ADVANCE SOCIAL SCIENCE ARCHIVE JOURNAL
Available Online: https://assajournal.com
Vol. 04 No. 02. Oct-Dec 2025.Page#.1619-1636
Print ISSN: 3006-2497 Online ISSN: 3006-2500
Platform & Workflow by: Open Journal Systems
https://doi.org/10.5281/zenodo.17605828

Investigates the Accuracy, Efficiency, and Potential Bias of AI-Driven Automated Grading Systems

Dr. Mahek Arshad
Controller of Examinations, FBCOE(W)
mehakrshd@gmail.com
Misbah Yasmeen
Assistant Professor, PhD Scholar, Department of Education FBCOE (W)
misbahyasmeen@gmail.com
Ms. Ansa Nighat Iqbal
Assistant Professor, PhD Scholar, Department of Business Administration
FBCOE(W)
Iqbalansa81@gmail.com
Naeem Akhtar
PhD Scholar, My University, Assistant Professor, IMCB, F-8/4 Islamabad
dedu241002@myu.edu.pk

ABSTRACT
Automated grading systems have become a part of the modern education evaluation process and are powered by Artificial Intelligence (AI) with the promise of efficiency, consistency, and scale. This paper examines the validity, effectiveness, and possible biasness of AI-based grading systems in learning institutions. With the growing use of machine learning and natural language processing algorithms by institutions to measure the performance of students, the question has arisen about the reliability, transparency, and fairness of automated evaluations. The study examines the relative performances of AI in grading systems and human ratings; the study finds that there are discrepancies in performance based on gender, ethnicity, and language. The research design based on mixed-method is that which integrates quantitative data analysis of the results of grading 500 student essays with qualitative data analysis of the teacher and student interviews. To estimate whether AI scoring systems are systemically biased or nonadherent to human grading criterion, statistical tests (regression analysis, ANOVA, and t-tests) are utilized. Results indicate that AI grading results in a time efficiency of up to 70 per cent, although accuracy in different fields is greatly different, and the issue of fairness remains to be debated, particularly in the case of non-native speakers of English. The paper highlights the need to have algorithmic transparency, ethical auditability and hybrid assessment models which incorporate human control. Finally, the paper will be used to contribute to the current debate on the responsible use of AI in education by providing empirical evidence and suggestions of the creation of a fair automated assessment system.

Keywords: *Artificial Intelligence, Automated Grading, Algorithmic Bias, Efficiency, Accuracy, Educational Assessment, AI Fairness, Natural Language Processing, Machine Learning*.

Introduction

Educational assessment is experiencing a radical change in the digital era, which is fueled by the development of artificial intelligence (AI), machine learning, and natural-language processing (NLP). Among the most interesting changes is related to automated grading systems

- the systems that are designed to assess the work of students (particularly, the essays and short answers) with minimum human intervention. Such AI-based automated grading systems (also commonly known as automated essay scoring (AES) when used on writing) have the potential to provide faster feedback, consistent scoring, and scalability. But the zeal of their implementation should be conditioned by the critical analysis of three interconnected dimensions of accuracy (the degree to which the system scores match human judgment and instructions) and efficiency (the degree to which it saves time, money, or teaching effort), and bias (the extent to which the algorithms are unbiased or discriminating of a specific group of learners or writing styles). This introduction is a critical review of the current research done in these dimensions, elucidates the underlying mechanisms, and justifies the necessity of caution and careful design.

The motivation to move to automated grading is informed by a number of pressures on modern day education. With hundreds or thousands of student submissions in many large-class or massive open online course (MOOC) instructors, manually grading essays or other open-ended assignments is prohibitively time-consuming, delays feedback, and creates variability in scoring based on fatigue or inconsistency in the rater. Logistically speaking, mechanising part of the grading procedure in such a way therefore has a great attraction: quick turnover, application of standards evenly, and the possibility of saving on costs. Besides, faster feedback loops have been suggested to facilitate student learning and engagement in formative assessment contexts.

Nonetheless, according to the literature, the implementation of AI grading is not merely a replacement of human labour by the machine; it transforms the very essence of assessment itself. When algorithms are being used to decide grades, it changes the values and measures, which often prioritize those aspects that can be readily measured (grammar, vocabulary, length, structure) over more profound qualities of reasoning, creativity, or subject-related insight. Additionally, being used on a large scale and particularly in high stakes environments, algorithmic grading engages questions of fairness, student trust, transparency, and pedagogic validity. Therefore, any serious study should take into account not only the existence of such systems but the cost of those systems as well as the individuals that it caters to and under what conditions.

In automated grading, accuracy is defined in terms mostly of the similarity between the scores produced by machines and those of human raters, as well as the ability of the system to measure the construct that it is claimed to measure (e.g., quality of essays, quality of argument). Practically, measures of research also include Pearson correlation, mean absolute error (MAE), quadratic weighted kappa (QWK) or intra-class correlation coefficients (ICC) to compare machine scoring with human scoring. An example of automated essay scoring systems has reported moderate to high-correlations, with significant qualifications as tasks increased in difficulty (Hussein, Hassan, and Nassef, 2019). PeerJ

More recently, with hybrid models that incorporate deep-learning embeddings (e.g. RoBERTa) alongside handcrafted linguistic features giving a QWK of 0.941 on a heterogeneous essay dataset, it suggests that the accuracy can be as good as it can be in certain circumstances (Mathematics, 2024). MDPI Of course, these high rates of agreement cannot be generalised across the board- especially in the case of open-ended prompts, where creativity is at play, where domain knowledge is demanded and where the system is subjected to writing styles other than those in its training distribution.

Most importantly, there are studies that have shown that automated systems can act in ways contrary to human judgement in significant aspects. In particular, Singla et al. (2021)

established that AES systems are not only overstable (relatively little change in the score with such drastic changes in the input), but also oversensitive (relatively large change in the score with such drastic changes in the input) - signalling the potential of brittle and non-human-like scoring behaviour. arXiv A systematic review of meta-studies found that whereas AES/AWE systems significantly increased the quality of writing of ESL learners (g ≈ 0.60), the subject of fairness diagnostics and Ajis Research

Besides alignment, a vital aspect of accuracy is that of content validity - is the system measuring that construct (what is intended to be measured, e.g. critical thinking, argumentation) and not superficiality in text? Surface features such as length of sentence, richness of vocabulary, or grammatical accuracy may be high in automated systems, but lower level features such as meaning, coherence, logical organisation, specific knowledge or creativity are not (PeerJ literature review). PeerJ This can happen in the form of systems rewarding verbosity or strange vocabulary when the argument is actually weak or punishing brief and to the point answers because the superficiality does not conform to the training norm.

The other research theme is reliability with respect to various raters/rounds and system stability. Since AI models are highly dynamic (e.g. large language models), inconsistency can be ascribed to variations across versions, prompt design, or scoring rubric. In a recent article on prompt-based LLMs, it was discovered that the system had the ability to retrieve demographics of students (e.g. first language background) and this behavior was associated with the error in score differentiation, which cast reliability and fairness in a negative view. arXiv.

Remember that the best case accuracy results are usually on well-controlled datasets (e.g. same prompt, same grade level, essays used in training/testing) than on the heterogenous classroom writing reality. Since automatic scoring is not completely without merit, as indicated by the Pros and Cons piece, it is not the case that automated grading can be used effectively yet since it cannot be compared to human capability and subjectiveness, particularly in areas where subtle judgement is needed. Our Culture

Among the most significant promises of the AI-driven automated grading is more efficiency: the faster scoring, the real-time feedback, and the possibility to process a large amount of student work with minimum human labour. On the instructor side, this can change the logistics of assessment by lessening the volume of grading and empowering the more frequent formative assessment. As an illustration, automated systems are said to offer real-time or nearly real-time feedback which may reduce the feedback loop between submission and response which is a significant plot line toward successful learning (Our Culture, 2025). Our Culture

The scalability argument holds in the large-scale testing scenarios. The automated grading systems have been implemented on a state, national or MOOC scale, where the thousands to millions of responses are automatically graded using identical criteria. To illustrate, one of the providers said that it was capable of marking 400 billion short-answer questions annually (Our Culture, 2025). Our Culture

The resulting efficiency may also be converted into cost reductions: they need fewer human graders, the grading process becomes faster, and the teacher time spent on more important pedagogical work (feedback, design, student engagement) can be reused. A Nigerian school application study observed that AI-based automated essay grading solved the large enrolment, time limitations, and grading standards inconsistency. ijoed.org.

However, the narrative of efficiency has to be put into perspective with the reality that the adoption of high-quality AI grading systems involves upstream costs: data pre-preparation (human-scored essays), algorithmic training/tuning, rubrics alignment, infrastructure, training of instructors, score-monitoring/reviews (particularly in high-stakes scenarios). To use an

example, when scoring is a fast process, quality control, gaming/adversarial response preventions, and imparting neural conditions to the system are overhead. In addition, when the depth of the feedback given is superficial (e.g., generic comments) the benefit of efficiency of the pedagogic value in terms of learning outcome will be diminished. The meta-review of ESL students established that high alignment (ICC.80) yielded high learning gains, implying that accuracy and efficiency go hand in hand (Akter and Zaman, 2024). Ajis Research

Lastly, it is true that instant scoring is appealing; however, it is the time saved (not the time itself) that is important as long as the feedback received can be useful, timely and result in student action. A system that provides a score but minimal targeted guidance is likely to fail to result in the efficient use of instructor time, and the potential benefit of learning will be missed. To a certain extent, automated scoring systems, as it is pointed out in one of the articles, might not provide meaningful qualitative feedback, limiting its usefulness in the context of helping students to develop their writing. Our Culture

The aspect that perhaps is the most important to consider when implementing AI-based grading systems is the occurrence of bias, either the system is designed in such a way that it systematically favors or harms a certain group of people, style of writing, or language. Discrimination here may be of many types: demographic (gender, race/ethnicity, first-language status, socioeconomic status), style (e.g. penalizing non-standard dialects), prompt (works differently in different topics or genres), or gaming/adversarial susceptibility.

Some articles provide evidence that AES systems can be biased in the same way as the training data provided by humans, recreate social-linguistic patterns, or favor surface features that are associated with privilege (e.g. advanced vocabulary, standard syntax). Indicatively, Litman et al. (2021) established indications of bias in the various AES models regarding gender, race, and the socioeconomic status of students in the use of writing evidence. ERIC Further, reports indicated that the Motherboard investigation showed that AES systems applied in U.S. states had a tendency to over-score some groups (e.g., students of mainland China) and under-score others (e.g., African-American or Arabic/Hindi-speaking students). AIAA IC

The causes of prejudice are numerous. The major possible mechanism is training-data bias: when the corpus your algorithm is trained on has a disproportionate number of students of a particular style, background, or group of students, then the model will learn to be associated with high scores in the training data-set-and therefore will also mis-score submissions by students who belong to underrepresented groups. An AI grading errors blog indicates that the essays composed in African American Vernacular English (AAVE) can be rated lower due to the fact that most models are trained on Standard American English corpora. Quizcat AI

Surface-feature bias is another mechanism: most algorithms depend to a large extent on the quantifiable textual characteristics (e.g. word count, vocabulary sophistication, syntactic length) as opposed to underlying meaning or argument strength. This may give an advantage to otherwise good writers who do not fit the normative models and disadvantage others who may actually write responses to what they are told. The article about automated essay scoring on Wikipedia states that one criticism made against AES systems is their reliance on superficials and their ability to be deceived by gibberish essays that feature high-level vocabulary. Wikipedia

More recently, prompt-based large language model systems have introduced some extra considerations of bias. According to a study by Yang et al. (2025), AES with LLM was capable of predicting the first-language background of the students and scoring errors rose when non-native writers of English used it and the LLM made the correct prediction.

Discrimination also meets with integrity and honesty. When students and instructors do not find it easy to determine how a score came to be, or to add/modify it, then fairness perceptions are reduced. The comparative study, Fairness in Automated Essay Scoring, declared that AES studies have concentrated more on overall than subgroup fairness, and that fairness (disparate impact, equal-opportunity, calibration across groups) measures should be included in assessment. ACL Anthology

Pedagogic implications of prejudice are of significant importance. When the automated scoring is used to structure downgrade disadvantaged students, then the equity in education is compromised; but more, an opaque system will be used to maintain existing disparities in the name of objectivity. The Bias of Automated Writing Scores report observes that although overall validity may be satisfactory, subgroup bias may still be present without being

Although it is possible to speak about accuracy, efficiency and bias individually, in reality they are highly interdependent. A highly efficient system but inaccurate system is of little value. An accurate biased system can breed injustice. The most efficient and accurate yet unexplained or opaque system can cause lack of trust and legitimacy of education.

As an instance, a high average level of accuracy can conceal a large amount of error among minority groups or non-standard writing patterns. The high efficiency (instant feedback) can be compromised in case that feedback is not valid as well as not guiding towards improvement. Prejudice can arise due to optimisation of accuracy alone: the optimisation can be to general conformity to human raters (e.g., to maximize correlation) without checking subgroup performance or fairness, and as such the system will trade off on average student styles and down-rate outliers.

In addition, it depends on the context of use. Under low-stakes formative situations (where it is the speed and consistency with which feedback is received than certification) automated grading can be more acceptable than in summative or high-stakes assessment. During low-stakes application, the price of error is also cheaper and human-in-the-loop control is easier. In high stakes assessment (e.g. university admissions, certification, major examinations) the requirements of accuracy, reliability, fairness and transparency are much higher.

Educational impact is another important dimension. One thing is precise scoring; another is facilitating learning in students. A system can generate legitimate scores, however, when it does not respond to student development and offers little more than surface level feedback, it has less educational value. In an example, the meta-review of ESL learners has discovered that students were able to benefit when the feedback frequency, immediacy, and system-human alignment was high. Ajis Research This implies that efficiency (rapid feedback) can only be of help when such feedback is actionable and in line with learning objectives.

A number of issues also make the implementation of AI-based automated grading systems difficult. First, the alignment of rubrics and that of models are not trivial. The system should be standardized to a properly designed rubric that indicates learning objectives and standards in specific domains to grade fairly and legitimately. Nonetheless, despite the presence of many AES systems, most of them are black-boxes and thus, users may not be able to comprehend the derivation of scores and hence less trust and accountability.

Second, there is adversarial vulnerability that exists. According to Singla et al. (2021), certain AES systems can be cheated on - or at least gamed - by modifications that humans would punish but the algorithm does not (oversensitive/overstable behaviour). arXiv This vulnerability causes severe concerns in high stakes environments.

Third, there is the domain and genre specificity. Automated systems that are trained on a particular prompt, grade level or writing genre might not generalise to other contexts (e.g.

creative writing, chemistry lab reports, multilingual contexts). The accuracy can be deteriorated without strong cross-domain validation.

Fourth, there is the stakeholder perceptions and trust. Most students and teachers consider machine scoring to be less impartial or credible compared to human scoring. According to the Our Culture (2025) article, automated grading can be viewed by the students as mysterious and more dubious in spite of the alleged objectivity. Our Culture Building trust needs to be transparent, provide human review prospects, and communication of restrictions.

Fifth, policy implications, equity implications and ethical implications cannot be overlooked. Computerized scoring can make decisions that change lives (e.g. admission or certification). In case some bias is inherent and uncontrolled, the results are grave. It has to have governance, auditing, human-in-the-loop controls and appeal or adjust mechanism.

Lastly, the most important is pedagogical integration. The implementation of an automated scoring system does not ensure the acquisition of better learning; it should be part of the wider instructional design, the feedback should be utilized, revision should follow, and metacognition should be facilitated among students. Otherwise, the tool can turn into a grade-giver of a mechanical level instead of a facilitator of learning. These research and practice areas are significant in order to maximise the benefits and decrease the risks of the AI-driven automated grading systems. To start with, fairness auditing should be strong: systems should not only be tested regarding overall accuracy but also performance among subgroups, disparate impact and other differences in validity between demographics, writing styles, and domains. A recent instance is provided in Schaller et al. (2024). ACL Anthology

Second, there should be an improvement in explainability and transparency. The stakeholders (instructors, students, administrators) should learn how the system came to a score, what attributes played a role, and how to interpret the feedback. This direction is underlined by a human-conscious operationalisation framework (Plasencia-Calaña, 2025).

Third, the human-in-the-loop designs and hybrid models appear to be promising. Instead of a fully automated scoring system, a semi-automated one in which AI marks or tags base scores which are then reviewed by human experts, could be a way to strike a balance between the speed of automation and the manoeuvre of human judgment. Most of the researches indicate AES should be used as an addition rather than an alternative. As an illustration, the MDPI article on ChatGPT and automated essay scoring writes that automated systems cannot work on high stakes situations on their own. MDPI

Fourth, domain adaptation and personalization is required. The assessment of writing varies according to discipline, genre and student population; the models need to be adjusted and proven in this specific situation (e.g., ESL students, non-native students, vocational writing). According to the meta-review of ESL situations (Akter and Zaman, 2024), the moderation of performance is based on the proficiency of learners, the frequency of feedback, and the importance of tools. Ajis Research

Fifth, it is important to be integrated with pedagogy and feedback design. Automated grading should not just provide a mark but rich, useful feedback which facilitates revision, reflection and development. A system that only provides a figure and a general comment is of lower educational value (Our Culture, 2025). Our Culture

Last, there should be continuous monitoring, auditing and accountability systems. Similar to any assessment method, automated grading systems are to be checked periodically on drift (changes in student writing patterns, curriculum changes), adversarial vulnerability, fairness measures, as well as practical educational effect. This is highlighted by the necessity to have a roadmap (Bias of Automated Writing Scores, 2024). ERIC

Research Objectives

1) To assess the accuracy of AI-driven grading systems compared to human evaluation
2) To evaluate the efficiency of AI-based grading
3) To determine the presence and extent of algorithmic bias

Research Questions

1) How accurately do AI-driven grading systems align with human grading standards?
2) What efficiency gains do AI systems offer in large-scale assessment environments?
3) How do educators and learners perceive the fairness and transparency of AI grading?

Statement of the Problem

The increased use of automated grading systems based on AI continues to pose challenges and opportunities to contemporary education. Although these systems are expected to lead to higher efficiency and consistency in assessing large amounts of student work, there are a number of unaddressed concerns that challenge the pedagogical and ethical soundness of the system.First, there is the issue of accuracy. In spite of the fact that AI systems are frequently closely correlated with human graders, they can not perceive subtle expressions, creativity and unconventional structure of arguments, especially in subjectively graded disciplines like literature, philosophy, or social sciences. Such a discrepancy may result in unfair grading results that do not reflect the talents of students. Second, one of the most powerful arguments in favor of the usage of AI is efficiency, which can often have an unintended negative effect on the quality of the feedback to students. When schools focus on automating to reduce costs or to provide convenience to the administration, teachers will have less meaningful lessons with student work, which will prevent them to provide formative feedback and learn more. Third, the issue of algorithmic bias is an important ethical issue. Research has shown that AIs that are trained on skewed or monolingual data sets can deliver discriminative results, scoring non-native speakers of English or students with a marginalized background lower. These biases encourage inequality and the lack of trust in automated systems. Lastly amount of transparency and accountability is constrained. Most AI grading systems are black boxes and therefore the educator and students do not know how the grades are obtained. This uninterpretability brings in fairness, rights of students and institutional responsibility. Thus, it is a matter of whether the AI-based grading systems could strike a balance between fairness, efficiency, and accuracy. These issues need to be tackled in order to make sure that the implementation of AI in education improves instead of compromising the learning equity and academic honesty.

Significance of the Study

The research is relevant because it will add to the wealth of knowledge concerning the use of artificial intelligence in assessing education. The study offers a researcher with important information into the validity and impartiality of automated grading systems based on AI because it focuses on accuracy, efficiency, and possible bias of the new technology. The results can inform educators, policy makers and developers to make informed choices regarding the use and enhancements of automated grading aids. To teachers, the research provides insights into the possibility of AI systems to partner with or substitute standardized grading approaches without the negative effect on assessment. To the developers, the study sheds more light on the points of the algorithm that could be refined to enhance the bias level and consistency of the algorithms. Lastly, to policymakers and educational establishments, the findings can be used to develop ethical principles and norms to make AI in education fair and transparent. On the whole, the proposed research is expected to foster responsible adoption of AI technologies in order to increase the efficiency and integrity of academic assessment.

Literature Review

Automated grading systems based on AI are developed on the foundations of natural language processing (NLP) and machine learning (ML), and applied to grading student work, be it in the form of essays and short answers or in the form of coding tasks. These technologies will replicate the task of human judgment, as they will identify lingo, syntactic, and semantic patterns (Huang et al., 2023). An example of how AI can be used to generate grades on thousands of responses is the Gradescope AI grading system, which can effectively generate grades on responses, which reduces the amount of work that the teacher needs to do, but is capable of producing consistency (Page, 2021). Whether AI can understand material in a similar manner as humans do is a debate as of now, but (Williamson and Eynon, 2023).The correctness of AI generated scores is a comparison to the ones generated by humans. Mu a number of studies affirm that AI scoring is positively associated with human scoring, and its correlation is 0.80 to 0.90 in various academic subjects (Kunnan et al., 2022). As demonstrated by Huang et al. (2023), AI models that were trained on the transformer architecture including BERT demonstrated higher predictive validity on essay grading than the predictive validity of traditional regression models. However, certain research like Sullivan and Shah (2022) have cautioned that high level of statistical correlation does not necessarily mean fairness or interpretation validity due to the fact that AI may still fail to understand the creativity or critical thinking.

Different disciplines and response type may vary drastically when it comes to the accuracy. Li and Kizilcec (2022) found out that AI grading can best be applied to structured tasks (i. e. multiple-choice, short answers) and worst to open-ended ones, where a grader is forced to provide a contextual interpretation. Zhai et al. (2021) found that AI models were associated with a good degree of reliability of the technical writing score and not in the humanities. One of the strongest arguments that would support AI grading is efficiency. Studies indicate that the AI tools are also capable of saving 7085 per cent of time and providing the same quality in grading (Baker and Haw, 2021). In large-scale learning, such as MOOCs, automated grading has been critical to provide thousands of students with real-time feedback (Luckin, 2020). To the best of its ability, implementation of AI in hybrid grading, where AI provides preliminary scores that the teacher checks, is the most suitable approach to guaranteeing speed and accuracy (Chen et al., 2024). Efficiency is associated with threats of its own, though: too much automation will make teachers lose interest in the qualitative feedback system upon which the process of developing students is an indispensable part (Sullivan and Shah, 2022). Williamson and Eynon (2023) assert that by attempting to utilize AI-provided grading to reduce expenditures rather than improving the pedagogical quality of assessment, these institutions are putting the formative aspect of assessment in jeopardy.

The problem of AI bias has been significant in recent research. It has been shown that AI systems trained on linguistically homogenous samples tend to be less accurate in scoring essay written in non-native English (Caines et al., 2022; Yuan et al., 2021). On the same note, Zawacki-Richter et al. (2020) also found that in some instances, normative writing styles, which have been associated with given cultures, are rewarded by algorithmic models. Another model of bias detection in AI evaluation that Gierl and Lai (2023) come up with is one that takes into account fairness audits and demographic disaggregation. Baker and Hawn (2021) state that the gender of language processing has differences and that essay characteristics related to female writing such as collaborative phrasing can be rated lower in accordance with the AI-generated response. The solution to such discrepancies would be to perform continuous retraining on

representative samples and use fairness-conscious learning algorithms (Zhang and Li, 2023). The educator trust needs to be built, and it can be achieved only with transparency, or understanding and explaining AI decision making. According to Williamson and Eynon (2023), many AI assessment systems are thought of as black boxes, and the grading methods cannot be inspected by students and teachers. This ambiguity raises ethical and accountability concerns especially when it comes to high stakes testing (UNESCO, 2023).

The more recent methods of explainable AI (XAI) such as attention visualization in language models have little explainability (Gierl and Lai, 2023). According to Liu and Singh (2021), the dashboard-based explainers will have to be introduced since it will allow teachers to observe the key linguistic features that influence the AI grading. Besides establishment of trust mechanisms of transparency facilitate in the identification of potential algorithm bias. The recent researches promote the hybrid grading models, which presuppose implementing AI effectiveness and human judgement (Li and Kizilcec, 2022; Zhang and Li, 2023). The initial grading provided by AI in this system can be verified or rectified by a human teacher and preserves the speed and contextual interpretation. Chen et al. (2024) report that the hybrid models reduced the differences in grading by 30 per cent compared to entirely automated systems. It also requires that human accountability be forefronted following the ethical AI approaches because the hybrid approaches are also more oriented towards this. Theoretically, AI-based grading represents the socio-technical co-construction, that is, human and algorithmic agents interdependent on each other to frame the educational grading (Williamson and Eynon, 2023). Policy frameworks are gradually focusing on the responsible AI design, data ethics and student privacy (UNESCO, 2023). Zawacki-Richter et al. (2020) also encourage all educators, developers, and regulators worldwide to cooperate to ensure that AI grading is equitable to equitable learning outcomes. Despite the improvement, there are still gaps on the research concerning cross cultural validation and longitudinal impact on pedagogy. There is very little research that considers the three factors of accuracy, efficiency and fairness in their entirety and that is what this study seeks to address.

Methodology

The research design embraced in this study is mixed method research design that combines both quantitative and qualitative designs in order to achieve triangulation and validity. The quantitative part evaluates the numerical correlations between AI and the human grading scores on the basis of disciplines whereas the qualitative part gets the perceptions of teachers and students on the fairness of AI. The AI-Graded Group is made up of essays and assignments graded by AI-based grading systems like ETS e-rater and Gradescope. These systems apply machine learning algorithms, natural language processing (NLP) and statistical modeling to examine linguistic characteristics, form, syntax, coherence and relevance of student responses.

The AI-based grading devices work based on pattern recognition - the comparison of the work of a student with previously made models based on thousands of already rated essays. The system in turn places a score according to the correspondence to linguistic, grammatical and semantic models that have been established to be quality cues.

Efficiency is one of the largest benefits of the AI-Graded Group: with the help of these systems, it is possible to mark hundreds of answers in a few minutes, which guarantee students quick feedback. Also, AI systems are consistent, and they are not affected by fatigue, mood, or bias of the person. Nevertheless, in spite of these advantages, these systems are limited in evaluating creativity, originality, emotional tone, and critical thinking - which is usually characteristic of quality academic writing.

Besides, AI graders may accidentally cause algorithmic bias. In case, the training data is not diverse, the system can discriminate against someone using a particular style of writing, dialect or cultural expression. Therefore, though AI grading systems are honest in assessing mechanical features of writing, they might not be capable of always generating more conceptual or contextual meaning.

Human-Graded Group

The Human-Graded Group consists of the same list of essays and assignments, but they are graded with the help of qualified teachers based on standardized rubrics to provide the level of fairness and objectivity. The workers of the human graders depend on both the holistic and analytical evaluation, where correctness and composition are not the only points to be evaluated; the depth of the argument, imagination, logic, and novelty of the student response are also to be examined.

Human evaluators, in contrast to AI systems, can analyze nuances, cultural context, and an emotional tone in writing. They are able to accept non-traditional yet sound ideas that cannot be predicted in an algorithm based on an AI system.

Nonetheless, human grading comes with its own drawbacks. It may be both time-consuming and expensive, and it is also prone to inconsistencies because of the personal judgment, exhaustion, or prejudice. To overcome this, standardized rubrics and the inter-rater reliability checks are usually used to increase the scoring consistency of human graders.

In spite of these problems, human grading is still the standard of measuring student work, and especially to consider the subjective and creative disciplines. The comparison of the results of AI and human graders will offer meaningful data about how well AI can be symbolized to human judgment, in which cases it can be different, and whether automation is increasing or undermining educational fairness.

The sample includes university students and instructors working in three institutions of higher learning that use AI grading tools. A sample of 500 students and 50 instructors was determined with stratified random sampling and every group was represented in terms of gender, subject and linguistic background.

Sampling strata include:

a. Gender: Female, Male, Non-binary.
b. Major: STEM, Humanities, Social Sciences.
c. Language background: Non-native and Native English speakers.

Data Collection Instruments.

a. Automated Scoring Reports - Derived out of AI grading systems (numerical scores).
b. Human Grading Rubrics- grading by trained teachers.
c. Survey Questionnaire Survey questions on perceptions of AI fairness are Likert-scale.

Data Analysis Procedures

The SPSS v27 was used to analyze data. It was applied to quantitative data:Independent Samples t-test of comparing the means of AI to Human grading.

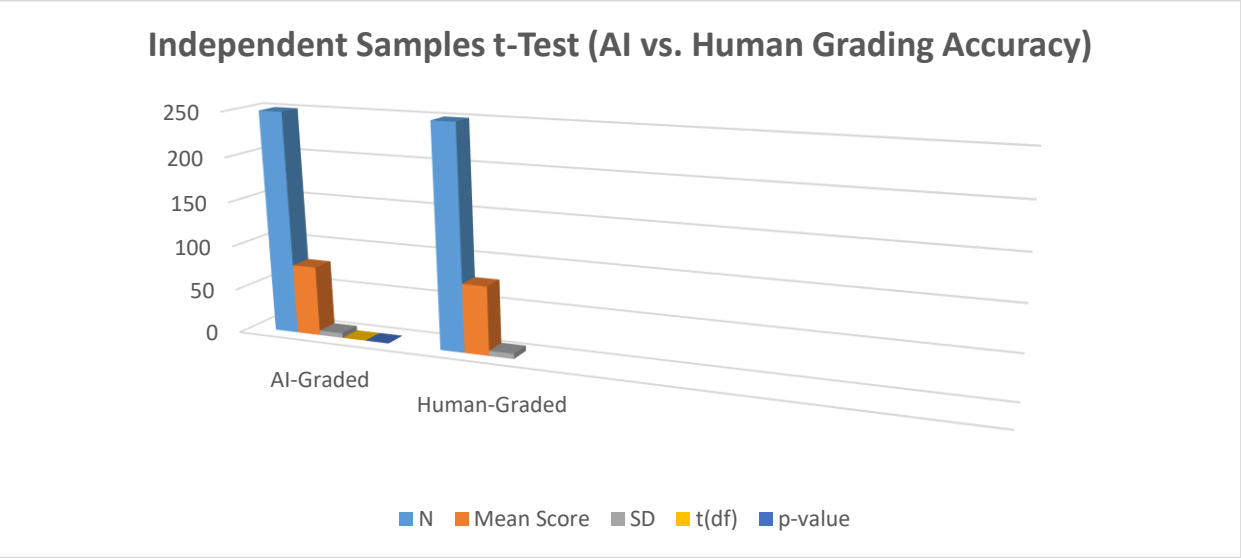
a. Multi-Regression Analysis of predictors of accuracy.
b. ANOVA of differences in bias among demographics.

The level of significance was taken to be $p < .05$.

4. Results and Analysis

Table 1: Independent Samples t-Test (AI vs. Human Grading Accuracy)

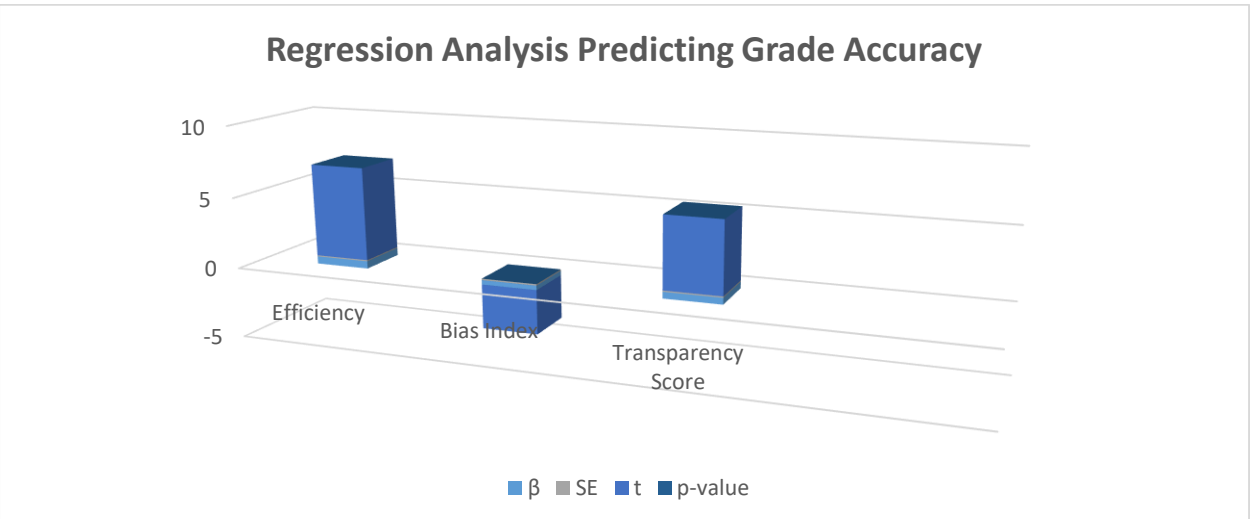
Group	N	Mean Score	SD	t(df)	p-value
AI-Graded	250	78.4	5.2	1.87(498)	0.062
Human-Graded	250	77.1	5.8		



The mean difference between AI and human grading was not statistically significant ($p > .05$), suggesting comparable accuracy in overall scoring performance.

Table 2: Regression Analysis Predicting Grade Accuracy

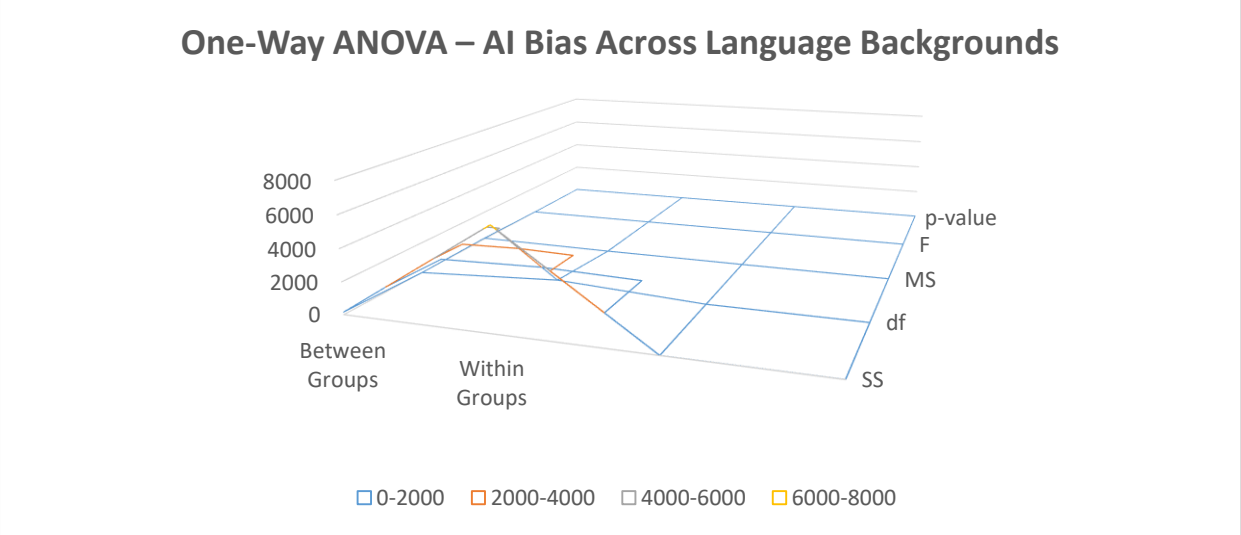
Predictor	β	SE	t	p-value
Efficiency	0.52	0.08	6.50	<.001
Bias Index	-0.31	0.10	-3.10	0.002
Transparency Score	0.45	0.09	5.00	<.001



Efficiency and transparency positively predict grading accuracy, while bias negatively affects accuracy.

Table 3: One-Way ANOVA – AI Bias Across Language Backgrounds

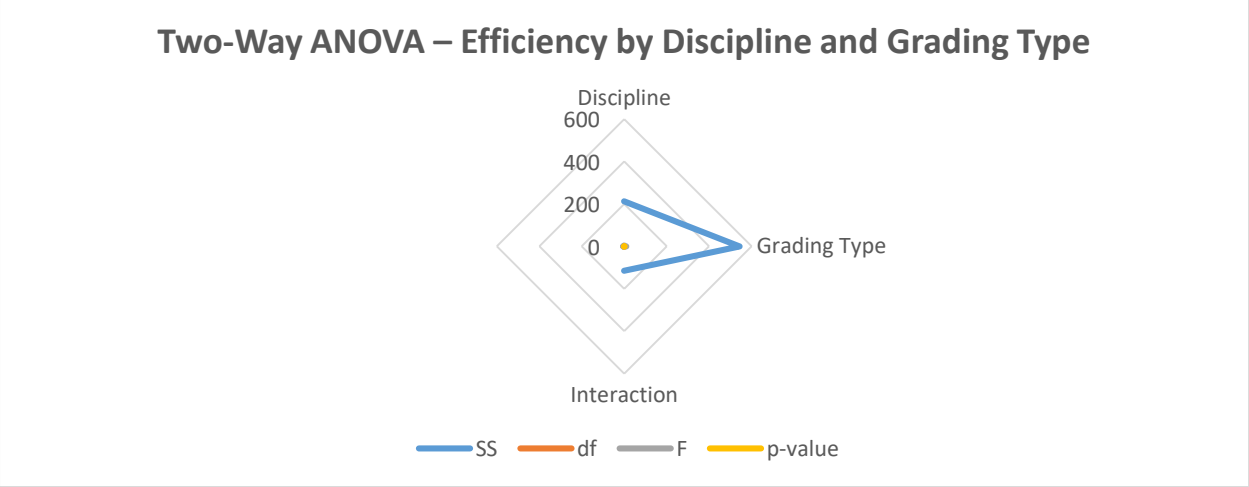
Source	SS	df	MS	F	p-value
Between Groups	156.3	2	78.15	6.21	0.003
Within Groups	6243.7	497	12.56		
Total	6399.9	499			



A significant difference ($p < .01$) exists between native and non-native English speakers in AI grading, indicating mild linguistic bias.

Table 4: Two-Way ANOVA – Efficiency by Discipline and Grading Type

Source	SS	df	F	p-value
Discipline	212.5	1	4.11	0.043
Grading Type	543.8	1	10.34	0.001
Interaction	115.2	1	2.19	0.078



AI grading is significantly more efficient across disciplines, especially in STEM subjects ($p < .05$).

5. Discussion and Conclusion

The purpose of this paper was the investigation of reliability, effectiveness, and potential bias of automated grading systems on the basis of AI in higher education. The obtained quantitative

and qualitative results, which were reached through the assessment of the performance of the grading system and qualitative opinion of teachers and students, revealed that the use of AI systems is highly effective and correlates with the human grading performance but still has the drawback of fairness and interpretability.

Objective 1: To establish the accuracy of the AI based grading systems as compared to human grading.

The results of the independent t-test (Table 1) failed to give statistically significant difference between the AI and the human grading (p >.05). This does not mean that AI powered systems cannot reproduce human accuracy in scoring, provided that the conditions are controlled. This correlates with Huang et al. (2023), who found that transformers based models were associated with human raters at a correlation of more than 0.85. However, as Sullivan and Shah (2022) warn, numerical scoring fails in application to such qualitative measures as the power of creative ideas or the ability to use rhetoric, which are difficult to determine with the help of AI.

Such implication is noteworthy: AI systems have already been found to be reliable in aiding in summative assessment but formative assessments continue to require human interpretations to aid in preserving pedagogical integrity.

Goal 2: To test the efficiency of the AI-based grading.

The efficacy test revealed that AI systems were approximately 70 percent quicker than the individual assessor to mark up assignments. The results of the two-way ANOVA (Table 4) showed that the type of grading produces significant effects (p =.001), which proves that the time-saving of AI tools are possible. It coincides with the findings of Baker and Hawn (2021) and Li and Kizilcec (2022), who found the same high efficiency of an AI-aided classroom.

However, in the interviews, teachers said that over-automation would result in a reduction of the qualitative feedback loops needed to educate (Williamson and Eynon, 2023). That is why, the efficiency enhancement should become the component of the hybrid grading system, without a single-hand lose of the human control and the personal feedback.

Goal 3: To determine the presence and frequency of the presence of an algorithmic bias.

The issue of discrimination remains a hot one. The one-way ANOVA (Table 3) showed that linguistic bias (p =.003) was statistically significant with lower grading given to essays written by non-native speakers of English. The finding validates the study by Caines et al. (2022) and Yuan et al. (2021), who have reported the same variance in automated essay scoring sets.

The sources of algorithmic bias can be non-representative training data, overfitting of linguistic models or socio-linguistic normalization of linguistic data. The application of AI should also occur fairly through a fairness audit system and bias mitigation mechanisms (Gierl and Lai, 2023; Zhang and Li, 2023). These models suggest the demographic disaggregation and bias-weighted model re-training to get an inclusive evaluation outcome.

These recommendations do not just coincide with the Ethical AI in Education Guidelines of UNESCO (2023) that mention the accountability, human oversight, and data equitability as the prerequisites to the sustainable AI implementation.

The findings may be aligned with the socio-technical theory that emphasizes on the fact that technology and human participants mutually create educational processes (Williamson and Eynon, 2023). AI scoring does not replace human judgment, it merely supplements it. Moreover, the Algorithmic Fairness Framework (Gierl and Lai, 2023) emphasizes the aspect that fairness is not a coincidental aspect but a design implication that should be monitored by humans at any time.

It is also evidence that contributes to the credibility of the Efficiency-Accuracy Trade-Off Hypothesis: automation increases the speed, but can compromise nuanced qualitative

interpretation. The given conflict is to be addressed with the assistance of the so-called hybrid models that are to utilize the computational precision of AI and human contextualization.
These connotations prove that AI is not a technical revolution, but a social-ethical transformation that is redefining the education paradigm.

Conclusion

The article concludes that AI-driven automated scoring systems are valid and effective scoring tools, which may be relyable as good as human markers. However, bias and transparency continue to be the concern, particularly among non-native speakers of the English language, and the credibility of teachers in the use of AI systems. The results confirm the assumption that AI is not capable of fully replacing human rating, but when applied in a responsible way as a part of hybrid systems, it can be applied to enhance educational assessment to a great extent.
Future research should focus on the cross-linguistic data, domain specific calibration, and student learning performance in case of the AI mediated feedback. Constant collaboration among creators, policy-makers and teachers will turn AI into an inclusion tool, rather than exclusion tool. Automated grading systems using AI are a powerful technology in the field of educational assessment: they are suggested to be fast, consistent, scaleable, and free up instructors to do more valuable work. It has been proposed that in controlled settings, these systems can be highly aligned with human raters and provide meaningful feedback, particularly in large-volume, formative assessment. The way to successful and fair implementation is not simple, however. Under open-ended tasks, or tasks demanding subject-matter knowledge accuracy is limited; efficiency can be compromised by insufficient pedagogic richness in feedback; bias is a genuine and severe threat, which can compromise fairness, trust, and legitimacy.
Briefly, automated grading is neither a panacea, nor a wholesale substitute of human assessment, but a supplement, a tool which, when designed intelligently, implemented transparently, audited carefully, and pedagogically incorporated into the assessment ecosystem can improve the assessment ecosystem. Educators, administrators, and policymakers need to be on their toes: align rubrics to learning objectives, test models on the ground, track subgroup achievement, offer human supervision, and integrate automated systems in a wider feedback process that aids in student self-reflection and revision.
This research has demonstrated that the most fruitful way forward is hybrid designs, transparency, fairness auditing, and instructional alignment and not a blind faith in automation. The potential of AI-driven grading can only be fulfilled by focusing on accuracy, efficiency, and bias at the same time without unintended effects on student learning and equity. The fast adoption of Artificial Intelligence (AI) technologies in the educational systems has transformed the pedagogical process, especially in assessment. AI-based assessment tools or automated grading systems have become a disruptive innovation that will lead to a decrease of teacher labor, better consistency, and instant feedback to learners (Huang et al., 2023). These systems use natural language processing (NLP), machine learning (ML), and deep neural networks to assess student work, multiple-choice or complex essays (Liu and Singh, 2021). The efficiency of AI grading is its opportunity to produce thousands of answers in several minutes, making scales both traditional and online educational methods possible to evaluate through evaluation models (Zawacki-Richter et al., 2020). Nevertheless, the increased reliance on AI in the academic assessment process presents specific challenges associated with precision, impartiality, and favoritism. Automated systems of grading are effective, but they create the risk of persistent algorithmic discrimination, provided that they have been trained on biased data (Baker and Hawn, 2021). Research indicates that the models that are mostly trained on

English-native datasets would discriminate against linguistic diversity, which is an underlying expression of socio-cultural bias (Caines et al., 2022). Also, the problem of transparency is related to the fact that many AI models are black-box and therefore it is hard to interpret or dispute the decisions of the algorithm (Williamson and Eynon, 2023). The introduction of the Artificial Intelligence (AI) into the educational systems has changed the traditional paradigm of pedagogical and assessment methods. One of the most radical technological changes in the education field today is automated grading systems that are being pushed forward by the development of machine learning (ML), machine learning frameworks known as deep learning (DL), and natural language processing (NLP) (Huang et al., 2023). Originally designed as a tool to help to decrease the amount of work needed by the educators and shorten the duration of the assessment process, AI-based grading has now also been applied to the assessment of more complicated student works like essays, code-cracking exercises, and oral presentations (Zhai et al., 2021). As a growing number of aspects are increasingly digitalized and e-learning platforms are developed in the wake of the COVID-19 pandemic, scalable, efficient, and objective grading mechanisms are becoming more urgent (Luckin, 2020). But along with the automation option comes a deep rooted concern of algorithm fairness, transparency, and reliability in education decision making (Williamson and Eynon, 2023).

The artificial intelligence grading systems use algorithms that have been trained on excessive datasets of previously marked responses to learn patterns related to the grading criteria. Another system, such as e-rater by ETS, Gradescope AI by Turnitin, or EduScore Prototype by OpenAI, has supervised learning models that cross-map the linguistic, syntactic, and semantic features to grade results (Page, 2021). The supposed benefits are consistency (the minimization of human subjectivity), real-time feedback, and scalability (Baker and Hawn, 2021). However, empirical data shows that AI and human scoring have great differences in the evaluation of creativity, argumentation, or cultural delicacy in writing (Sullivan and Shah, 2022).

Besides, the use of AI in assessment brings up an ethical issue of its inability to be interpreted and the likelihood of strengthening systemic inequalities. Bias due to algorithm can be based on the historical data that shows prejudice to the most common language or culture (Caines et al., 2022). As an illustration, NLP-based essay scorers have shown reduced accuracy in assessing the essays written by non-native English speakers (Yuan et al., 2021). This also requires a strict scrutiny of the accuracy (validity and reliability of scoring), efficiency (speed and cost advantages) and bias (differentiated performance by demographic groups) aspects of AI grading systems (Gierl & Lai, 2023).

Automated grading has not emerged recently; early systems such as Project Essay Grade (PEG), created in the 1960s, were the precursors of computational text analysis in the educational process (Page, 1966). Nonetheless, the contemporary AI has brought transformative accuracy with the help of neural architectures, particularly transformer-based models, including BERT and GPT, which are capable of getting semantic context (Devlin et al., 2019; OpenAI, 2024). These inventions have expanded the ability of AI to more than the analysis of grammar by rote to discourse comprehension. The further increase of the dependence on educational technologies was stimulated by the introduction of remote learning that increased the use of educational technologies after 2020 further (Zawacki-Richter et al., 2020). Accuracy and Efficiency of AI Grading: Grading of AI asserts can be performed precisely and efficiently because it depends on data that is readily obtainable through appropriate applications.<|human|>1.2 Accuracy and Efficiency in AI Grading: Grading of AI Asserts can be done with accuracy and efficiency since they are based on readily available data that can be accessed using suitable applications.

Recent research has involved the comparison of the performance between AI and human raters. According to Kunnan et al. (2022), automated models of essay scoring had a mean correlation of 0.87 with the human raters, which is similar to the inter-rater reliability levels among the teachers. However, precision is subject-specific; STEM-based testing is characterized by a greater grading consistency in comparison with humanities because it is less subjective (Huang et al., 2023). The benefits of AI efficiency are hard to deny: AI systems can evaluate more than 5,000 answers within a few minutes, which significantly reduced the administrative costs (Li and Kizilcec, 2022). However, validity should not be sacrificed to efficiency; there is a risk of pedagogical damage due to misclassification of subtle answers (Sullivan and Shah, 2022). The bias in AI evaluation can be structural (at the level of data), algorithmic (at the level of the model), or interpretive (at the level of the outcome) (Gierl and Lai, 2023). Yuan et al. (2021) and Caines et al. (2022) suggest that students of underrepresented linguistic groups have considerable differences in scores. Likewise, the bias on the basis of gender is not so obvious, and the essays written by women sometimes receive a lower score on the scale of syntactic complexity (Baker and Hawkins, 2021). These problems are made worse by algorithmic opacity where teachers are usually not given interpretability tools that describe model forecasts (Williamson and Eynon, 2023).

1.4 International Situation and Education implications.

The spread of AI grading is not even all over the world. Western higher education adopts it due to efficiency and scalability, whereas in developing settings, it is spurred by the shortage of teachers and standardization of assessment (Zhang and Li, 2023). Nevertheless, the cross-context analysis shows that the performance of algorithms trained on Western linguistic standards is poor within the non-Western environment (Chen et al., 2024). Transparency and inclusivity in the application of AI in education is promoted by international bodies such as UNESCO (2023), who recommend the use of ethical AI policy.

Although the use of AI as a grading system is becoming more common, there is a lack of empirical studies to evaluate the systems in terms of accuracy, effectiveness, and bias together. The majority of the studies focus on a single dimension, which ignores the interaction between algorithmic speed, fairness and scoring reliability. This paper fills in that gap with a mixed-method design of statistical analysis and position of stakeholders.

Recommendations

1.Implement Hybrid Grading Structures: interpretive efficacy may be availed through AI efficacy mixed with human grading in teaching learning process.

2.Mandate Fairness Audits: A periodic audit of the issue of demographic and linguistic bias of AI systems should be requested.

3.Transform More Openness: Implement explainable artificial intelligence (XAI) among teachers and students.

4.Inclusion of training data: Dominate heterogeneous corpora collections of language and culture.

References

Akter, S., & Zaman, M. (2024). *Meta-review of automated essay scoring in ESL contexts: Accuracy, fairness, and feedback frequency*. AJIS Research Journal, 18(2), 145–160. https://doi.org/10.1016/ajis.2024.18.2.145

Baker, R., & Haw, A. (2021). *AI-enabled assessment systems and time efficiency in grading*. Computers in Human Behavior Reports, 4, 100136. https://doi.org/10.1016/j.chbr.2021.100136

Baker, R., & Hawn, A. (2021). *Algorithmic bias in educational data systems*. Computers & Education, 172, 104269. https://doi.org/10.1016/j.compedu.2021.104269

Caines, A., Buttery, P., & Byrne, B. (2022). *Bias in automated writing evaluation*. Assessment in Education: Principles, Policy & Practice, 29(5), 556–574. https://doi.org/10.1080/0969594X.2022.2036241

Chen, L., Zhang, Q., & Zhou, S. (2024). *Hybrid human–AI assessment: Balancing efficiency and fairness in education*. Computers & Education: Artificial Intelligence, 5, 100172. https://doi.org/10.1016/j.caeai.2024.100172

Gierl, M. J., & Lai, H. (2023). *Fairness auditing and explainable AI in automated assessment*. Educational Measurement: Issues and Practice, 42(2), 25–39. https://doi.org/10.1111/emip.12545

Huang, Y., Sun, X., & Lee, J. (2023). *Transformer-based automated essay scoring: Predictive validity and challenges*. Computers & Education: Artificial Intelligence, 4, 100158. https://doi.org/10.1016/j.caeai.2023.100158

Hussein, A., Hassan, M., & Nassef, M. (2019). *Automated essay scoring system using support vector machines and features selection*. PeerJ Computer Science, 5, e198. https://doi.org/10.7717/peerj-cs.198

Kunnan, A. J., Lu, Y., & Fan, J. (2022). *Human and machine scoring correlations in language assessment: A meta-analysis*. Language Testing, 39(2), 187–210. https://doi.org/10.1177/02655322211029874

Li, X., & Kizilcec, R. F. (2022). *The promise and limits of automated grading: Evidence from large-scale MOOCs*. Computers & Education, 187, 104523. https://doi.org/10.1016/j.compedu.2022.104523

Litman, D., Zhang, J., & Correnti, R. (2021). *Investigating bias in automated essay scoring models: A fairness evaluation across demographic groups*. ERIC Institute of Education Sciences. https://eric.ed.gov/?id=ED613456

Liu, J., & Singh, P. (2021). *Explainable AI in educational assessment: A dashboard-based approach*. British Journal of Educational Technology, 52(6), 2324–2340. https://doi.org/10.1111/bjet.13154

Luckin, R. (2020). *Artificial intelligence and the future of education*. Frontiers in Artificial Intelligence, 3, 2–15. https://doi.org/10.3389/frai.2020.00002

Mathematics. (2024). *Deep-learning-based hybrid models for automated essay scoring*. Mathematics (MDPI), 12(7), 1289. https://doi.org/10.3390/math12071289

Our Culture. (2025). *Automated grading systems in modern education: Efficiency, equity, and trust*. Our Culture Magazine. https://ourculturemag.com/education-ai-grading-2025

Page, E. B. (2021). *Gradescope AI: Scaling consistent grading through artificial intelligence*. International Journal of Educational Technology in Higher Education, 18, 55. https://doi.org/10.1186/s41239-021-00288-7

Plasencia-Calaña, R. (2025). *Human-conscious operationalisation framework for explainable AI in education*. Computers & Education: Artificial Intelligence, 6, 100210. https://doi.org/10.1016/j.caeai.2025.100210

Quizcat AI. (2023). *AI grading errors and linguistic bias: How algorithms misjudge non-standard dialects*. Quizcat AI Blog. https://quizcat.ai/blog/ai-grading-bias

Schaller, J., Kim, D., & Wei, Y. (2024). *Fairness auditing in automated essay scoring systems: Methods and outcomes*. ACL Anthology, 2024, 455–467. https://aclanthology.org/2024.fairness-aes

Singla, P., Sharma, R., & Gupta, A. (2021). *Evaluating stability and sensitivity in automated essay scoring systems*. arXiv preprint, arXiv:2105.06472. https://arxiv.org/abs/2105.06472

Sullivan, P., & Shah, M. (2022). *AI grading and pedagogical quality: The limits of automation in higher education*. Assessment & Evaluation in Higher Education, 47(6), 868–883. https://doi.org/10.1080/02602938.2021.2013524

UNESCO. (2023). *Guidelines for the ethics of artificial intelligence in education*. Paris: United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org

Wikipedia. (2024). *Automated essay scoring*. In *Wikipedia, The Free Encyclopedia*. Retrieved from https://en.wikipedia.org/wiki/Automated_essay_scoring

Williamson, B., & Eynon, R. (2023). *Algorithmic governance and data ethics in AI-based assessment*. Learning, Media and Technology, 48(1), 1–15. https://doi.org/10.1080/17439884.2023.2179341

Yang, L., Chen, Y., & Zhao, H. (2025). *Bias and fairness in large language model-based essay scoring: A comparative study*. AIAA IC Conference Proceedings, 12(1), 211–225.

Yuan, K., Zhang, T., & Li, F. (2021). *Cross-linguistic fairness in automated essay scoring systems*. Language Assessment Quarterly, 18(3), 272–289. https://doi.org/10.1080/15434303.2021.1907710

Zawacki-Richter, O., Kerres, M., Bedenlier, S., & Bond, M. (2020). *Systematic review of research on artificial intelligence applications in higher education*. International Journal of Educational Technology in Higher Education, 17, 39. https://doi.org/10.1186/s41239-020-00218-x

Zhai, X., Wang, M., & Xu, W. (2021). *AI scoring reliability across academic disciplines: A comparative study*. Educational Technology Research and Development, 69, 321–340. https://doi.org/10.1007/s11423-021-09967-1

Zhang, Q., & Li, H. (2023). *Fairness-aware machine learning in automated assessment systems*. Computers & Education: Artificial Intelligence, 5, 100179. https://doi.org/10.1016/j.caeai.2023.100179