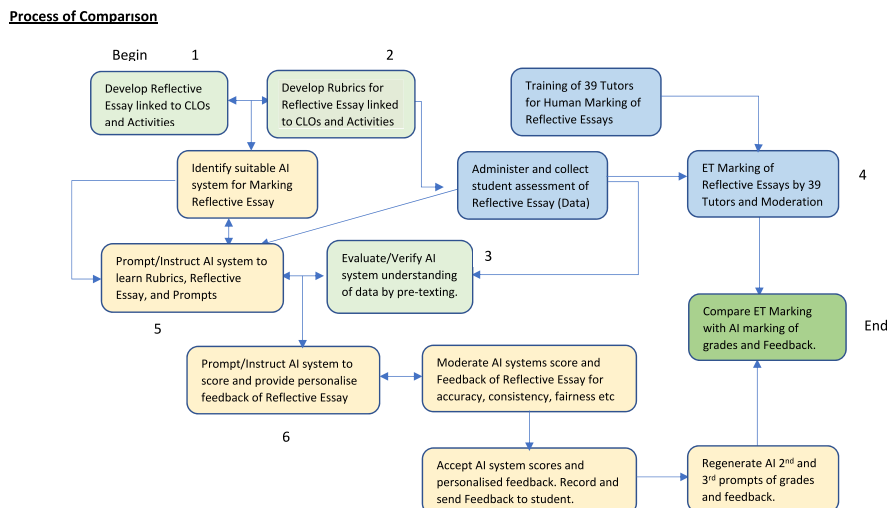


Fig. 1. Process of comparing expert tutor (ET) evaluation of reflective essays with marking of generative AI tool.

from a total of 766 submitted scripts (datasets). These assessments are well-documented and structured as scaffolded tasks for students, with clear and detailed rubrics. Marking conducted consistently over several years, was carried out by 39 trained tutors, each evaluating 16 scripts, with two tutors assessing an average of 32 scripts collaboratively. All student submissions underwent independent moderation and thorough marking. A subset of moderated scripts, along with the rubrics, was used to pre-test ChatGPT's ability to score independently and provide personalised feedback, without relying on a large dataset for learning assessment and feedback provision. These reflective assessments were deemed suitable for evaluating ChatGPT's marking capabilities, providing the basis for estimating the parameters measured in the sample data used in the study.

3.4. Tools and data collection

The course examiner and project lead de-identified all data points (scripts), leaving only responses and feedback in the essay scripts. The scripts were serially numbered and regrouped by the 39 Expert Tutor markers before being passed to the researcher for use in a case study. Using Microsoft Excel's Data Analysis Statistical Tool, a random sample of 117 scripts was selected from the 766 grouped by tutors. Three scripts were randomly chosen from each of the 39 tutor groups. Each script categorized into final grades 1 through 4 was randomly numbered without a predictable pattern to ensure impartial selection. This process aimed to equally distribute potential biases across all scripts, ensuring impartial scoring. Nine incomplete scripts from the sampled 117 were excluded from the analysis dataset, leaving 108 scripts.

3.5. Data analysis

The analysis considered various parameters, including the rubric criteria for "Depth," "Analysis," "Logic and Clarity of Writing," and students' "Feedback" on their reflections. A reliability analysis assessed the consistency of Expert Tutors (ET) marking against the criteria. Cronbach's Alpha for all 766 scripts (across all the parameters) was 0.728, indicating acceptable consistency in measuring the intended constructs. Standardized items yielded an Alpha value of 0.741. For the 108-script sample dataset, Alpha values ranged between 0.694 and 0.775 (ChatGPT scores) and 0.687–0.713 (ET scores), signifying acceptable internal consistency. Computed values for ChatGPT score (G1) and regenerated score (RG2) showed Cronbach's Alpha Coefficients of 0.775 (0.853) and 0.819 (0.905) respectively, indicating good internal consistency relative to ET scores. The standardised items in brackets showed slightly higher internal consistency values compared with the measures for the Expert Tutors scores for the population (766) and sample (108). Thus, both ET and ChatGPT measured similar constructs, forming a good basis for the analysis.

Inter-rater reliability using Cohen's Kappa or intra-class correlation coefficient was used to measure the agreement among the 39 Expert Tutors scoring of the reflective essay scripts. This provided a baseline measure of reliability to ensure that the test scores reflect more than just random errors (providing some understanding of the ETs scoring of the four rubric criteria). All 766 (population) and 108 (sample) valid assessment scripts were considered in estimating the parameters to be measured. Descriptive statistic: Mean score, standard deviation, and Pearson's correlation coefficients were used to compare Expert Tutor and ChatGPT scores aiding in understanding agreement. Statistical tests, such as Paired T-tests or Wilcoxon signed-rank test, compared Expert Tutor and ChatGPT scores to determine significance, with effect sizes calculated using Cohen's *d* or Hedges' *g*, addressing research Hypotheses 1 and 2. This provided insight into the magnitude of any observed differences.

An error/sensitivity analysis was done to identify AI (ChatGPT) scoring errors compared to Expert Tutors, examining varying scores and feedback against rubrics. Effects of prompt variations on ChatGPT

scoring were evaluated, aiding in understanding scoring stability and generalisability, addressing Hypothesis 3 - Expert Tutors' objectivity and accuracy compared to ChatGPT.

3.6. Ethics considerations

The university's human research ethics office approved the project (initially won as a teaching and learning research grant). Prior to the researcher's access, the project lead and program coordinator ensured that all students' data were de-identified, and scripts were serially numbered. The study therefore used anonymous submissions, markers' comments, feedback prompts, and awarded marks for the analysis. The study also considered written comments on "Depth," "Analysis," and "Clarity of Writing" provided independently by 39 markers. All feedback comments were tagged based on rubrics and categorized accordingly.

4. Results and findings

4.1. Reliability or consistency of scoring methods

HQ1: There is a consistent pattern in scores and feedback provided by Expert Tutors (ET) and ChatGPT across all scripts.

The Cohen's Kappa Intraclass Correlation Coefficient (ICC) was used to measure the reliability or consistency of measurements on the same scripts by both ET and ChatGPT, as depicted in [Table 1](#). Two scenarios were considered: Single Measures and Average Measures. For Single Measures, the ICC value was 0.349, suggesting about 34.9% of the variability in measurements can be attributed to differences between ET and ChatGPT scores, with a 95% Confidence Interval between 0.315 and 0.385. This value indicates a low or weak agreement (an ICC value between 0.5 and 0.75 are considered moderate agreement or consistency among the measurements). The F-test value (3.680) with degrees of freedom ($df_1 = 765$, $df_2 = 3060$), and significance level (p -value) of 0.000, reveals statistically significant agreement ($p < 0.000$), although weak to moderate. In contrast, the ICC for Average Measures was 0.728, suggesting about 72.8% of the variability in measurements can be attributed to differences between ET and ChatGPT scores, with a 95% Confidence Interval between 0.697 and 0.758, suggesting moderate agreement (consistency in ET and ChatGPT measurements). The F-test, degrees of freedom, and significance levels all confirm statistical significance. Both ICC values for Single and Average Measures are statistically significant ($p < 0.001$). The substantial 72.8% agreement in the Average Measure indicates significant agreement between ET and ChatGPT measurements. Narrower confidence intervals for Average Measures suggest higher precision, reflecting a reasonable level of reliability and consistency in measurements, with a higher ICC value indicating good agreement.

4.2. Comparison metrics: measuring expert tutor and ChatGPT scores

HQ2: There is no significant difference between scores assigned by Expert Tutors and ChatGPT (AI).

This hypothesis aimed to compare the level of agreement or disparity between Expert Tutors (ET) and ChatGPT-G1 and ChatGPT-RG2 prompt scores using provided assessment criteria and rubrics. There are two parts. [Table 2](#) illustrates the mean distribution of the scores (Part 1: Descriptive Statistics) and [Tables 3 and 4](#), include correlation coefficients between ET and ChatGPT scores (Part 2: Inference Statistics). In comparing the Total grades, ET scored a mean of 3.296, suggesting the most comprehensive assessment, while ChatGPT-G1 scored the lowest ($M = 2.069$), and ChatGPT-RG2 ($M = 2.274$) fell in between. ET generally indicated better performance than ChatGPT-G1. Variability in the scores generated by all three approaches, confirmed by Standard Deviation (SD) ([Table 2](#)), was less in ChatGPT-G1 and ChatGPT-RG2

Table 1
Interclass correlation coefficient.

Intraclass Correlation Coefficient							
Measures	Intraclass Correlation ^b	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	0.349 ^a	0.315	0.385	3.680	765	3060	0.000
Average Measures	0.728 ^c	0.697	0.758	3.680	765	3060	0.000

Note.
Two-way mixed effects model where people effects are random and measures effects are fixed.
^a The estimator is the same, whether the interaction effect is present or not. ^b Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance. ^c This estimate is computed assuming the interaction effect is absent because it is not estimable otherwise.

Table 2
Descriptive Statistics (Population = 766 scripts; ET/ChaGPT sample = 108 scripts).

Assessment Criteria Score	ET - Population		ET - Sample		ChatGPT-G1 Score		ChatGPT-RG2 Score	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Depth	0.819	0.247	0.829	0.238	0.569	0.269	0.606	0.256
Analysis	0.838	0.238	0.898	0.202	0.477	0.277	0.593	0.266
Logic/Clarity	0.877	0.220	0.889	0.2088	0.565	0.2377	0.575	0.2427
Feedback	0.635	0.413	0.653	0.407	0.491	0.0569	0.500	0.2900
Total Score	3.170	0.709	3.296	0.6918	2.069	0.8273	2.274	0.8568

Table 3
Distribution of measures of association between expert tutor (ET) and ChatGPT-G1/RG2 scoring (N of valid cases = 108).

Depth: Symmetric Measures		ET-Depth *G1-Depth				Depth * RG2-Depth			
		Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance	Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance
Interval by Interval	Pearson's R	-0.068	0.098	-0.700	0.485 ^c	0.072	0.099	0.742	0.460 ^c
Ordinal by Ordinal	Spearman Correlation	-0.071	0.099	-0.731	0.466 ^c	0.064	0.098	0.658	0.512 ^c
Measure of Agreement	Kappa	-0.052	0.064	-0.845	0.398	-0.002	0.067	-0.024	0.981
Analysis: Symmetric Measures		ET-Analysis *G1_Analysis				Analysis * RG2Analysis			
		Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance	Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance
Interval by Interval	Pearson's R	0.083	0.084	0.854	0.395 ^c	0.133	0.088	1.386	0.169 ^c
Ordinal by Ordinal	Spearman Correlation	0.080	0.085	0.831	0.408 ^c	0.135	0.086	1.408	0.162 ^c
Measure of Agreement	Kappa	0.044	0.031	1.202	0.230	0.062	0.046	1.224	0.221
Logic/Clarity of writing: Symmetric Measures		ET-Logic/Clarity * G1_Logic/Clarity				Logic/Clarity * RG2Logic/Clarity			
		Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance	Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance
Interval by Interval	Pearson's R	0.052	0.071	0.539	0.591 ^c	0.120	0.078	1.243	0.217 ^c
Ordinal by Ordinal	Spearman Correlation	0.063	0.073	0.652	0.516 ^c	0.125	0.077	1.294	0.199 ^c
Measure of Agreement	Kappa	0.084	0.039	1.837	0.066	0.084	0.041	1.746	0.081
Feedback on Reflection: Symmetric Measures		ET-Feedback *G1 Feedback				Feedback * RG2Feedback			
		Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance	Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance
Interval by Interval	Pearson's R	0.211	0.050	2.220	0.029 ^c	0.336	0.087	3.678	<0.001 ^c
Ordinal by Ordinal	Spearman Correlation	0.287	0.091	3.088	0.003 ^c	0.320	0.089	3.477	<0.001 ^c
Measure of Agreement	Kappa	0.105	0.049	2.336	0.020	0.150	0.058	2.872	0.004

Table 4
Pair sample size effect analysis.

Pair	Score Criteria	Standardiser	Point Estimate (D)	95% Confidence Interval		Significance	
				Lower	Upper	One-Sided p	Two-Sided p
Pair 1	ET-Depth – G1Depth	0.2540	1.021	0.710	1.332	0.000	0.000
Pair 2	ET-Depth - RG2Depth	0.2475	0.898	0.611	1.185	0.000	0.000
Pair 3	ET-Analysis – G1Analysis	0.2429	1.735	1.385	2.084	0.000	0.000
Pair 4	ET-Analysis - RG2Analysis	0.2370	1.289	0.983	1.595	0.000	0.000
Pair 5	ET-L/Clarity – G1L/Clarity	0.2238	1.448	1.120	1.776	0.000	0.000
Pair 6	ET-L/Clarity- RG2L/Clarity	0.2266	1.385	1.070	1.700	0.000	0.000
Pair 7	ETFeedback-G1Feedback	0.4611	0.351	0.107	0.596	0.002	0.004
Pair 8	ETFeedback-RG2Feedback	0.3583	0.426	0.199	0.654	0.000	0.000
Pair 9	ET-Total/4 – G1Total/4	0.7649	1.568	1.255	1.880	0.000	0.000
Pair 10	ET-Total/4 – RG1Total/4	0.7829	1.270	0.990	1.551	0.000	0.000

scores.

4.2.1. Part 1: descriptive statistics

The comparison between Expert Tutors (ETs) and ChatGPT scores of individual criteria items (Table 2) indicates ET provided higher mean scores across all criteria (“Depth,” “Analysis,” “Logic/Clarity of writing,” “Feedback,”), particularly scoring high in “Logic/Clarity of Writing,” and “Analysis” with variability in “Feedback” scores than ChatGPT-G1 (first prompt scoring) and ChatGPT-G2 (regenerated with same prompt). Generally evaluating “Logic/Clarity” more positively compared with “Feedback” on reflections confirms diversity in perceptions among the ETs scoring.

The Expert Tutors’ “Analysis” overall mean score of 0.898 suggests that almost all students (individual scripts) provided a reflective analysis that effectively met the rubric criteria. In contrast, ChatGPT-G1 ($M = 0.477$) and ChatGPT-RG2 ($M = 0.593$) did not consider most of the scripts to have effectively met the rubric criteria for “Analysis.” Notably, the ETs appear to be more satisfied with the individual student scripts compared with ChatGPT. Comparatively, on regenerated prompts, ChatGPT-RG2 scored slightly higher than ChatGPT-G1, although it was expected that, the same results with grades would have been scored showing consistency in grades awarded.

Overall, the relatively high score of the Expert Tutors suggests that the students on average performed well on all assessment criteria with high mean scores compared with all ChatGPT scores. The relatively small SD also suggests that the scores for the criteria were clustered closely around the mean, indicating consistency in the scoring and students’ performance on all the criteria items.

4.2.2. Part 2: inference statistics

The analysis indicates that there is no significant association between Expert Tutor (ET) scores and ChatGPT-G1 scores for Depth, as evidenced by the Pearson chi-square statistic (0.758) with 2 degrees of freedom (df) and a p-value of 0.685. Similarly, the Interval-by-Interval Pearson’s R-value of -0.068 (Table 3) suggests a weak negative correlation between the two scores, but it is not statistically significant (p-value = 0.485). Cohen’s Kappa measure of agreement yields -0.052 , indicating slight agreement but not statistically significant (p-value = 0.398). The same trend is observed with ChatGPT-RG2 scoring, with a weak, not statistically significant correlation between ET and ChatGPT scoring. This is evidenced by chi-square with a p-value (0.460) at a 0.05 significant level. A Cohen’s Kappa p-value of 0.981 also confirms the lack of agreement was not statistically significant. The analysis concludes that there is no statistically significant relationship between Expert Tutor scoring and ChatGPT-G1/RG2 scoring across all criteria (Table 3). The correlations between the variables are weak and not statistically significant. Additionally, for the “Analysis” and “Logic/Clarity” criteria, there is no significant relationship between the variables, with weak and insignificant correlations.

For “Feedback” on reflections, the Pearson chi-square statistic (10.629) indicates no significant association between ET and G1 scoring

at the 0.05 significance level. However, a weak positive correlation is observed (Pearson’s $R = 0.211$, p-value = 0.029). When comparing ET and RG2, a statistically significant association is found (p-value = 0.003) along with a weak positive correlation (Pearson’s $R = 0.336$, p-value < 0.001). Cohen’s Kappa measure suggests fair agreement between ET and G1 scoring (0.105, p-value = 0.020), and fair agreement between ET and RG2 scoring (0.150, p-value = 0.004). It can therefore be argued that the difference in the scores between the categories of the ET and ChatGPT are unlikely to be due to chance.

The statistical significance of effect sizes (paired *t*-test or Wilcoxon signed-ranked test) confirms significant differences (large effect size) between ChatGPT and Expert Tutor scoring across all criteria variables, except for feedback (Table 4). The p-values (both one-sided and two-sided) for all pairs were less than 0.001, indicating significant differences not occurring by chance (Appendix 2). This implies that ChatGPT scores are significantly more consistent, and objective compared to ET scores. The finding provided insight into the extent of the differences between the paired variables, ChatGPT and Expert Tutor scoring (Table 3). The significant difference among all pairs of variables implies that the variables are not equivalent and possess distinct characteristics or effects.

4.3. Comparing errors made between ChatGPT and Expert Tutor Scoring

HQ3: ChatGPT demonstrated consistent and effective scoring comparable to Expert Tutor.

This section explored the possible effect of misclassified scripts and possible patterns identified in the scoring of the scripts. An error analysis was conducted by two (2) academics who compared the scoring of ChatGPT and Expert Tutors (ET). They each, independently reviewed six randomly selected scripts from the pool of 108 sampled scripts to identify errors or inconsistencies. The comparison revealed issues with grading consistency over time, arbitrary script generation, and repetitive feedback.

4.3.1. Misclassified scripts

ChatGPT did not misclassify any uploaded scripts. It can be argued that ChatGPT understood, but more so identified correlations between variables like task, rubric, and prompts, and summarised it. Each script was accurately scored based on predefined criteria (“Depth,” “Analysis,” “Logic/Clarity of writing,” “Feedback on Reflection,”). The context was properly captured, with a correct interpretation of the script’s intent. ChatGPT appropriately recognised each script and scored by the four (4) rubric criteria (without misclassification), providing a summary of grades awarded from each criterion. ChatGPT appropriately recognised scripts without the student’s “Feedback on Reflection” and scored them as “0”. It provided personalised feedback including the student’s strengths and weaknesses in their reflective writing. However, with a large dataset (more than 4 scripts), ChatGPT self-generated additional student IDs, and scores and occasionally provided generic feedback

unrelated to the scripts. In instances where scoring rationale lacked personalised feedback on students' responses, these issues were rectified after 2 or 3 prompt regenerations.

4.3.2. Consistency

Both Expert Tutors (ET) and ChatGPT-G1/RG2 scoring showed inconsistencies in grades and feedback over time. However, ChatGPT demonstrated more consistency and rigour in scoring and personalised feedback (See Appendix A1). Overall, the ETs provided five (5) formative feedback on "Reflection", "Submission" and generalised structured feedback on "Depth," "Analysis," and "Logic/Clarity of writing." In some cases, ET feedback and grades were inconsistent (Table A2). For instance, in script number 12805, although it received a score of 1.0 for "Depth," the feedback indicated a lack of depth in justifications according to the rubric criteria. ChatGPT scored it as 0.5, a consensus agreed upon by independent reviewers. They found most ET comments inconsistent with the awarded grades across all six independently reviewed scripts. On the contrary, when ChatGPT was prompted to regenerate scores and feedback, the value of personalised feedback remained consistent, but there were slight differences in the graded scores. They changed to either 0.5 or 1.0.

4.3.3. Generalised repetitive feedback

In the ETs scoring process, inconsistencies, especially in personalised feedback, were apparent. Among the 766 of the total scripts, 144 (18.7%) received no personalised feedback despite being graded by the rubric criteria (with scores), raising questions about the grading rationale. In other cases, personalised feedback was not provided across all the rubric criteria, although they were scored (Table A1). Examples of some ET repetitive feedback on "Depth," "Analysis," and "Clarity/Logic" are shown in Appendix Table A1. Out of 799 scripts in the Reflection Comments section, 32% received no feedback, while the rest (68%) received generalised comments (Table 5, Appendix Table A2). There was a lack of consistency across all the scripts. In contrast, ChatGPT's personalised feedback seemed more detailed and directive (on how to improve) across all assessment criteria, but consistency also varied. While some students received feedback on their reflective writing strengths and weaknesses, others did not. Feedback on strengths and weaknesses sometimes appeared in regenerated prompts.

Notably, some ETs provided detailed, criterion-specific personalised feedback, offering clear guidance on what the students needed to do to improve, while others gave more generalised feedback, lacking specificity. For example, 49.1% of the scripts received no feedback on "Depth," and "Analysis," (50.5%), and "Clarity/Logic," (34.9%). Although structured feedback often prompted students' reflection, it was somewhat generic. ChatGPT initially provided consistent accurate grades and personalised feedback, when prompted one script at a time, with a return time of 15 s maximum. However, after assessing multiple sets of scripts at a time (more than 3 or 4 scripts), scoring became inconsistent, and feedback was repeated based on the same prompts. Even when feedback accurately reflected performance, scores sometimes didn't align, a pattern also observed in ET grading.

Table 5
Generic structured feedback on reflection.

Reflection – Comment	No. Std	%
Greater focus on the reflection criteria would lift this reflection to a higher level.	18	2.3
This is a good attempt. You have some but not all of the aspects asked for in the criteria	172	22.5
Well done! This is an excellent reflection.	331	43.2

4.4. Sensitivity analysis: general reviewers' observations

To evaluate ChatGPT's robustness against Expert Tutors (ETs) scoring, a sensitivity analysis was conducted. The aim was to understand how changes in identified factors affect outcomes and shed light on the reliability, robustness, and limitations of ChatGPT marking compared to ETs. Factors considered include accuracy, consistency, speed, and biases. This was done by two independent Reviewers, each reviewing 6 scripts scored by both ETs and ChatGPT.

The two independent reviewers were unanimous in their view that Expert Tutors (ETs) consistently awarded higher scores to students than expected based on the rubric criteria in most scripts. ETs rarely assigned a "0" grade, even when the criteria were not adequately or fully met. For instance, 10 scripts expected to receive a "0" grade were given "0.5" grades. In two criteria, reviewers agreed that students' reflections on "Depth" and "Analysis" which should have been graded a "1" were graded "0.5" instead, indicating inconsistencies in grading. ET feedback lacked specificity on student errors and improvement suggestions, despite generic structured questions, and sometimes questions for the students to think about. Personalised feedback for each criterion was absent, with generic feedback often mismatched with student responses (See Appendix Table A2). Similarly, ChatGPT scoring also showed inconsistencies. Although expected to maintain consistent grades and feedback upon regeneration, some grades were slightly altered while the ideas and values of feedback remained consistent, albeit with softened tone adjustments. Some criteria received lower grades ("0") when they should have received "0.5" or "1.0" according to rubrics criteria. In contrast, ChatGPT's scoring tended to be more critical, although it demonstrated some inconsistency between the initial (G1) and regenerated (RG2) scores compared to the ETs scoring. According to the Reviewers, ChatGPT provided accurate and more consistent personalised feedback across all criteria and tended to be stricter in scoring overall than ETs.

4.4.1. ChatGPT learning of rubrics and scoring scripts

ChatGPT accurately learnt the assessment task, rubric criteria, rubric descriptors, and grading criteria for scripts meeting all criteria (Figs. 2 and 3). It generated a rubric summary in a table format. ChatGPT took 5 s to learn both the rubric and prompt used for the script grading.

Scoring of Scripts: ChatGPT processed each uploaded script by its unique script numbers and scored them based on labelled rubric criteria referenced as "Aspects" (Criteria), and, Reflections on "Depth," "Analysis," "Clarity/Logic of Writing," and "Feedback on Reflection" (Fig. 3). The scripts were accurately identified and scored by ChatGPT. The scoring and personalised feedback provision took approximately 13–15 s per script. Notably, ChatGPT provided justifications for scoring each criterion, a feature not consistently found in ETs' grading. Depending on the prompt, ChatGPT offered a rationale for each criterion's score and personalised feedback on areas for improvement. In some instances, specific comments were provided on the writing's strengths and weaknesses along with suggestions for improvement.

Insights into Experience: ChatGPT can serve as an initial step in providing students objective and precise feedback, despite observed consistency limitations in regenerated grades and feedback. Examiners using ChatGPT for written assessments could benefit from comparing initial and regenerated scores and feedback against rubric criteria. Though this process may require a few minutes, it proves worthwhile. In this study, comparing ChatGPT and ETs' scores and feedback assisted in validating the enhancement of human judgment accuracy and objectivity. However, the study suggests it is premature to rely solely on ChatGPT for grading reflective essays. Limitations, such as the number of scripts processed at a time, could be addressed by inputting one or two student reflection scripts at once, avoiding continuous processing.

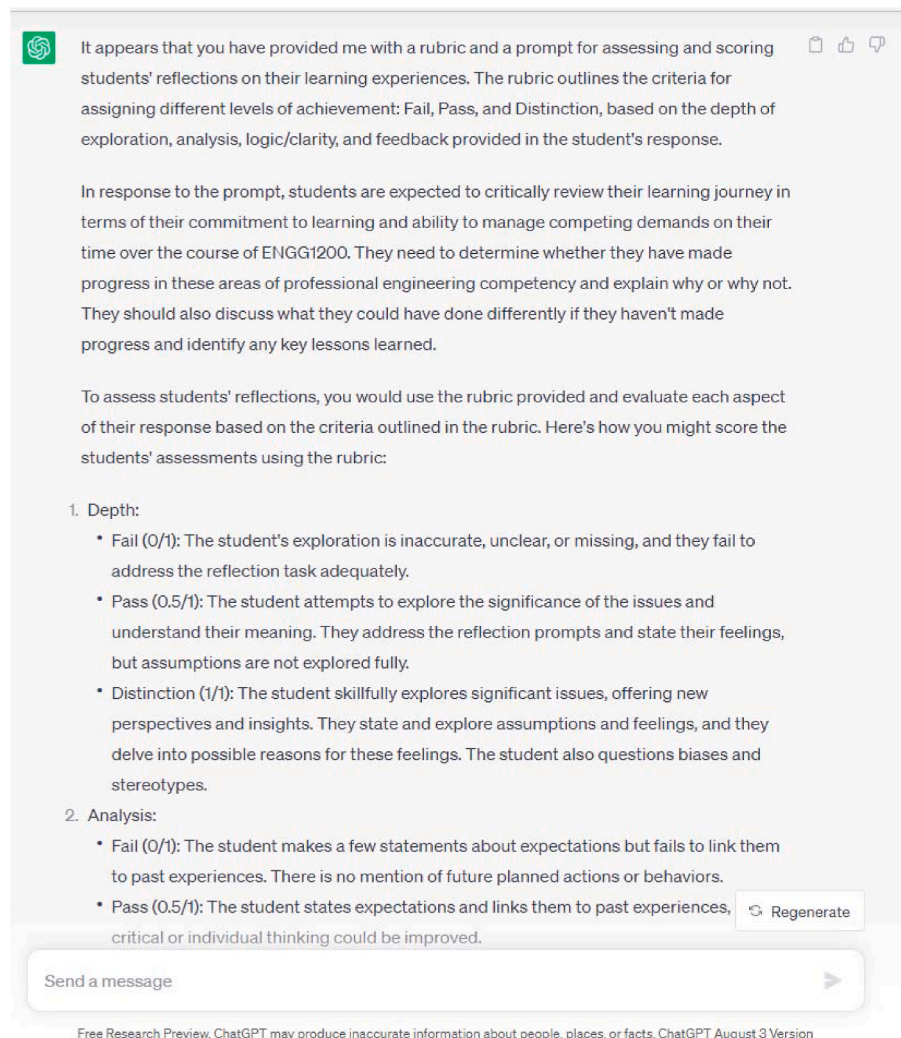


Fig. 2. Example of ChatGPT learning of assessment task and rubrics for scoring scripts.

5. Discussion





The findings present a comparison between Expert Tutor (ET) and ChatGPT scoring of students' reflective scripts using descriptive and inference statistics. The results indicate a significant internal consistency between the two scoring methods, suggesting that ChatGPT is effective in scoring reflective scripts similar to Expert Tutors. While descriptive statistics show a positive performance in Expert Tutor scoring compared to ChatGPT, mean and standard deviation distributions alone are not accurate measures for assessing precision and efficacy, sensitivity/error analysis would be required. The weak correlation between ET and ChatGPT scoring highlights this discrepancy. Initially, there was a belief that rubrics with ranges might not be accurately scored by ChatGPT, but this was proven false as demonstrated by proficiency in scoring within ranges. The decision to maintain absolute binary values in the rubrics aligns with the Expert Tutors' approach for consistency.

The narrower confidence interval for the Intraclass Correlation Coefficient (ICC) suggests greater precision in the estimated ICC, contributing to increased reliability in the analysis. This leads to the conclusion that there is a substantial level of reliability and consistency in the measurements between ET and ChatGPT, attributed to the higher ICC value. However, it's important to contextualize this agreement within specific applications and domains, recognizing that the higher ICC value indicates strong agreement but needs consideration based on the

context.


5.1. Subjectivity in marking/scoring reflective essays


Expert Tutors (ETs) have higher mean scores compared to ChatGPT, suggesting a preference for ETs in scoring. However, a sensitivity analysis reveals that ChatGPT adheres more strictly to the rubrics, compared to the ETs, who sometimes awarded marks inconsistently with their feedback. These variations in ETs marking reflect individual perspectives and experiences in marking, as observed by an independent reviewer "... individual examiners often bring into the marking their own unique experiences and perspectives when interpreting and marking the content of written assessments." They may be influenced by particular phrases, sentences, words, or even their context or knowledge of the students' experiences. Such a personalised approach may add depth (a layer of personal touch and insight) to the evaluation but may also lead to false conclusions favouring ET scores over ChatGPT. This highlights not only the challenges involved in creating fair and accurate assessments but also in the complexity of evaluating writing for structure, coherence, and clarity while providing timely and constructive feedback (Zak & Weaver, 1998). Hence, relying solely on statistics to judge performance may lead to misleading conclusions. The flexibility of human evaluation, influenced by the 39 different tutors with unique interpretations (though guided by the rubrics), explains the higher scores given by the ETs. Despite their contextual insight and empathy,

1/2 >  Certainly, here's a summary of the rubric presented in a table format:   

Aspect	Fail (0/1)	Pass (0.5/1)	Distinction (1/1)
Depth	Exploration inaccurate, unclear, or missing.	Some attempt; mostly addresses prompts.	Skillful exploration; new insights, assumptions explored.
Analysis	Few statements, no linkage to past, future.	Expectations linked to past, room for improvement.	Critical thinking, innovation, clear linkage to past.
Logic/Clarity	Short, unconnected, incomplete examples.	Clear, concise, well-organized, brief reasons.	Well-written, fluid, original, logical, supported.
Feedback	Personal taste evaluation, negative tone.	Identifies strengths/weaknesses, lacking depth.	Objectively identifies, thoughtful, improvement suggestions.

Use this table as a quick reference guide when assessing students' responses based on the rubric provided. Match each aspect of the student's work against the corresponding criteria in the table to determine the appropriate score (Fail, Pass, Distinction) for each aspect.

 Regenerate

Send a message 

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT August 3 Version](#)

Fig. 3. Example of ChatGPT learned rubric summarised in a table for scoring scripts.

human markers may overlook strict adherence to the rubric, a strength of AI tools like ChatGPT.

The involvement of AI, like ChatGPT, represents a paradigm shift in assessment. It offers efficiency gains and improved accuracy, which would enable examiners to focus on enhancing teaching, providing personalised feedback, and aiming for authenticity. Prior research supports AI's capacity to alleviate workload and enhance grading precision (Luckin et al., 2022; Munir et al., 2022). However, a balanced approach incorporating human evaluation is recommended. ChatGPT's ability to evaluate multiple scripts within seconds while providing relevant feedback underscores its efficiency. This efficiency significantly reduces examiners' workload, potentially saving nearly 30 min per script. It can therefore be argued that ChatGPT has the potential to reduce subjectivity in assessment, enhancing grading accuracy (Maier & Klotz, 2022; Munir et al., 2022; Zak & Weaver, 1998).

5.2. ChatGPT scores reflective essays as expert tutors (ETs)

The comparison between ChatGPT and ETs scores (Table 3) provides insight into the essay evaluation process and the factors contributing to significant differences between the two methods. The clear internal consistency in the evaluated scripts by both ETs and ChatGPT suggests non-random fluctuations but rather a systematic occurrence. Both approaches to marking assessed the same constructs outlined in the rubrics, indicating that much of the variability in measurements can be attributed to alignment within each method. ChatGPT's measurements consistently aligned with the grading procedures of the ETs, indicating its proficiency in assessment. This alignment is evident as both ETs and ChatGPT scored within the rubric criteria range of grades 0 to 1, evaluating the same constructs. This contradicts Elliot and Klobucar (2013) scepticism regarding automated systems' ability to assess coherence, organization, and logic in essays accurately.

The alignment in ET and ChatGPT measurements underscores ChatGPT's potential and ability to learn from rubrics, grade tasks, and provide personalised feedback, which can be used by students to improve on their comprehension and knowledge acquisition, writing

skills and encourage engagement (Gibbs & Simpson, 2005). This finding aligns with Zawacki-Richter et al. (2019) observations on AI-assisted marking's potential to enhance student learning through valuable feedback. ChatGPT's ability to efficiently grade all 108 scripts and provide personalised feedback within a day, means students could receive prompt feedback. While differing in focus and conclusions, Ouyang et al. (2022) and Chiu et al. (2023) also emphasize the use of AI in assessments, echoing the potential benefits of personalised learning, improved outcomes, and enhanced grading efficiency. Thus, ChatGPT's scoring of reflective essays parallels ETs' evaluations, consistent with prior research on AI scoring of written assessments (Mizumoto & Eguchi, 2023; Richardson & Clesham, 2021).

5.3. Effectiveness and variation in performance in scoring reflective essays

ChatGPT has demonstrated its ability to comprehend both content and context in written reflections and effectively assess them according to rubric criteria when provided with suitable prompts. It can recognize and evaluate creativity and originality, improving its capability through self-learning. The ability to analyse the "Depth", "Analysis" of reflections, and assess the "Logic and Clarity in writing", and "Feedback" on reflections, stands in contrast to some earlier doubts about AI's ability to automation of written assessments (Elliot & Klobucar, 2013).

Despite its proficiency, potential variations in measurements may stem from subjective factors inherent in both Expert Tutors (ETs) and ChatGPT scoring methods. The variability in measurements can be attributed to subjective factors among ETs and inconsistencies in ChatGPT's performance over time. That is, ETs, despite having rubrics, sometimes provide feedback inconsistent with assigned grades, indicating subjective influences among the 39 ETs. Similarly, ChatGPT's performance may decline over time with repetitive scoring, necessitating caution to maintain alignment with students' reflections, in ensuring that aspects of the criteria are not overlooked. Processing 2–3 essay scripts at a time yields more accurate scores and feedback compared to processing more than 4 scripts simultaneously. Over time, ChatGPT's performance issues may be resolved with system

improvements. The alignment between ETs and ChatGPT measurements again highlights ChatGPT's potential for effective grading tasks (Kalervo et al., 2022; Swiecki et al., 2022). Its potential to enhance students' comprehension and knowledge acquisition (Santamaría Lancho et al., 2018), enhancing motivation, engagement, and learning outcomes (Gibbs & Simpson, 2005; Lee et al., 2022) is important for higher education learning. While some limitations exist (Maier & Klotz, 2022) ChatGPT's experience suggests these can be overcome with appropriate prompts.

We can therefore argue that the efficiency gains and accuracy improvements presented by ChatGPT may have significant implications for learners, as timely personalised feedback encourages supportive and engaging learning environment (Lee et al., 2022; Maier & Klotz, 2022). Thus, ChatGPT's potential has demonstrated that, given the appropriate prompts, can be used to overcome human limitations in marking assessments.

6. Challenges, limitations, implications, and recommendations to higher education learning

6.1. Challenges

The reliability of ChatGPT's scoring over time poses notable concerns, particularly regarding systems fatigue and processing more than four scripts simultaneously. To mitigate this, a practice was adopted to review every graded script and personalised feedback for consistency. While some may view this as counterproductive, insights from this study show significant time savings despite the additional review step. Undoubtedly, ChatGPT has proved to have the ability to overcome challenges of understanding textual nuances and meaning, which were previously shown to limit AI's ability to assess more intricate writing aspects (Burstein et al., 2004; Celik et al., 2022). It is worth noting that, the effectiveness of ChatGPT heavily depends on appropriate prompts and consistent outcome reviews.

The researcher's experience in this study highlights a sense of detachment from students' writing during the script uploading and scoring process, hindering flexibility. Understanding students' perspectives and learning experiences is limited, indicating an area of improvement in flexibility. Thus, the examiner may lose the learning experience from the student's reflection of their worldview and how that influences the way they solve problems. With many institutions lacking policies around AI use in assessments, careful consideration is critical. To effectively support examiners, further research is needed to explore the strengths, limitations, and appropriate use of generative AI systems in teaching and assessment, and its alignment with institutional procedures. Course coordinators must be supported in adhering to such institutional policies.

6.2. Limitations to the study

A major limitation was the absence of a machine-learning process for the AI system to learn from perfectly marked scripts with personalised feedback, and use it to grade scripts accurately from scratch. This absence may have potentially led to system fatigue over time (while processing more than 4 uploaded scripts at a time). Additionally, the human prompts provided for ChatGPT are subjective and interpretative of the creativity and originality of the essay scripts. Thus, the accuracy of marked scores and appropriateness of feedback depends on the prompts recognised by ChatGPT as applicable to all scripts. Nonetheless, ChatGPT has shown potential for scoring essays and providing feedback comparable to Expert Tutors. However, consistent improvement of prompts is important for success.

6.3. Implications for higher education learning

This study argues that ChatGPT's potential to enhance learning

assessment tasks by using rubrics to evaluate scripts and provide personalised feedback means it can not only personalise learning experiences and offer real-time performance feedback, but useful in managing grading workload. It can free up time for other teaching and learning activities, as evidenced by the example of scoring the 108 scripts in a day. However, a deep understanding of ChatGPT's workings, effective prompts, and script review is important for accurate scoring. Examiners must train themselves on effectively prompting ChatGPT. Ethical concerns regarding data ownership and usage also need attention.

6.4. Recommendations

The study recommends that higher education institutions collaborate with AI system providers to ensure that the development of AI tools and resources does not compromise effective teaching and learning for educational purposes (academic integrity). AI tools can be customised to mark written assessments for large class cohorts of students. In addition, while ChatGPT has the potential to reliably score written essays and provide feedback comparable to that of Expert Tutors, our experience suggests that a consistent iteration of improved prompts was critical for the success achieved in this study. Similarly, the feedback from ChatGPT can be used to identify the strengths and weaknesses of struggling students through reflections on their experiences in the Engineering program.

7. Conclusion

This study which was prompted by challenges in traditional essay grading and the need for personalised feedback to enhance students' learning, evaluated the effectiveness of generative AI (ChatGPT) in comparison to human intelligence (human marking) for grading reflective essays. The findings show that, with the right prompt, ChatGPT can score reflective written essays effectively. It revealed:

- Both Expert Tutors and ChatGPT demonstrated consistency in the processes of marking and providing personalised feedback based on task instructions and rubrics. However, both also showed inconsistencies and struggles to offer sufficient detailed feedback over time.
- Statistically significant differences existed between the Expert Tutors' and ChatGPT's marking approaches. While AI marking seemed consistent and objective, Expert Tutors offered flexibility and nuanced judgments not accounted for by ChatGPT's current prompt variations.
- ChatGPT's effectiveness relied on the suitability of prompts provided, with the potential for system fatigue leading to generalised feedback lacking specificity to individual student needs. The absence of machine learning for perfectly marked scripts over time may have contributed to this limitation.

The study underscores the importance of moderating ChatGPT-scored scripts and their personalised feedback. It emphasises AI's potential to enhance written assessment evaluation in higher education but advocates for careful consideration of ethical and institutional policies. Institutions should invest in AI for marking reflective essays, balancing its advantages with responsible usage, while working with AI service providers. Further research is needed to explore the benefits and challenges of AI integration in assessment and feedback processes.

CRediT authorship contribution statement

Isaiah T. Awidi: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

I extend my gratitude to Dr. Stephen Hall, Head of the initial Teaching and Learning Grant Project, for providing the data used in this evaluation, as well as to all faculty colleagues involved in the project. I am also thankful for the valuable critique and proofreading provided by Ms. Cathy Tames and Dr. Neil Martin.

Appendix

Appendix Table A1

Expert Tutor Generalised Repetitive Feedback.

Depth Marker Comment	Frequency	Percent
1. No Feedback Provided	376	49.1
2. Pretty surface-level stuff you are talking about. No feelings stated	44	5.7
3. Try discussing your experience with your team members. They may have helpful suggestions.	56	7.3
4. Try standing back from the experience. What other ideas relate to the experience? Why/how did it happen? What factors contributed? How do you feel about it	47	6.1
5. Try thinking about why/how did it happen? What factors contributed? How did you feel about it?	116	15.1
6. Try using the reflection prompt as a guide for finding a suitable experience/activity/problem or issue	55	7.2
7. You mentioned a problem, how did you feel about it?	72	9.4
Total	766	100.0
Analysis Marker Comment	Frequency	Percent
1. No Feedback Provided	387	50.5
2. Some initial exploration of issues and events within the team and yourself, but little in-depth thought evident and assumptions not explored	31	4.0
3. Try analysing the interaction, event, or episode you described.	15	2.0
4. Try standing back from the experience. What other ideas relate to the experience? Why/how did it happen? What factors contributed? How do you feel about it	48	6.3
5. What can be concluded, in a general sense, from these experiences and the analysis you have undertaken?	33	4.3
6. What can be learned from this experience? What can be concluded about your own specific, unique, personal situation or ways of working?	65	8.5
7. What might this experience mean in the context of your course?	23	3.0
8. What sense can you make of the situation? Bring in ideas from outside the experience to help you. What was really going on? (You could consider for example: further observations, details, factors, relevant other knowledge, suggestions from others, new information, formal theory).	52	6.8
9. What was your role? What feelings and perceptions surrounded the experience? How would you explain the situation to someone else? What other perspectives, theories, or concepts could be applied to the situation?	26	3.4
10. What will you do as a result of this experience? How will you use it to inform your future	86	11.2
Total	766	100.0
Clarity Marker Comment	Frequency	Percent
1. No Feedback Provided	267	34.9
2. Check that your grammar and spelling are correct before submitting	66	8.6
3. Make sure that information is organised into concise manageable paragraphs that cover one key idea.	96	12.5
4. Try presenting information in a way that shows the connection between statements	24	3.1
5. Try providing more information for each point	34	4.4
6. Well written and has a nice flow. Great work	279	36.4
Total	766	100.0

Appendix Table A2

Sensitivity Analysis - Comparison of ChatGPT and Expert Tutor Scoring

Random Coded Student	G1-Depth	G1-Depth Comment	G1-Analysis	G1-Analysis Comment	G1-L/Clarity	G1-L/Clarity Comment	G1-Feedback	G1-Feedback Comment	G1-Total/4	Overall Feedback Comment.
12,805	0.5	The essay briefly explores the issues of commitment and managing competing demands on time. It mentions personal flaws and changes made to address them. However, it lacks deeper exploration and new perspectives.	0.5	The essay discusses the changes implemented to manage time and states progress in commitment and managing demands. It links past experiences to expectations, but critical thinking and innovative connections are lacking.	0.5	The essay is relatively clear and organized, with reasons and examples provided. However, it could be more concise and could better meet the requirements outlined in the instructions.	0.5	Strengths and weaknesses are identified, but the justification lacks depth. Suggestions for improvement are not explicitly mentioned.	2	Overall, the essay meets the requirements to pass but falls short in terms of depth, critical thinking, and feedback. The final score is 2/4, which corresponds to a "Pass" level.

(continued on next page)

Appendix Table A2 (continued)

Random Coded Student	G1-Depth	G1-Depth Comment	G1-Analysis	G1-Analysis Comment	G1-L/Clarity	G1-L/Clarity Comment	G1-Feedback	G1-Feedback Comment	G1-Total/4	Overall Feedback Comment.
ET-12805	1.0	Try using the reflection prompt as a guide for finding a suitable experience/activity/problem or issue	0.5		1.0	Well written and has a nice flow. Great work	0.5		3 (3.5)	Very well written. A strong, in-depth view on the improvement of organisation skills gained. The writing is mostly clear, with a few slip ups here and there, and although it has great overall depth, there are some justifications lacking in depth. Using the marking rubric provided, I believe its fair to give the student
19,202	1	The essay delves into the issues of commitment and managing competing demands, providing personal insights and reflections. It explores past experiences, changes made, and the impact on learning and time management.	1	The essay demonstrates critical thinking and makes innovative connections between past experiences and the expectations of ENGG1200. It clearly identifies progress and offers plans for further improvement.	0.5	The essay is generally clear and well-organized. However, it could benefit from being more concise and meeting all the requirements outlined in the instructions.	0.5	Strengths and weaknesses are objectively identified, but suggestions for improvement could have been more specific and detailed.	3	Overall, the essay excels in depth and analysis, meeting the criteria for a distinction. However, it falls slightly short in terms of logic/clarity and feedback. The final score is 3/4, which corresponds to a "Distinction" level.
ET-19202	0.5	You mentioned a problem, how did you feel about it?	1.0	What sense can you make of the situation? Bring in ideas from outside the experience to help you. What was really going on? (You could consider for example: further observations, details, factors, relevant other knowledge, suggestions from others, new information, formal theory).	1.0		1.0		3.5	1a) Very honest reflection of your own weaknesses regarding your studies. It also appeared to be somewhat realistic towards the end where you still admitted you had some improving to do in the future.1b) Again, to reiterate what was said above, you have a very honest and productive way of self-criticism, through which you seem to have learnt to use to drive you forward. With the particular method of time management, you mentioned (completing a set portion every night) will be particularly beneficial, because you will gradually form a habit of doing such work every night, which will make it easier to

(continued on next page)

Appendix Table A2 (continued)

Random Coded Student	G1-Depth	G1-Depth Comment	G1-Analysis	G1-Analysis Comment	G1-L/Clarity	G1-L/Clarity Comment	G1-Feedback	G1-Feedback Comment	G1-Total/4	Overall Feedback Comment.
13,703	0.5	The essay demonstrates some attempt to explore the significance of commitment to learning and managing competing demands. However, there is room for improvement in terms of exploring assumptions and providing new perspectives.	0.5	The essay links expectations to past experiences and acknowledges the importance of managing competing demands. However, there is no mention of future planned action or behaviours, which could enhance the analysis.	0.5	The essay is mostly clear and organized. However, there are some unconnected statements, and the examples/reasons provided are incomplete and do not fully meet the requirements.	0.5	The essay identifies strengths and weaknesses appropriately, but the justification lacks depth. Suggestions for improvement are not explicitly provided.	2	motivate yourself each time you go to complete said work. In conclusion, I believe that it sounds like you have a very healthy approach to your learning and given your ability to self-criticise you will only continue to improve in your organisation and productivity. *1.5
ET-13703	0.5	Pretty surface level stuff you are talking about. No feelings stated	0.5	Try analysing the interaction, event or episode you described.	1.0		0.5		2.5	This reflection explores the issue of time and people management in a way that is relevant to the student and their very well. This piece is clear and easy to read whilst covering most aspects of the task required. Overall, this essay meets the requirements for a Distinction grade (3/4).
13,505	1	The essay skilfully explores the significance of commitment to learning and managing competing demands, providing new insights and perspectives. Assumptions, feelings, and reasons for changes are explored effectively.	0.5	The essay links expectations to past experiences and highlights the importance of reminding oneself of goals. However, there is no mention of making connections beyond the ENGG1211 course, which could enhance the analysis.	1	The essay is well-written, clear, concise, and demonstrates a logical structure. Viewpoints and interpretations are well-supported.	0.5	The essay identifies strengths and weaknesses appropriately, but the suggestions for improvement are not explicitly provided. The approach is generally sensitive.	3	Overall, this essay meets the requirements for a Distinction grade (3/4).
ET-13505	0.5	You mentioned a problem, how did you feel about it?	1.0	What sense can you make of the situation? Bring in ideas from outside the experience to help you. What was really going on? (You could consider for example: further	1.0	Well written and has a nice flow. Great work	0		2.5	Filled out all criteria. Great!

(continued on next page)

Appendix Table A2 (continued)

Random Coded Student	G1-Depth	G1-Depth Comment	G1-Analysis	G1-Analysis Comment	G1-L/Clarity	G1-L/Clarity Comment	G1-Feedback	G1-Feedback Comment	G1-Total/4	Overall Feedback Comment.
12,506	0.5	The essay explores the issue of commitment to learning and time management. It acknowledges the challenges faced during the course, such as team members not pulling their weight and difficulties in organizing team meetings. The exploration of assumptions and feelings is minimal. The depth of exploration could be improved	0.5	observations, details, factors, relevant other knowledge, suggestions from others, new information, formal theory). The essay connects expectations to past experiences and suggests improvements for the future, such as dividing the workload more fairly and utilizing a Gantt chart. However, the critical thinking and innovative connections to beyond ENGG1211 are lacking. The analysis demonstrates some understanding of the task but does not fully meet the requirements for a Distinction	0.5	The essay is relatively clear and well-organized, although it could benefit from more examples and reasons. It meets the requirements for a Pass level	0.5	The essay identifies areas for improvement and provides some suggestions. However, the feedback lacks depth in terms of justification. It meets the requirements for a Pass level	2	Overall, the essay demonstrates an adequate level of reflection but falls short in terms of depth, critical thinking, and innovative connections. The essay achieves a Pass level
ET-12506	1.0	Try discussing your experience with your team members. They may have helpful suggestions.	0.5	What can be concluded, in a general sense, from these experiences and the analysis you have undertaken?	0.5	Try presenting information in a way that shows the connection between statements	0.0		2	This student has addressed the questions however could have gone into more depth. They have structured it well with an intro, body, and conclusion. More could have been written to provide greater depth and understanding. Overall, the essay provides some reflection but lacks depth and critical thinking. The essay achieves a Pass level (2.0 out of 4.0) based on the rubrics.
14,508	0.5	The essay briefly touches on the issue of commitment to learning and time management. It mentions the fear of letting team members down and the challenges faced in completing tasks due to unforeseen setbacks. However, there is limited exploration of assumptions and feelings. The depth of exploration could be improved	0.5	The essay links expectations to past experiences and acknowledges the need for time management and dealing with unpredictable situations. However, the analysis lacks depth in terms of critical thinking and making connections beyond ENGG1211. It meets the requirements for a pass-level	0.5	The essay is relatively clear in its communication, although some statements are brief and could be further developed. It meets the requirements for a pass-level	0.5	The essay identifies the progress made in commitment to learning and the ability to manage time. However, the feedback lacks in-depth justification and suggestions for improvement. It meets the requirements for a pass-level	2	Overall, the essay provides some reflection but lacks depth and critical thinking. The essay achieves a Pass level (2.0 out of 4.0) based on the rubrics.
ET-14508	0.5	Try using the reflection prompt as a guide for	0.5	What sense can you make of the situation? Bring in	1.0	Try providing more information for each point	0.0		2	-No feedback given-

(continued on next page)

Appendix Table A2 (continued)

Random Coded Student	G1-Depth	G1-Depth Comment	G1-Analysis	G1-Analysis Comment	G1-L/Clarity	G1-L/Clarity Comment	G1-Feedback	G1-Feedback Comment	G1-Total/4	Overall Feedback Comment.
		finding a suitable experience/activity/problem or issue		ideas from outside the experience to help you. What was really going on? (You could consider for example: further observations, details, factors, relevant other knowledge, suggestions from others, new information, and formal theory).						

Note: ChatGPT generated repetitive feedback over time, comparable to Expert Tutor grading and providing personalised feedback.

References

- Al Darayseh, A. (2023). Acceptance of artificial intelligence in teaching science: Science teachers' perspective. *Computers and Education: Artificial Intelligence*, 4, Article 100132.
- Al Ghatrifi, M. O. M., Al Amairi, J. S. S., & Thottoli, M. M. (2023). Surfing the technology wave: An international perspective on enhancing teaching and learning in accounting. *Computers and Education: Artificial Intelligence*, 4, Article 100144.
- Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.
- Boud, D. (2007). Reframing assessment as if learning were important. In *Rethinking assessment in higher education* (pp. 24–36). Routledge.
- Brookhart, S. M. (2017). *How to give effective feedback to your students*. ASCD.
- Bruscia, K. E. (2016). Types of objectivist research. In *Introduction to music therapy research* (pp. 59–60). Barcelona Publishers Barcelona.
- Bunch, M. B., & Cizek, G. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications.
- Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine*, 25(3), 27–27.
- Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends*, 66(4), 616–630.
- Chamberlain, C., Button, A., Dison, L., Granville, S., & Delmont, E. (2004). The role of short answer questions in developing higher order thinking. *Per Linguam: A Journal of Language Learning= Per Linguam: Tydskrif Vir Taalaanleer*, 20(2), 28–45.
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8, 75264–75278. <https://doi.org/10.1109/ACCESS.2020.2988510>
- Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4, Article 100118. <https://doi.org/10.1016/j.caeai.2022.100118>
- Earl, L. M., & Katz, S. (Eds.). (2006). *Leading schools in a data-rich world: Harnessing data for school improvement*. Corwin Press.
- Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In *Handbook of automated essay evaluation* (pp. 38–57). Routledge.
- Fadel, C., Holmes, W., & Bialik, M. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Boston, MA: The Center for Curriculum Redesign.
- Fitzgerald, M. (1994). Why write essays? *Journal of Geography in Higher Education*, 18(3), 379–384. <https://doi.org/10.1080/03098269408709282>
- Foltz, P. W. (2020). Practical considerations for using AI models in automated scoring of writing. *Application of Artificial Intelligence to Assessment*, 101.
- Gibbs, G., & Simpson, C. (2005). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, (1), 3–31.
- Haines, C. (2021). *Assessing students' written work: Marking essays and reports*. Routledge.
- Holmes, W., Bialik, M., & Fadel, C. (2020). *Artificial intelligence in education*.
- Hounsell, D. (1995). Marking and commenting on essays. In F. Forster, D. Hounsell, & S. Thompson (Eds.), *Tutoring and demonstrating: A handbook* (pp. 51–64).
- Joksimovic, S., Ienthaler, D., Marrone, R., De Laat, M., & Siemens, G. (2023). Opportunities of artificial intelligence for supporting complex problem-solving: Findings from a scoping review. *Computers and Education: Artificial Intelligence*, Article 100138.
- Kalervo, G., Thompson, G., Swist, T., Kitto, K., Rutkowski, L., Rutkowski, D., Hogan, A., Zhang, V., & Knight, S. (2022). *Automated essay scoring in Australian schools: Key issues and recommendations*.
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3, Article 100074. <https://doi.org/10.1016/j.caeai.2022.100074>
- Krendl, K. A., & Lieberman, D. A. (1988). Computers and learning: A review of recent research. *Journal of Educational Computing Research*, 4(4), 367–389.
- Lee, Y.-F., Hwang, G.-J., & Chen, P.-Y. (2022). Impacts of an AI-based chatbot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation. *Educational Technology Research & Development*, 70(5), 1843–1865. <https://doi.org/10.1007/s11423-022-10142-8>
- Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large language models write reflectively. *Computers and Education: Artificial Intelligence*, 4, Article 100140. <https://doi.org/10.1016/j.caeai.2023.100140>
- Luckin, R., Cukurova, M., Kent, C., & du Boulay, B. (2022). Empowering educators to be AI-ready. *Computers and Education: Artificial Intelligence*, 3, Article 100076. <https://doi.org/10.1016/j.caeai.2022.100076>
- Maier, U., & Klotz, C. (2022). Personalised feedback in digital learning environments: Classification framework and literature review. *Computers and Education: Artificial Intelligence*, 3, Article 100080. <https://doi.org/10.1016/j.caeai.2022.100080>
- Mizumoto, A., & Eguchi, M. (2023). *Exploring the potential of using an AI language model for automated essay scoring*. Available at: SSRN 4373111.
- Munir, H., Vogel, B., & Jacobsson, A. (2022). Artificial intelligence and machine learning approaches in digital education: A systematic revision. *Information*, 13(4), 203.
- Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies*, 27(6), 7893–7925.
- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527.
- Richardson, M., & Clesham, R. (2021). Rise of the machines? The evolving role of artificial intelligence (AI) technologies in high stakes assessment. *London Review of Education*, 19(1), 1–13.
- Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.
- Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach*. Pearson Series.
- Santamaría Lancho, M., Hernández, M., Sánchez-Elvira Paniagua, A., Luzón Encabo, J. M., & de Jorge-Botana, G. (2018). Using semantic technologies for formative assessment and scoring in large courses and MOOCs. *Journal of Interactive Media in Education*, 2018(1).
- Shermis, M. D. (2010). Automated essay scoring in a high stakes testing environment. *Innovative Assessment for the 21st Century: Supporting Educational Needs*, 167–185.
- Shermis, M. D. (2022). Anchoring validity evidence for automated essay scoring. *Journal of Educational Measurement*, 59(3), 314–337.
- Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.
- Shermis, M. D., Burstein, J., Elliot, N., Miel, S., & Foltz, P. W. (2016). *Automated writing evaluation: An expanding body of knowledge*.
- Southworth, J., Migliaccio, K., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence*, Article 100127.
- Stephen, T. C., Gierl, M. C., & King, S. (2021). Automated essay scoring (AES) of constructed responses in nursing examinations: An evaluation. *Nurse Education in Practice*, 54, Article 103085.
- Sweller, J. (2003). *Evolution of human cognitive architecture* (Vol. 43). Academic Press.
- Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3, Article 100075. <https://doi.org/10.1016/j.caeai.2022.100075>
- Tindall-Ford, S. A., & Sweller, J. (2020). *Advances in cognitive load theory: Rethinking teaching*. Milton Park, Abingdon, Oxon: Routledge 2 Park Square, OX14 4RN.
- Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, 85(4), 475–511.
- Warburton, N. (2020). *The basics of essay writing*. Routledge.

- Wiggins, G. (1990). The case for authentic assessment. *Practical assessment, research, and evaluation*, 2(1).
- Wiggins, G. (2012). Seven keys to effective feedback. *Feedback*, 70(1), 10–16.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13.
- Winstone, N., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge.
- Xia, Q., Chiu, T. K., Zhou, X., Chai, C. S., & Cheng, M. (2022). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, Article 100118.
- Yildirim-Erbasli, S. N., & Bulut, O. (2023). Conversation-based assessment: A novel approach to boosting test-taking effort in digital formative assessment. *Computers and Education: Artificial Intelligence*, 4, Article 100135.
- Zak, F., & Weaver, C. C. (1998). *The theory and practice of grading writing: Problems and possibilities*. SUNY Press.
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 1–27.
- Zhang, S. (2021). Review of automated writing evaluation systems. *Journal of China Computer-Assisted Language Learning*, 1(1), 170–176.