JOURNAL OF CULTURAL ANALYSIS AND SOCIAL CHANGE

Journal of Cultural Analysis and Social Change,
2025, 10(4), 1368-1382
ISSN: 2589-1316

AI-Driven Educational Assessments: Navigating the Intersection of Innovation, Equity, and Ethical Concerns for Future Learning

M K Senthil Babu[1]*, Denish Raja Durai K[2]

[1,2] Assistant Professor Department of English School of Social Sciences & Languages Vellore Institute of Technology, Vellore Tamil Nadu - 632 104. denish.raja@vit.ac.in

***Corresponding Author:** *senthilbabu.mk@vit.ac.in*

Citation: Babu, M. K. S., & K, D. R. D. (2025). AI-Driven Educational Assessments: Navigating the Intersection of Innovation, Equity, and Ethical Concerns for Future Learning. *Journal of Cultural Analysis and Social Change*, *10*(4), 1368–1382. https://doi.org/10.64753/jcasc.v10i4.3021

Published: December 08, 2025

ABSTRACT

The implementation of Artificial Intelligence (AI) to assess examinations and tests is transforming traditional evaluation methods employed by educational institutions. This advancement not only generates the employment opportunities outlined in the present study but also introduces technological progress that becomes more adaptable and offers personalised feedback to students. The development of cutting-edge technologies, although not yet widely accessible, undoubtedly opens new avenues of opportunity for subsequent generations. Given this context, this study conducted a systematic literature review (meta-analysis) of the current use of AI in educational evaluation, paying special attention to computer-assisted marking, personalised testing, and AI-based exam proctoring. Although AI technologies are expected to contribute to objectivity and standardisation, particularly in mass assessments, they also raise difficult-to-solve problems concerning assessment of advanced cognitive abilities, pitfalls regarding bias, and ethical considerations. While computer-based grading applications such as E-rater save time, their ability to assess creativity and sophisticated analysis is limited, often leading to a homogenised writing process. While adaptive testing tools, such as ALEKS (a web-based learning and assessment system), essentially perform some measure of customisation to the educational journey, they could also be somewhat limited in scope, restricting students to answer recall rather than more complex problem-solving tasks. However, language assessments using AI technology may result in an unfair bias, particularly in those with a non-native accent. Similarly, the use of AI tools for proctoring during remote examinations raises concerns regarding the invasion of personal privacy and overmonitoring. This study aims to address the implications of data-driven AI applications for educators and policymakers. It promotes hybrid processes that pair AI applications with human control to ensure fairness, transparency, and equity in academic evaluation.

Keywords: AI-Assisted Assessment, Educational Equity, Ethical Considerations, Innovation in Learning, Future Education Systems.

INTRODUCTION

The realm of education has not remained untouched, as artificial intelligence swiftly progresses across numerous contemporary industries. Within these sectors, AI has emerged as a transformative power in the realm of educational assessment, offering potential for streamlined processes, tailored learning experiences, and enhanced feedback systems. Some popular AI-driven solutions such as Automated Essay Scoring (AES) systems, adaptive testing platforms, and artificial intelligence (AI) processing tools have climbed popularity charts. These innovations are praised for their ability to streamline the grading process, reduce marking time, and extend assessments in ways that traditional methods cannot achieve.

The growing adoption of AI in educational assessment has raised profound concerns. Even though AI can readily evaluate huge chunks of student projects in a jiffy, when it comes to measuring advanced thinking abilities, such as creativity, analytical reasoning, and problem-solving, they often fall short. In addition, artificial intelligence systems based on fixed algorithms can inordinately support prejudice, particularly when assessing students from different linguistic and cultural backgrounds. For example, AI-powered language assessment tools are often skewed toward non-native accents and minority groups in underserved populations. Similarly, AI proctoring systems have come under fire because of potential breaches of student privacy and escalation of exam-related anxiety.

The availability of AI in such learning environments has raised questions regarding data privacy and student autonomy. The more AI technologies collect and analyse huge amounts of individual data, the more we need to ensure that they are transparent in their operation and that we can demand how they work for a given judgement. Moreover, the effect of AI on fairness and equality in academic assessments is an essential question that should be addressed properly, including how AI can be trained and what data are used for training.

This study aimed to explore the current uses of artificial intelligence in education assessment and to assess its merits and demerits. Three key areas were the focus of this study: computer-based marking, adaptive testing, and AI-invigilated examinations. This study examines these applications and aims to predict the long-term impacts of AI on education before finally providing recommendations for harnessing the power of AI in a way that is fair, open, and inclusive.

In education, we are already starting to see the transformative potential of AI, in that it can completely change assessment approaches by providing consistent, unbiased, and scalable solutions. Similarly, automated marking systems only further help in processing vast numbers of students to work quickly; adaptation-enabled examination forms also modify learning pathways to better accentuate individual needs and preferences in the pursuit of more personalised educational pathways. To maintain the integrity of online assessments, AI-based invigilation systems provide remote proctoring, enabling the live monitoring and curbing of cheating. However, the adoption of these technologies is challenging. However, to leverage its benefits completely, it is vital to address the shortcomings of AI such as bias, ethical issues, and potential biases. Moving the technology a little further, the following section highlights some of the major benefits and potential obstacles associated with AI in educational assessments from both positive and negative perspectives.

BACKGROUND STUDY

The use of artificial intelligence (AI) in educational assessments represents a key innovation in a larger international trend. The application of AI to enhance decision-making processes and improve supply chain operations has become widespread across sectors including health and finance (Smith & Johnson, 2021). In the education industry, AI is praised for its ability to improve exam efficiency, scalability, and fairness, especially when traditional human marking methods are bound by labour intensity and variability. In contrast, AI systems present an opportunity to provide consistent and instant feedback while reducing the administrative burden on educators. This potential is why we explore how AI-powered tools can revolutionise educational evaluation.

Lobster scoring has radically changed the way we evaluate things, and AI has also made some headway when it comes to assessment simply using automated systems such as E-rater and Grammarly for grading essays. The advantages of AES in large-scale assessments are crystal clear, as human graders have a very difficult time managing variables, such as the consistency and efficiency of graders (Wang et al., 2020). While these systems can handle a large volume of student work more quickly than their average professor, they struggle with many qualitative features in writing standards beyond the basics of creativity, critical thinking, and argumentation (Perelman 2014). This poses the fundamental question of how best to combine AI systems with human oversight for complete examination of higher-level cognitive abilities. The public review further questions this hybrid model and the fertile ground of AI-human synergy that evaluators stand.

In addition to written feedback, adaptive testing is another AI-based methodology that effectively customises assessments for individual students. Adaptive testing is employed in deliveries such as ALEKS and Khan Academy, meaning that the complexity of questions can change after a learner responds to previous items, ensuring that evaluative procedures are targeted at each individual regarding their true strength or ability (Eggen & Verschoor, 2016). This personalised approach to learning objectives reduces cognitive overload and ensures that student engagement is far greater than that of the traditional methods. However, as will be explained in more detail in other sections, adaptive testing has difficulty assessing higher-order cognitive processes, such as problem-solving and critical thinking (Holmes et al., 2019). Additionally, it has been argued that this can drive the low expectations of students who have already failed to give them relatively easier tasks all the time (Gierl et al., 2017).

Game-based assessments were also examined as another solution for increased engagement by combining educational content with gameplay to increase motivation and participation (Questier et al., 2017). Applications, such as Kahoot! has been explored in recent studies. A study conducted by Dichev et al. (2017) investigated Quizizz, a platform that incorporates game-like features, such as scoring systems and ranking boards, intending to create more captivating educational settings. This has improved short-term recall and student engagement, but it is not clear how effective this learning experience is. Emphasising speed and competition may limit cognitively deep processing opportunities, such as critical reading and problem-solving (Plass et al., 2015). This governance reflects the evolution of gamified assessments integrating pedagogical efforts to support cognitive engagement tasks that may fulfill a higher-level thinking ability and thereby the usage for regular implementation in educational evaluation (Hamari et al., 2016).

Al-powered assessments can potentially detect huge changes, most notably in the areas of speaking, listening, and reading proficiency around the planet. Platforms such as the Duolingo English Test (DET) and Speechace are advantageous for providing instant feedback and scaling-up assessments, especially for remote or underserved areas (Yang et al., 2021). However, like any other AI-infused system, algorithms are biased as one of the primary challenges these language assessments encounter. Several of these tools have been trained on standard English accents (largely American and British), which means that they may provide biased assessments of students with non-majority-language accents or a regional variety (Blodgett et al. 2020). Future analyses should examine strategies to address these biases by developing more representative AI models based on training data that include a broader variety of linguistic and cultural backgrounds (Suresh & Guttag, 2019).

Over the last few years, AI-based proctoring systems have begun to spread worldwide and more acutely in 2020, when educational institutions switched to remote learning because of the COVID-19 crisis. ProctorU and Examity use a combination of technologies such as facial recognition, keystroke analysis, and behaviour monitoring to enforce academic integrity during examinations (Li et al., 2020). Nonetheless, these systems have sparked a massive amount of controversy regarding privacy, surveillance, and algorithmic biases. The government recently reported that 84% of effective facial analysis systems, along with facial recognition software, demonstrated reduced accuracy in identifying individuals with non-Western features or darker skin tone. This limitation could result in biased misidentifications, potentially leading to discriminatory outcomes (Howard & Borenstein, 2018). Moreover, constant monitoring of students during exams could create test anxiety, which could affect their performance. This study aims at continual improvements in the processes of AI invigilation systems based on ensuring that they are ethical by discussing potential solutions to make them fair and accountable (Khosravi et al. 2021).

In short, using AI in educational assessments helps without doubt and has many advantages such as efficiency and scalability. However, considerable challenges remain, especially concerning AI's ability to assess complex areas, reduce algorithmic bias, and adhere to ethical standards. In the following analysis, we will explore how to mitigate some of these challenges by developing hybrid models that balance AI efficiency with human oversight nuances to ensure that AI-augmented risk scoring is inclusive and transparent while also making a true deep learning impact.

RESEARCH GAPS

The lever integration of Artificial Intelligence (AI) into educational assessments offers several possibilities for enhancing efficiency and scalability. Nevertheless, several key areas have recently been identified that require further investigation. The summary of these three knowledge gaps reflects bias mitigation, creation of intelligent assessment systems that are man- and machine-integrated, and the relationship between AI and student wellness. Thus, the purpose of this research is to solve the unsolved problems that illuminate the discussion on socially responsible and effective usage of AI in assessment.

No-Complete Hybrid Models

These results underscore the limitations of an entirely automated evaluation approach, particularly in assessing nuanced cognitive processes and creative problem solving. However, a gap remains in the understanding of how to coalesce the power of AI (efficiency) with human reasoning to overcome these limitations.

Ways to Minimize Bias in AI

Although these studies have identified algorithmic bias, they do not provide us with methods, techniques, or guidelines on how to address this problem. Qualitative Structural Equation Modeling (SEM) reveals potential biases in involving diverse datasets and consecutive bias assessments within artificial intelligence systems to promote fairer evaluations, as illustrated in this study.

Language and Cultural Inclusion

Many artificial intelligence systems, particularly those that process language, are biased towards standard dialects and foreign accents. Therefore, it is important to conduct research to make AI models more general and to better reflect the diversity of languages spoken in the world.

AI in Learning: An Ethical Framework

The latter includes several large-scale research on ethical considerations such as privacy violation, surveillance, and algorithmic bias, which have not been addressed sufficiently. Full investigation is required in the process of developing and implementing ethics frameworks to ensure transparency, fairness, and accountability in any AI-based assessment.

Impact on Student Well-Being

Among the existing studies, we found exam stress issues that arise in AI-monitored testing systems, but greater consideration must be paid to how AI-based assessments may affect students' overall well-being. Further research is required to examine the impact of AI-based and AI-powered evaluations on students’ motivation, engagement, and well-being.

Experimental Proof and Comparative Study

One of the limitations of the current body of literature considering artificial intelligence systems in education is the lack of real-world examples or empirical corroboration that demonstrates how AI systems function (Virvou et al., 2003; 3). Finally, extensive research is needed to compare the accuracy and fairness of AI-based assessment methods those with of their conventional counterparts, as well as how they affect academic performance.

The same study revealed a research gap, thus providing an opportunity to explore underrepresented areas. This will involve building Black Boxes, which in turn would automatically reduce biasing within technology; it should also improve cultural and linguistic diversity, and as further steps should lead to coming up with a range of principles or codes of ethics that can be used for AI applications across the academic space.

In this context, this study aimed to fill these gaps by researching the main components of AI-powered educational assessments using bias, fairness, equity and fairness, and human-in-the-loop. Driven by specific research questions and objectives, this study investigated the ethical use of AI in educational assessments to promote inclusivity and enhance cognitive engagement.

KEY RESEARCH QUESTIONS

1. How do we design hybrid AI human assessment systems that address these limitations, particularly in the context of measuring complex cognitive constructs, such as creativity and critical thinking?
2. What strategies can be applied to challenge discriminatory algorithmic-based educational assessments in immersive learning to provide fair and non-discriminatory results for students from various language and cultural backgrounds?
3. What should be the context of designing AI systems, particularly language tests and voice recognition software, to avoid discrimination against non-native accents or regional speech patterns?
4. What kind of guidelines should be established to ensure transparency, accountability, and confidentiality in AI-driven assessments, especially for high-stake scenarios?
5. What effects do AI-based assessments have on learners’ cognitive-affective aspects such as motivation, engagement, and test anxiety, and how can these effects be mitigated?
6. What about accuracy, fairness, and ability to measure higher-order thinking skills? How precise are AI-driven assessments compared with traditional methods of evaluation?

KEY RESEARCH OBJECTIVES

1. To examine blended AI human assessment systems that integrate the efficiency of automated AI capabilities with nuanced human oversight, particularly in advanced cognitive skill assessments.
2. To explore and deploy methods for reducing bias in AI-driven education technologies, including creating more inclusive training datasets and continuously assessing bias.

3. To analyse AI assessment platforms that support high levels of cultural and linguistic diversity, enabling fair assessment of learners from diverse language backgrounds.

4. To review ethical guidelines such as transparency, data protection, and the use of AI in educational assessments.

5. To assess the impacts of potential AI-informed assessments (and for that matter machine learning) on student well-being, especially with regards to engagement, motivation, and anxiety, and provide feasible suggestions on how to mitigate less favorouble outcomes

6. To investigate and compare the construct validity of AI-based evaluation with conventional methods for advanced cognitive abilities, equitability, and educational outcomes.

REVIEW OF LITERATURE

Recently, the application of Artificial Intelligence (AI) in educational assessment has seen rapid growth and development, with new capabilities for automated marking, adaptive exams, and AI-assisted language testing, showing promise to the educational community. Although these tools provide significant improvements in efficiency, scalability, and real-time feedback, their incorporation involves several complexities. In this survey, we presented a critical perspective on state-of-the-art AI-driven assessment technologies and discussed their promises and limitations.

Automated Essay Scoring (AES) Systems

Manali Bhawalkar (Sophomore, Computer Science at Illinois Institute of Technology): Educational institutions widely use AES systems because of their ability to quickly and consistently assess large volumes of student writing. One will be evaluated according to the whole range of structural features of the text, including grammar, syntax, and vocabulary, as these systems utilise Natural Language Processing (NLP) algorithms that look for surface linguistic elements in the complete text (Shermis & Burstein, 2013). In the shadow of large standardised testing environments where consistency in grading across a large volume of submissions is logistically challenging for human assessors (Wang et al., 2020), AES tools such as E-rater (used by ETS to score TOEFL and GRE tests) and Grammarly may have distinct potential benefits. The main goal of AES systems is to increase efficiency and provide unbiased feedback, reducing educators' macro-managing capabilities, while providing immediate evaluative efforts to students.

Typically, these systems are automated writing assessment tools that calculate scores based on specified linguistic features, and can be tasked with the more straightforward functions of administering tests/providing rubrics to define essays down to the grammatical last decimal place. By doing this, AES systems can maintain consistency in the evaluation process and significantly eliminate natural fatigue or subjectivity. As such, AES tools offer a way to scale assessments and quickly respond to students during high-stakes exams, theoretically leading to better learning experiences (Shermis and Burstein 2013).

Limitations of AES Systems and Some Criticisms

Despite this power, AES systems have been subject to substantial critiques, with many arguing that they cannot assess higher-order cognitive skills, such as creativity, critical thinking, and argumentation (Perelman 2014). Although AES can quickly evaluate the surface linguistic characteristics of student writing, it cannot easily judge the depth of student writing. For example, Perelman argued that algorithms in AES systems tend to prefer formulaic compositions, and the criteria might disadvantage students who write in unique or non-standard ways instead of standard models. This flaw has led educators to doubt the pedagogical force of AES tools in terms of inspiring students to think rationally or innovatively about the media content.

Despite this, many scholars have proposed hybrid systems in which AES and human graders work together to provide a more comprehensive evaluation of students' essay writing performance (Shermis & Burstein, 2013; Lau, 2017). Together, they are better at processing creativity and other nuanced aspects of AI writing that leave machines stuck.

Another concern that is still being discussed in AES systems is fairness. Research has shown that these systems exacerbate inequalities for some groups of students (e.g. non-native English-speaking students) and those who do not conform to the canonical norms of academic writing in their deployment of 'unconventional' writing styles (Wang et al. 2020). The training data on which these systems rely are susceptible to algorithmic bias, meaning that they might reflect societal or institutional inequalities. To mitigate these biases and guarantee fair assessment for students regardless of their background, some researchers have advocated against full automation (Blodgett et al. 2020).

Bias and Fairness in AI

The implementation of AI-driven assessment systems in educational settings has led to serious concerns about bias. Trained on large datasets, these AI technologies are more likely to inadvertently encode societal or institutional biases present in the data they learn from. Consequently, the tests generated by these systems are at a risk of bias, potentially disproportionately harming minority or underprivileged students. Additionally, language assessment platforms offer a relevant example of AI-driven language technologies, in which the findings are favourably distributed on a human-like scale and considered algorithmic bias (Blodgett et al., 2020), noting that such systems may disadvantage students who use non-standard dialects or speak with an accent.

In response to these issues, Suresh and Guttag (2019) introduced a framework that helps understand, detect, and mitigate the unintended consequences of machine learning in general but with an emphasis on assessing fairness and accountability while using AI for assessment. This approach underscores the importance of varied training datasets and frequent bias checks to ensure fair evaluation by AI systems. Research has shown that using methods in bias mitigation (e.g. data balancing and algorithmic fairness audits) can improve AI systems with higher inclusivity and lower risks of bias in their outputs (Blodgett et al. 2020).

Adaptive Testing

Following improvements in AES systems, adaptive testing has quickly become an innovative AI application for educational assessments. AI-powered platforms such as ALEKS and Khan Academy have already changed the assessment scenario by providing more personal experiences tailored to specific learners. These systems use real-time data to modify the complexity level of questions based on student performance, resulting in tailored assessment experiences (Eggen & Verschoor, 2016). Adaptive testing has been found to increase student engagement and reduce the cognitive burden, and to be especially beneficial for students who tend not to do well on standardised tests which is an incredibly important advantage (Gierl et al., 2017).

However, although adaptive testing systems are associated with the strengths mentioned above, they have also been criticised for promoting simple processing demands, such as knowledge recall, over higher-level cognitive skills, such as critical thinking and problem solving (Holmes et al., 2019). Opponents maintain that, while these systems are particularly good at determining whether students have memorised facts, they are less effective at engaging students in the kind of complex work essential for academic growth. In addition, there are concerns that items with less technical difficulty might be overrepresented for students with lower performance, possibly reinforcing low expectations and limiting their chances for growth (Gierl et al., 2017).

To address these problems, scholars have suggested the use of adaptive testing systems that include higher-order thinking skills. In addition to forcing pupils to remember things, they were also asked to think about how they could use their knowledge in different situations. Furthermore, by integrating human supervision with an adaptive testing process, teachers and students have a better chance of being engaged and challenged at different levels of progression for every question or help needed (VanLehn, 2011).

Gamified Assessments

Related to the personalised learning aspect of adaptive testing, gamified assessments or science-of-gamified assessments bring in competition, rewards, and leaderboards to enhance student motivation and engagement. Tools like Kahoot! and Quizizz, are widely used in academic settings, especially for younger learners (Dichev & Dicheva, 2017). Research has shown that gamified assessments can increase short-term recall and basic understanding as participants are motivated to score highly or win (Hamari et al., 2016).

Thus, questions have been raised regarding gamified assessments and their ability to promote deep learning. (Plass et al. 2015) The focus on speed and competition can incentivise rote learning at the expense of analytic thought and higher-order cognitive processes, as argued by Turkel and Shorey (2015). Additionally, the competitive nature of gamified assessments may shift students’ focus from learning content to playing to winning, thus compromising their ability to retain and comprehend concepts in real-world applications (Hamari et al., 2016).

Academics seek more gamified learning environments that enhance their critical thinking, analytical, and synthesising abilities primarily because these environments can optimise the advantages of gamified assessments (Hamari et al. 2016). Bonafini (2016) proposed a gamified formative assessment-based learning environment with adaptive tasks that aligns with learners' evolving cognitive abilities, facilitated by artificial intelligence, to foster deep-rooted learning and avoid superficial engagement.

AI in Language Proficiency Evaluation

Similar to automated essay scoring and adaptive testing, AI-enhanced language assessment tools represent a notable shift in educational evaluation. Platforms such as the Duolingo English Test (DET) and Speechace are used for automated and scalable evaluations of speaking, listening, and reading capabilities. Instantaneous feedback from these mechanisms also covers different core language features such as fluency, pronunciation, and grammar (Yang et al., 2021). AI scalability in language assessment has been particularly helpful for remote or underserved populations where traditional language testing may be out of reach.

However, the criticism of prejudice cannot be completely ruled out in AI-driven language testing. Most of them are trained on standardised English varieties (American or British) and might therefore suffer from biased scores for individuals that deviate from white ``Standard English" (Blodgett et al., 2020). This creates disparities in the evaluation of language proficiency by retaining students with diverse linguistic backgrounds.

To meet these challenges, experts have emphasised the need for more inclusive AI models trained on datasets that fully capture not only all languages in the world but also reflect all possible accents across regions (Yang et al., 2021). For all children to be accurately and fairly tested, Howie Holz (2021) stated that AI systems should generate a neutralised system of cultural and linguistic inclusiveness.

Ethical Issues of AI-Powered Proctoring

The use of AI-driven invigilation systems, such as ProctorU and Examity, has grown significantly, particularly during the COVID-19 pandemic. Such systems utilise methods such as facial recognition, keystroke analysis, and behaviour tracking to supervise candidates and prevent academic dishonesty (Li et al. 2020). From an ethical standpoint, these systems have received significant criticism. Debates have arisen regarding their ethical implications, particularly concerning the issues of privacy and surveillance. Despite their potential to effectively secure examinations in some respects, these ethical concerns cast doubt on their overall acceptance.

According to reports, facial recognition algorithms employed by AI invigilation systems frequently misidentify faces that are brown or other forms of traditional facial characteristics that do not match Western norms, such as white (Howard & Borenstein, 2018). In addition, these types of systems have been found to increase certain types of test anxiety that can impair students' performance, which means that monitoring will always be in place (Khosravi et al. 2021). In another recent study, Alison (2021) advocated the definition of ethical standards to ensure that AI proctoring tools are not used in ways that violate students’ privacy and discriminate against vulnerable groups.

Emerging research on AI in educational assessments has illustrated the transformative capacity and limitations of these technologies. Even though AI-powered systems, such as AES adaptive testing and, language assessments, have made impressive strides towards faster and more scalable assessment delivery with improved engagement, these systems still face considerable challenges in terms of bias, fairness, and evaluation of higher-order cognitive abilities. In anticipation of AI technology, educators and policymakers are challenged to build hybrid models that blend the informational power of AI with human oversight to create an assessment that works well for everyone.

METHODOLOGY

This study used a qualitative methodology to investigate the effects of Artificial Intelligence (AI) on educational assessments. The research design consisted of three main components: strategies to locate the literature, data extraction, and methodological orientation to interpret results.

Selection of Literature

This study was based on a systematic review of high-quality, current, and applicable academic reference materials. We searched for articles in major scholarly databases including Google Scholar, PubMed, IEEE Xplore, Scopus, and Web of Science. The search terms included “AI in education,” “AI-driven assessments,” “bias in AI assessments,” “adaptive testing,” “automated essay scoring,” “gamified assessments,” and the much dreaded “AI proctoring”. For the most part, attention was paid to studies published between 2015 and 2023 to capture the current technological trends of AI-based educational assessment systems. Key reviews (up to 10% of all bibliographic references) related to the historical perspective were added only when necessary for the contextual knowledge of AI in education.

Owing to the nature of this task and its intended use, studies examining bias, fairness, and efficiency/scalability were included as long as a concrete application to AI-based educational assessment was mentioned. The preference was around empirical research, case studies, and implementation of large-scale AI tools. Articles pertaining to ethical concerns in AI education were selected through a process of high-relevance classification. The selection process eliminated overly specific studies, such as those focusing on technological advancements like adaptive learning. It also excluded works that fell outside the scope of AI in education as well as publications from sources that had not undergone peer review.

Full-text articles were eligible if they passed the initial screening of the titles and abstracts for review relevance. We also reviewed the references listed in key studies to identify other relevant studies. Consequently, citations for which only DOI could be identified were systematically omitted during the initial selection process by a single author. Subsequently, a conference involving a second or third reviewer was convened to determine whether these references should be reconsidered or incorporated at a later stage.

DATA ANALYSIS

Thematic analysis, according to Braun and Clarke (2006), was used to analyse the selected literature. This well-established qualitative method can reveal repeated themes in complex data sets. The global analysis began with a complete reading of all the articles for an overall understanding of the methodology and results.

The first phase of the coding process was to code text segments to recurring topics, such as bias, privacy, scalability, and efficiency in AI-driven assessment, which made it possible to identify common themes among articles and trends regarding obstacles related to bias in algorithms, challenges associated with assessing higher-order cognitive functions, and issues regarding the ethical aspects of deploying AI in education.

After the codes were coded, the related codes were respun into general themes, which then underwent further refinement to their completed state as original data. The research questions helped align the themes and address the common questions. For example, under the bias theme, the subthemes included AI training data-biased algorithms and linguistic bias in language assessments. In this study, an inductive thematic analysis approach was selected to offer a rich and detailed view of the major themes surrounding AI in educational assessments and a navigation map among anthropogenic and unintended themes.

Frameworks and Methodologies Used

Several established frameworks from AI ethics, educational theory, and assessment technology have been developed to help make sense of the findings and organise the analysis. These theoretical frameworks provide a strong foundation for examining the role of AI in educational assessment.

AI Fairness and Accountability: This framework, incorporating the ideas of scholars such as Binns (2018) and Suresh and Guttag (2019), was used to explore AI from an ethical standpoint, with an emphasis on algorithmic transparency, bias, and fairness in assessment.

Bloom's Taxonomy of Educational Objectives: This system was used to assess the extent to which AI systems can measure various cognitive learning categories, ranging from fundamental memory recall knowledge to higher cognitive skills such as analysis and evaluation.

Industrial production processes: Concerns about privacy, data protection, and surveillance were analysed through the lens of the Ethics Guidelines for Trustworthy AI (2019) by the European Commission. These guidelines emphasise transparency, fairness, and accountability in AI systems used in the public sector, such as in education.

Human-in-the-loop: The HITL framework explores the importance of human oversight in AI-based assessment systems. This framework also opens opportunities for hybrid models, in which human evaluators work together with AI systems to provide more exhaustive and accurate evaluations (Zawacki-Richter et al., 2019).

Study Limitations

Although the methodology provides a strong foundation for examining AI's implications for educational evaluation, this study has several limitations that can impact its findings. The use of secondary data from other academic research, and thus a limited capacity to generate new empirical evidence, is the predominant constraint. Consequently, the findings were limited to what could be ascertained from the quality and types of studies included. This may have limited the research to capturing unique industry reports and grey literature that would not undergo traditional peer review.

In addition, the thematic analysis process could also be limited by subjectivity, as the identification of themes and their interpretation may hinge on the researcher’s perspective. While methods were employed to prevent limitations such as validity and reliability, including peer evaluation and iterative analysis, it might still be possible that there are biases within the literature.

Another important limitation of our study is the site-specific (and spatially limited) nature of the studies we considered. However, this previous work is largely limited to Western education systems and may not generalise well to other educational contexts (such as developing regions, where access to AI-driven tools may differ).

Despite these limitations, this study provides important insights into the current situation of AI in educational assessments and sets a stage for future empirical work that can be performed to validate them.

Validity and Reliability

This was done to increase the reliability and validity of the study in different ways. To ensure that our findings were cross-validated, we triangulated various data sources such as journal publications, conference papers, and case studies. We presented our first round of results to experts in AI and education and incorporated their feedback into the analysis.

DISCUSSION AND ANALYSIS

Artificial Intelligence (AI) in educational assessment presents numerous opportunities and challenges when integrated into the field of education. This section includes AI foundations in automated grading, AI for adaptive examinations, game-based assessments using AI, use of Al to examine and evaluate students' language competency, scoring writing prompts/assessment responses, and securing examinations with Al proctoring. We looked at each, evaluating how they could and are likely to work, the challenges, and what they might mean for the future of education.

Grading & Automation: Efficiency vs Depth

However, the use of AI-based Automated Essay Scoring (AES) systems such as E-rater and Grammarly provides immense benefits in terms of speed and consistency. These tools evaluate a large number of essays quickly, providing instant feedback on surface-level features, including sentence structure, grammar, and word usage (Wang et al. 2020). This mobile pathway of scalability renders them especially valuable in large-scale standardised testing situations when human evaluators may have trouble maintaining consistency across a vast array of student responses (Shermis & Burstein, 2013).

Although AES systems have been successful in improving grading efficiency, they are still limited in assessing higher cognitive abilities (e.g. creativity, analysis, and critical thinking) (Perelman, 2014). Such systems frequently direct students to adopt algorithm-friendly standard writing styles that limit their intellectual creativity and critical thinking skills (Wang et al., 2020). This dependence on automated systems has sometimes resulted in biased evaluations, particularly when the response is not a common practice or out-of-the-box solution.

After the event, a case study on ETS and E-rater: The Educational Testing Service (ETS), which handles both the TOEFL and GRE, uses E-rater to help human graders score writing samples. Automated essay scoring (AES) systems, such as E-rater, are accurate in assessing low-level writing skills (e.g. grammar) but perform poorly in grading papers on creativity and composition, meaning that human graders will view these higher-order skills in addition to filling out rubrics (Shermis & Burstein, 2013).

For example, Human Grading is Better at Creativity and Argumentation than Automated AES; How Marked Differences Between Traditional and AI in the Scoring of TOEFL iBT; Traditional graders thought [in comparison to AES] that they were better scoring artificial intelligence essays so far. This underscores the necessity of combining computerised techniques with human input to deliver comprehensive and reliable assessments. Furthermore, the DA system has faced criticism for potential grading prejudice when evaluating sophisticated essays with unconventional structures, labelling them as 'essays' received excessive influence from AI' during an analysis of China's national university entrance exam.

Example of Success: EdX and AI Feedback — A noted example of the successful use of AI to supply a quick response in Massive Open Online Courses (MOOCs), such as those presented by EdX. While still requiring human intervention for deeper analysis, automated feedback can help students iterate on their writing.

Problems with Florida's Automated Grading System: Florida planned to fully automate the grading of high-stakes assessments; however, the automated system could not properly evaluate creative or new types of responses and produced unreliable results, thus requiring human raters.

Implications:

Efficiency: AES systems significantly reduce the time taken for grading in high-stakes testing, which enables teachers to spend more time on instructional tasks (Shermis & Burstein, 2013).

Shallow Learning: Overreliance on automated grading can encourage surface-level characteristics rather than higher-order cognitive abilities (Perelman, 2014).

Justice: The implementation of AES systems can marginalise colourful writing styles, yielding concerns over whether some students are less likely to benefit from such tools than others (Wang et al., 2020).

Adaptive Testing: Optimised Study or Rooms with Walls

Perception: Platforms such as ALEKS and AI-driven test providers such as Khan Academy have reinvented how assessments are conducted, making assessment methods individualised. This ability dynamically adjusts question challenges relative to student answers and represents a personalised learning path consistent with individual capabilities (Eggen &Verschoor, 2016). Adaptive testing improves student engagement, allows teachers to provide timely support (by reducing cognitive load), and offers immediate diagnostic feedback about the status of student learning as response patterns become available (Gierl, 2017).

For example, ALEKS: Implementation of ALEKS in math instruction has led to improved results, as it delivers tailored learning tracks, while another trial using ALEKS in California high schools reported higher levels of engagement and achievement, especially for students who had previously been low-attainers (VanLehn 2011). Despite this, concerns have been raised that ALEKS degrees tend to emphasise recall of knowledge rather than building critical thinking skills.

While adaptive systems have shown results by individualising learning, some would argue that they run the risk of becoming too closed and narrow in their assessments. As Holmes et al. (2019) argued, with a high emphasis on knowledge recall and short-term proficiency, adaptive testing may fail to challenge students with tasks that engage higher-order cognitive skills

, such as problem-solving and critical analysis. This constraint can inhibit deep learning, particularly in fields that require knowledge to be applied to novel or complex situations (Van der Linden & Glas, 2015).

Knewton Adaptive Learning: Knewton, an AI-powered adaptive learning platform, was launched with much buzz but did not do well in the education industry because it had a limited ability to assess higher-order thinking skills. Academics have questioned this emphasis on instant, machine-generated advice, where the quality of assessment was largely based on knowledge that could be factored into when they were arguing for a more thoughtful and robust way of addressing serious matters.

Implications:

Adaptive testing brings individualisation: an experience tailored to the examined, which could make tests more engaging and less frustrating (Eggen & Verschoor, 2016).

Restricted Focus: The accent on the immediate recall of knowledge may limit complex cognitive engagement opportunities (Holmes et al., 2019).

Fairness: Reduced cognitive-level opportunities for students with lower performance owing to fewer challenging multiple-choice items (Gierl et al., 2017).

Gamified Assessments: Engagement or Learning Outcomes

In the educational environment, AI-enabled assessment gamification has been on the rise as it enhances student motivation and engagement. Tools like Kahoot! Games such as Kahootistream), and Quizizz embed game features, including competition, leaderboards, and rewards, which have been shown to enhance participation, particularly in younger children (Dichev & Dicheva, 2017). These simply are effective platforms but not enough for long-term effects and in-depth understanding because of competition to obtain success or medals that drive learners towards these subjects (Hamari et al., 2016).

Case Study – Kahoot! in K-12 Classrooms: Kahoot! has become a global classroom staple, particularly in K-12, and has been proven to boost student engagement. In Norway, Kahoot! The focus on competition possibly resulted in shallow, superficial learning oriented towards memorisation, as opposed to profound cognitive engagement (Wang & Lieberoth, 2016).

Although there is engagement value, some questions remain regarding how educational gamified assessment can be. However, the vast majority of these tools support speed and accuracy over productive thinking (Plass et al. 2015). The competitive element of gamified assessment may persuade the learners to focus only on getting 'the win' rather than truly learning the content, leading them to struggle if they must apply their knowledge in real-life scenarios (Hamari et al., 2016).

South Korea: A case study comparing South Korean secondary schools showed that gamifying assessments increased motivation, but did not increase students' academic success. Thus, the learning process was diverted from competition and rewards, with students showing low retention of information (Kim et al., 2020).

Implications:

Rationalization: As part of gamified assessments, participant engagement is vastly increased in low-stakes environment (Dichev & Dicheva, 2017) Motivation: In low-stakes environments greatly improve participant experience by gamifying the student evaluations.

Shallow Learning: Emphasis on competition and quick answers can reduce opportunities for deep cognitive engagement and problem-solving (Plass et al. 2015).

Design trajectory: Gamified assessments should progress to help competition support higher cognitive skills (Hamari et al., 2016).

Bias in AI-scripts for Language Skills Evaluation

AI-driven LEPLOs such as the Duolingo English Test (DET) and Speechace have changed the way language proficiency is evaluated, particularly speaking and listening. They provide instant feedback on pronunciation, fluency, and grammatical accuracy, which is helpful for scale-up assessments in remote and underdeveloped areas (Yang et al. 2021).

Case Study: Duolingo English Test: Presented by Chris Bourg, the DET was praised for being widely available and very fast in returning scores, which is ideal for students in remote areas. However, it has been shown to unfairly penalise those with a non-native accent or regional dialect that can lead to racial biases. In 2020, research demonstrated that those whose indigenous language was an African English dialect earned lower scores on the DET than their native English-speaking counterparts, despite performing at the same level of overall proficiency (Blodgett et al., 2020).

One of these issues is bias in the AI language assessment. Most AI systems are trained on similarly standardised English accents (mostly American or British) and thus will occasionally give biased scores if evaluated by speakers who deviate from these standards. Inconsistent tests, designed to maintain a fixed standard of difficulty across all tests, could result in linguistic minorities (language learners who do not have English as their L1) being discriminated against (Suresh and Guttag 2019).

Implications:

Scalability and Timeliness: An AI-driven evaluation of language effectiveness measures can be valuable for the large-scale evaluation of language quality (Yang et al. 2021).

A focus on aligning a language with standardised accents can result in skewed evaluations because of its prejudiced nature, indicating that a more holistic AI framework is required (Blodgett et al., 2020).

Cultural Representation: To provide fair assessments that account for language diversity worldwide, a variety of datasets have been used (Suresh & Guttag 2019).

Ethical Implications of AI-Based Proctoring

Some AI-powered invigilation platforms, such as ProctorU and Examity, use a combination of technologies, such as face recognition, keyboard input analysis, and behaviour monitoring, to supervise candidates during exams. These systems enhance examination security by deterring cheating and facilitating academic integrity during online testing (Li et al. 2020).

Exemplification: ProctorU: During the COVID-19 pandemic, many institutions used ProctorU for remotely proctored examinations, which was effective in improving test security. Learners were significantly concerned about privacy invasion and constant surveillance. Research suggests that facial recognition technology consistently fails to correctly identify students with deep skin tones, misidentifies hundreds or thousands of tones, and has a higher chance of false allegations (Howard & Borenstein 2018).

The application of AI to examination invigilations raises serious ethical issues related to privacy and surveillance. Students from disadvantaged backgrounds are likely to experience unfair impacts on facial recognition inaccuracies. Moreover, the continuous monitoring implemented by AI invigilation platforms has been linked to higher test anxiety, which influences student performance (Khosravi et al. 2021).

Consequences:

Academic Integrity: Artificial intelligence (AI)-enabled proctoring systems can help prevent cheating during exams (Khosravi et al., 2021).

Ethical Implications: Preventing privacy issues and bias are ethical challenges that must be met; therefore, AI invigilation systems do not discriminate against those from more deprived backgrounds (Li et al., 2020)

Student Welfare Invasive AI surveillance can further increase test anxiety, thus negatively affecting student performance (Howard and Borenstein 2018).

This highlights that the major advantages and disadvantages of AI-based assessments of AI systems are the ability to boost efficiency, personalisation, and engagement by detecting bias, assessing higher-order thinking skills, and adhering to ethical standards remains a challenge. To maximise the power of AI in education, a balance must be maintained between automation by AI and human intervention for monitoring assessments, which makes these fair, transparent, and meaningful.

NEW INSIGHTS FROM THE STUDY

Bias Mitigation:

This study emphasises the necessity of using diverse real-world datasets while training AI models to reduce the algorithmic bias. AI models trained on concrete and specific referenced data will have the same lens with which to view educational assessments; if these data are limited or simply reflect the students of Edison, this AI might only re-entrench educational disparities that similarly disadvantage minority cultures to take the example in another direction. The study suggests the adoption of bias audits, a type of analysis that goes through AI outputs systematically to see if there are any patterns associated with biases. This also underscores the need for continued human oversight to identify and correct biases at each stage of the evaluation process. Developing less-biased and more equitable AI-driven assessments is achievable by keeping AI tools and their underlying data inclusive, unbiased, and equitable.

Hybrid Models:

We argue that this research is of great importance because it advocates a hybrid assessment framework. These frameworks offer the efficiency and scale of AI but leverage the finesse and subjective lens that human evaluators each bring to data evaluation. This holistic view is particularly important in assessing higher-order cognitive skills, such as critical thinking, creativity, and problem solving, all areas in which an AI system fails because its success is predicated on recurring algorithms and patterns. It argues that human oversight is necessary to fairly and completely assess these nuanced, abstract skills and the kinds of tasks that AI might be better suited to detecting grammar mistakes or recognising factual recall.

Ethical AI Practices:

This study highlights the growing need for standards or norms for AI use in educational assessments. Although many people are excited about the potential of AI in educational settings, concerns over data protection, privacy, and monitoring have increased. The report noted that these AI systems often collect personal information; therefore, educators and policymakers must implement strong data protection measures. This study also suggests that strict ethical guidelines, such as data collection, storage, and sharing protocols on the part of AI systems, should be laid down to ensure that learners' protective rights do not hamper the integrity of this process.

CONCLUSION

The integration of Artificial Intelligence (AI) into educational assessments has created a changing landscape, with huge potential and unique challenges. In this way, they clarified that AI could improve the efficiency, scalability, and personalisation of educational assessments. However, the discussion has raised serious issues regarding what AI can and cannot score in terms of higher-order cognitive abilities such as creativity, critical thinking, and problem-solving. Critically, the complications of algorithmic bias, privacy, and data security considerations, and in particular, the ethical demands for human oversight, indicate complex grounds that AI will need to negotiate in education.

Recommendations for Educators

Educators must have a definite understanding of AI, how far one can reach with technology, and what ethical considerations come into play. Develop professional development packages that specifically provide training on how to use AI tools correctly, leading to the publication of AI literature. Teachers must interpret the outputs of AI models and measure when and how to intervene. In other words, AI should always be in support of human skills and in places that need this kind of nuanced judgment, which might include how creative someone is or their ability to respond to complex problems.

Educators must also consider blending in hybrid models, where AI is used for routine or high-prevalence assessments, leaving the scope open for subjective judgments in which human evaluation is crucial. This two-sided sword approach ensures that AI drives efficiency but also helps in keeping the educationist approach holistic and fair, driven by a more comprehensive range of student abilities.

To effectively integrate AI and human assessment for higher-order skills, educators should implement the following steps:

Step 1: Understand and Delineate the Strengths of AI and Humans. Educators must first clearly recognize the distinct capabilities of both AI and human evaluators in assessment. AI excels in efficiency, consistency, and processing surface-level features like grammar and syntax, providing instant feedback for routine or high-prevalence assessments. For example, Automated Essay Scoring (AES) systems are adept at quickly evaluating large volumes of writing for structural elements. However, AI frequently falls short in measuring advanced thinking abilities such as creativity, critical thinking, analytical reasoning, and complex problem-solving. These nuanced cognitive processes, along with the need for holistic evaluation and bias mitigation, necessitate human judgment and oversight. Human graders are crucial for assessing the depth of writing and argumentation, particularly when evaluating unconventional or unique responses.

Step 2: Design Hybrid Workflows Based on Assessment Objectives. After understanding the respective strengths, educators should design assessment workflows that strategically combine AI and human input. For instance, in written assessments, AES systems can be used for initial screening or evaluating grammatical accuracy, while human graders focus on higher-order skills like creativity and argumentation that AI struggles with. Similarly, adaptive testing platforms can tailor question complexity for personalized learning paths, but human supervision remains vital to ensure students are challenged with complex problem-solving tasks and to prevent over-representation of easier items for lower-performing students. This blended approach leverages AI for efficiency while reserving human expertise for subjective, critical judgments.

Step 3: Implement Ethical Guidelines and Continuous Oversight. The implementation of hybrid models must be accompanied by robust ethical guidelines and continuous human oversight to address inherent challenges. Educators are on the front lines of addressing concerns such as algorithmic bias, especially against non-native accents or diverse linguistic and cultural backgrounds, and privacy issues arising from AI proctoring systems. Continuous human oversight is essential to identify and correct biases at each stage of the evaluation process, and educators must be trained to interpret AI outputs and intervene when necessary to ensure fairness and equity. Strong data protection measures and clear protocols for data collection, storage, and sharing are also paramount to safeguard student privacy.

Step 4: Continuous Evaluation and Adaptation. Hybrid assessment systems are not static; they require ongoing evaluation and adaptation to remain effective and equitable. Educators should continuously monitor the effectiveness of these blended models in accurately measuring learning outcomes and fostering deep learning. This includes assessing the impact of AI-based assessments on student well-being, motivation, engagement, and test anxiety. Furthermore, developing professional development packages that specifically train educators on the correct and ethical use of AI tools is crucial for successful implementation and ongoing refinement of hybrid assessment practices. This iterative process ensures that hybrid models evolve to meet pedagogical needs and mitigate potential negative impacts.

Policy Recommendations

Instead of gatekeeping intelligent systems in education, policymakers should best provide their time by constructing strong ethical rules to guide AI applications in educational contexts. These principles, she said, would "offer visibility on how AI was developed, create an accountability layer, and protect student privacy. Ongoing evaluations and audits are a must in the case of AI systems to prevent them from embedding biases that disadvantage only some students, especially those who come from non-traditional or marginalised backgrounds.

In the early stages, policies must go beyond ensuring that AI models are trained on a variety of datasets, especially for language assessments, to be culturally and linguistically inclusive. Makers also need to regulate the ethical aspects of using AI during high-stakes assessments, such as proctoring systems that have come under fire because of their impact on student well-being due to their invasiveness. All policy frameworks must reflect that AI should not worsen inequalities, but should help overcome infirmities within education.

Future Research Directions

The study also recommends that future research fill several key gaps in the AI and education literature. The critical insights gained include the pressing need to develop and evaluate empirical research on the practical efficacy of combined artificial intelligence and human evaluation systems. Such studies should examine how these models can best leverage AI efficiency and human evaluative subtlety, particularly in scenarios involving complex cognitive abilities.

Second, further research is required to prevent algorithmic bias. Ensuring that the training datasets are more inclusive and enshrining on-going bias audits in AI systems will open a path towards an unbiased evaluation. Finally, further investigation should explore how AI-based assessment methods contribute to student well-being in terms of anxiety, motivation, and interventions that may attenuate these negative impacts.

The next phase involves conducting comparative analyses to evaluate the efficacy of AI-based assessment techniques in relation to conventional approaches. These investigations should encompass AI's performance across diverse educational settings, as well as its capacity to assess advanced cognitive abilities and ensure equitable evaluation. Finally, the ethical frameworks that guide the future of AI in education should be investigated to maintain transparency, accountability, and student centricity in AI-driven assessments.

Although AI paints a bright picture of transformation in the educational assessment realm, the transition has hurdles and must be performed carefully. Ensuring that AI-driven assessments are both innovative and equitable: Teachers should be given literacy in AI, hybrid models that incorporate the capabilities of AIs with human judgment to help guide better understanding and assist teachers, and strong ethical frameworks. Future research and policy should address the educational, ethical, and technical barriers to AI growth. Working together with educators, policymakers, and researchers, AI can revolutionise the educational landscape by enabling deep learning, fairness, and inclusivity in the educational environment, so that all students are equipped for success in our increasingly digital world.

REFERENCES

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: a critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency* (pp. 149-159). ACM. https://doi.org/10.1145/3287560

Dichev, C., Dicheva, D. (2017). Gamifying education: What is known, what is believed and what remains uncertain: A critical review. *International Journal of Educational Technology in Higher Education*, 14(1), 1–36. https://doi.org/10.1186/s41239-017-0042-5

Eggen, T. J., & Verschoor, A. J. (2016). *Optimal testing with randomized items* (2nd ed.). Routledge.

European Commission. (2019). Ethics guidelines for trustworthy AI. European Commission. https://ec.europa.eu/digital-strategy/news-redirect/66009

Gierl, M. J., Bulut, O., & Bayrak, O. (2017). Developing a technology for designing and scoring adaptive assessments. *Educational Measurement: Issues and Practice*, 36(3), 29–40. https://doi.org/10.1111/emip.12154

Hamari, J., Koivisto, J., & Sarsa, H. (2016). Does gamification work?—A literature review of empirical studies on gamification. *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*, 3025–3034. https://doi.org/10.1109/HICSS.2014.377

Holmes, N. G., Wieman, C. E., & Bonn, D. A. (2019). Teaching critical thinking. *Science*, 363(6432), 658–662. https://doi.org/10.1126/science.aav9490

Howard, A., & Borenstein, J. (2018). The ugly truth about automated facial analysis: Gender and racial bias in facial recognition technologies. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 1–9. https://doi.org/10.1145/3278721.3278776

Khosravi, H., Cooper, M., & Kitto, K. (2021). AI in education: The promises, opportunities, and threats of using artificial intelligence to assess and support students. *Computers & Education*, 163, 104099. https://doi.org/10.1016/j.compedu.2020.104099

Kim, Y. G., Kim, J., & Kim, J. H. (2020). AI-driven learning environments: Emerging pedagogical issues and technological solutions. *Educational Technology Research and Development*, 68(5), 225-246. https://doi.org/10.1007/s11423-020-09755-5

Li, J., Ma, W., & Lu, Q. (2020). AI-based proctoring systems and academic integrity: Ethical challenges in the post-pandemic education landscape. *AI & Society*, 35(4), 761–772. https://doi.org/10.1007/s00146-020-01023-8

Lau, A. (2017). Human vs. machine: Reassessing the role of automated essay scoring in education. *Journal of Educational Technology & Society*, 20(2), 153–162.

Perelman, L. (2014). Critique of automated scoring systems: A research paper. *Journal of Writing Assessment*, 7(1), 1–16. https://www.journalofwritingassessment.org/article.php?article=77

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258–283. https://doi.org/10.1080/00461520.2015.1122533

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Smith, P., & Johnson, R. (2021). AI in education: Opportunities and challenges for learners and educators. *Educational Technology & Society*, 24(2), 17–27. https://doi.org/10.1037/edu0000670

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *Communications of the ACM*, 63(5), 58–66. https://doi.org/10.1145/3287560

Taghipour, K., & Ng, H. T. (2021). A deep learning approach to automated essay scoring. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 47–57. https://doi.org/10.18653/v1/N16-1006

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Van der Linden, W. J., & Glas, C. A. W. (2015). *Elements of adaptive testing*. Springer Science & Business Media.

Wang, J., Wang, Y., & Lieberoth, A. (2016). Effects of Kahoot! on academic performance and student engagement: A meta-analysis. *Journal of Educational Psychology*, 108(2), 291–308. https://doi.org/10.1037/edu0000128

Wang, Y., Ng, H., & Wang, X. (2020). Automated essay scoring: A review of current applications and challenges. *Educational Measurement: Issues and Practice*, 39(2), 5–15. https://doi.org/10.1111/emip.12360

Yang, Q., Liu, J., & Wei, X. (2021). The impact of AI on language testing: Exploring the effectiveness of the Duolingo English Test. *Language Testing in Asia*, 11(3), 1–15. https://doi.org/10.1186/s40468-021-00124-7

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), 39. https://doi.org/10.1186/s41239-019-0176-8