

ARTIFICIAL INTELLIGENCE-ASSISTED ASSESSMENT AND DIGITAL LITERACY IN HIGHER EDUCATION: A THEMATIC ANALYSIS OF VALIDITY, RELIABILITY, AND FAIRNESS

¹Isaac Bamikole Ogunsakin

Dept. of Educational Foundations, Faculty of Education,
Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria.
sakinbamikole2019@gmail.com
<https://orcid.org/0009-0002-9385-1738>

²Oluwatosin Motunrayo Kachi

(corresponding author)

Department of Educational Management and Planning,
Faculty of Education, Federal University Lokoja, Kogi State, Nigeria
oluwatosin.kachi@fulokoja.edu.ng
<https://orcid.org/0009-0000-8949-4913>

Abstract

The fast integration of generative artificial intelligence (GenAI) and huge language models into higher education is changing assessment processes while raising questions about validity, reliability, fairness, and academic integrity. This thematic literature review looked at recent research on AI-assisted assessment and AI literacy in higher education, with a focus on the psychological and ethical consequences of AI use. A structured literature search was undertaken via Scopus, Web of Science, Google Scholar, and Taylor & Francis Online using predetermined search terms and eligibility criteria to discover relevant peer-reviewed publications published between 2023 and 2026. Thirty papers met the inclusion requirements and were combined into a structured theme framework. The synthesis revealed four interconnected dimensions: assessment design and construct representation; algorithmic consistency and score invariance; sociotechnical equity and prejudice; and governance, academic integrity, and human-in-the-loop safeguards. The results show that GenAI can increase scoring, feedback, and assessment efficiency, but its validity and reliability are dependent on assessment design, model attributes, prompting, context, and human supervision. Variations in AI models and prompts, disparities in AI literacy and access, and potential algorithmic and sociolinguistic biases could undermine fairness and comparability in AI-assisted evaluation. The review emphasizes the importance of

transparent assessment design, systematic validation, AI literacy development, fairness safeguards, and significant human monitoring. Future research should focus on interdisciplinary and longitudinal examinations of AI scoring consistency, construct representation, fairness, and the validity of AI-assisted assessment in a variety of educational situations.

Keywords: Artificial intelligence; AI-assisted assessment; digital literacy; validity and reliability; academic integrity

Introduction

Artificial intelligence (AI), particularly generative AI (GenAI) and large language models (LLMs), is revolutionizing higher education assessment by automating scoring, feedback, question production, adaptive assessment, and learning analytics. Despite these advantages, one major worry is whether AI-assisted assessment produces genuine, trustworthy, and fair proof of student learning. Two concepts are crucial to this problem. AI-as-Evaluator, also known as Automated Measurement, is the process of using AI to evaluate student work. Validity, dependability, scoring consistency, agreement with human judgments, and fairness are among the most important problems. Evidence indicates that LLM-based scoring can differ depending on prompting tactics, assessment situations, and writing features (Kim, 2025; Kostic et al., 2024; Pack et al., 2024; Yavuz et al., 2025)..

AI-as-Student-Collaborator, also known as Construct Interaction, includes students utilizing GenAI for evaluation purposes. The question is whether performance still reflects the desired concept when replies include student knowledge, AI literacy, prompting abilities, and access to AI. Where these elements are not intended outcomes, they can add construct-irrelevant variation (Kaldaras et al., 2024; Xia et al., 2024). Differences in students' AI knowledge, experience, access, and evaluation abilities influence assessment interpretation (Hornberger et al., 2023; Beckman et al., 2025; Kong et al., 2025).

AI literacy and justice are consequently critical in both models. LLM usage necessitates critical thinking, fact-checking, and human oversight (Kasneci et al., 2023), whereas AI literacy varies across students and involves knowledge, application, assessment, creation, and ethical use (Hornberger

et al., 2023; Kong et al., 2025; Mansoor et al., 2024; Rožman et al., 2025). Unequal AI access and AI-related abilities may potentially result in disparities. AI-detection algorithms have demonstrated significant inadequacies in differentiating between human and AI-generated text (Weber-Wulff et al., 2023), raising concerns regarding fairness and academic integrity (Bond et al., 2024; Ogunleye et al., 2024; Xia et al., 2024).

This thematic review synthesizes 30 studies using the Automated Measurement and Construct Interaction perspectives, focusing on four dimensions: assessment design and construct representation; algorithmic consistency and scoring invariance; sociotechnical equity and bias; and governance, academic integrity, and human-in-the-loop safeguards. It investigates the consequences for the validity, reliability, fairness, and interpretation of AI-assisted assessments.

Methodology

Research Design

This study adopted a qualitative thematic literature review design. The approach was considered appropriate because the purpose of the study was to identify, compare and interpret recurring patterns in contemporary scholarship concerning AI-assisted assessment and digital/AI literacy in higher education. Unlike a meta-analysis, which seeks to statistically combine quantitative findings, thematic analysis allows researchers to examine conceptual relationships, methodological differences and recurring concerns across heterogeneous studies. This was particularly suitable because the selected literature included empirical studies, psychometric investigations, systematic reviews, scoping reviews and conceptual contributions. The review was guided by a measurement-oriented perspective. Three major assessment qualities formed the analytical foundation: validity, reliability and fairness. The review incorporated digital and AI literacy and academic integrity because they directly affect how AI-assisted assessment is designed, implemented and interpreted. The review focused primarily on literature published between 2023 and 2026. This period was selected because it corresponds with the rapid expansion of generative AI and the widespread emergence of LLM-based educational applications.

Search Strategy

The literature search was conducted across Scopus, Web of Science, Google Scholar, and Taylor & Francis Online. Search terms focused on artificial intelligence, generative AI, higher education, assessment, AI literacy, digital literacy, validity, reliability, fairness, bias, and academic integrity. Key search combinations included “artificial intelligence” AND “assessment” AND “higher education,” “generative AI” AND assessment, “ChatGPT” AND assessment AND university, “AI literacy” AND “university students,” “AI assessment” AND validity, “AI assessment” AND reliability, “AI assessment” AND fairness, and “AI detection” AND academic integrity. Boolean operators and related terms, including automated essay scoring, AI-supported feedback, and learning analytics, were incorporated to broaden the search and capture relevant interdisciplinary literature. Studies were screened against predefined inclusion and exclusion criteria, with eligible peer-reviewed studies published between 2023 and 2026 included in the final synthesis.

Inclusion and Exclusion Criteria

Studies were included if they were published between 2023 and 2026, had verifiable scholarly records, focused on AI or AI-supported assessment in higher education, and addressed at least one review theme, including AI literacy, assessment, validity, reliability, fairness, bias or academic integrity. Empirical studies, systematic and scoping reviews, and relevant theoretical works were considered where DOI or stable publication information was available. Studies were excluded if they focused on non-educational AI applications, did not address higher education, lacked sufficient bibliographic information, duplicated another record or had limited relevance to the review objectives.

Study Identification and Selection

The initial search produced a broad pool of potentially relevant studies. Titles and abstracts were screened to remove clearly unrelated publications, while potentially relevant studies were assessed against the inclusion and exclusion criteria. Full texts or publisher records were examined where relevance was unclear, and bibliographic details were checked to minimise duplication and identification errors. The final review included 30 studies covering AI literacy, automated scoring, AI detection, academic integrity and broader assessment issues across different methodological approaches.

Data Extraction

A structured literature matrix was used to record the author and year, geographical context, research design, participants or data source, AI technology, assessment application, AI literacy, validity, reliability, fairness, academic integrity, major findings and implications. The matrix supported comparison across studies and identification of recurring findings, methodological differences and areas of disagreement. Particular attention was given to distinguishing empirical evidence from conceptual arguments, given the limited long-term evidence on AI-assisted assessment.

Quality and Relevance Assessment

Studies were assessed based on methodological clarity, conceptual relevance and usefulness to the review objectives. For empirical studies, attention was given to research design, sampling, measurement procedures and evidence supporting validity and reliability claims. AI literacy studies were examined for instrument development and validation, while automated scoring studies were assessed for human-AI agreement, score stability, model differences and prompt effects. AI-detection studies were examined for classification accuracy and potential false positives and negatives. Reviews provided broader field perspectives, while empirical and psychometric studies supplied specific evidence on assessment quality.

Thematic Analysis

The selected literature was analysed thematically through familiarisation, initial coding, grouping related codes, theme development, and interpretation of relationships among themes. The coding process focused on assessment design and construct representation, algorithmic consistency and scoring invariance, sociotechnical equity and bias, and governance, academic integrity, and human-in-the-loop safeguards. Related concepts included AI literacy, construct validity, automated scoring, scoring consistency, prompt sensitivity, model variation, linguistic and cultural bias, digital inequality, transparency, and human oversight. These codes were synthesised into four interconnected themes: (1) Assessment Design and Construct Representation, (2) Algorithmic Consistency and Scoring Invariance, (3) Sociotechnical Equity and Bias, and (4) Governance, Academic Integrity, and Human-in-the-Loop Safeguards. The analysis also examined cross-cutting relationships, particularly how AI literacy influences assessment performance and how validity, reliability, and

fairness interact in AI-assisted assessment. This approach enabled the review to move beyond describing individual studies towards developing an integrated framework for understanding the conditions required for valid, reliable, fair, and educationally meaningful AI-assisted assessment in higher education.

Trustworthiness

Several strategies were used to enhance the trustworthiness of the review, including explicit inclusion criteria, structured data extraction, bibliographic verification, and comparison of findings across studies. Contradictory findings were retained rather than eliminated because differences between studies may reveal the conditions under which AI-assisted assessment performs well or poorly. For example, evidence of high agreement between an LLM and human raters was not automatically interpreted as evidence of validity, just as consistent AI-generated scores were not automatically interpreted as evidence of fairness. The analysis therefore maintained a clear conceptual distinction between validity, which refers to whether the assessment supports the intended interpretation; reliability, which concerns whether measurement is sufficiently consistent; and fairness, which addresses whether assessment operates equitably across relevant groups and contexts.

Ethical Considerations

The study relied exclusively on published literature and did not involve direct interaction with human participants. Formal participant consent was therefore not required. Nevertheless, ethical principles were observed through accurate representation of previous research, appropriate citation, careful interpretation of findings and avoidance of unsupported claims. Particular care was taken not to overstate the capabilities or limitations of AI technologies beyond the evidence provided by the reviewed studies.

Methodological Limitations

The review has a few limitations. First, the search concentrated mostly on English-language publications, potentially omitting pertinent studies published in other languages. Second, the rapid evolution of AI technologies means that findings may become less relevant as models and versions evolve. As a result, AI performance should be evaluated in light of the specific model, version, and assessment conditions used. Third, while

the review evaluated a variety of scholarly databases and sources, such as Scopus, Web of Science, ERIC, ScienceDirect, and Google Scholar, variations in database coverage, indexing procedures, and search capability may have influenced the research found. Finally, the inclusion of publisher platforms and other scholarly sources for extra searching and bibliographic verification may have resulted in differences in coverage. These limitations should be noted when interpreting the findings and determining the review's comprehensiveness.

Table 1: Characteristics and Thematic Classification of the Reviewed Studies

No.	Study	Year	Primary focus	Method/evidence	Theme
1	Hornberger et al.	2023	University students' AI literacy	Test development/validation	AI literacy
2	Kasneji et al.	2023	LLMs in education	Critical review	AI literacy/Assessment
3	Weber-Wulff et al.	2023	AI-text detection	Empirical evaluation	Fairness/Integrity
4	Memarian & Doleck	2023	FATE in AI and higher education	Systematic review	Fairness
5	Moorhouse et al.	2023	GenAI and university assessment	Policy/document analysis	Assessment
6	Laupichler et al.	2023	AI literacy measurement	Delphi/validation	AI literacy
7	Knoth et al.	2023	AI literacy/course assessment	Instrument evaluation	AI literacy
8	Pack et al.	2024	Automated essay scoring	Empirical psychometric study	Validity/Reliability
9	Xia et al.	2024	GenAI and assessment	Scoping review	Assessment
10	Bond et al.	2024	AI in higher education	Meta-systematic review	Assessment/AI
11	Ogunleye et al.	2024	Generative AI in education	Systematic review	Assessment
12	Kaldaras et al.	2024	Assessment validity in GenAI era	Measurement analysis	Validity
13	Uto & Aramaki	2024	Automated essay scoring	Many-facet modelling	Reliability
14	Fleckenstein et al.	2024	Detection of AI-generated writing	Experimental study	Fairness

15	Mansoor et al.	2024	University AI literacy	Multinational survey	AI literacy
16	Lee et al.	2024	University students' AI literacy	Empirical study	AI literacy
17	Lintner	2024	AI-literacy scales	Systematic review/COSMIN	AI literacy
18	Chiu et al.	2024	Objective AI-literacy testing	Rasch validation	AI literacy/Validity
19	Perkins et al.	2024	AI Assessment Scale	Framework	Assessment/Integrity
20	Kostic et al.	2024	LLM automated essay evaluation	Case study	Assessment
21	Yavuz et al.	2025	LLM-based EFL essay grading	Human-AI comparison	Validity/Reliability
22	Kim	2025	GPT-4 essay scoring	Empirical study	Validity/Reliability
23	Hornberger et al.	2025	Cross-national AI literacy	Multinational study	AI literacy
24	Kong et al.	2025	AI literacy	Scale/empirical study	AI literacy
25	Rožman et al.	2025	AI literacy	Cross-national study	AI literacy
26	Beckman et al.	2025	GenAI divide	Empirical study	AI literacy/Fairness
27	Karkouliau et al.	2025	ChatGPT and academic integrity	Qualitative/empirical	Integrity
28	Tian & Zhang	2025	AI dependence and critical thinking	Empirical study	AI literacy

29	Ncube et al.	2026	AI and student assessment	Systematic review	Assessment/Integrity
30	Bulut et al.	2024	AI in educational measurement	Measurement/ethical review	Validity/Reliability/Fairness

Thematic Literature Review

Thematic Synthesis

The thematic synthesis was reorganized into four interconnected dimensions that capture the major measurement and governance issues surrounding AI-assisted assessment in higher education: (1) Assessment Design and Construct Representation, (2) Algorithmic Consistency and Scoring Invariance, (3) Sociotechnical Equity and Bias, and (4) Governance, Academic Integrity, and Human-in-the-Loop Safeguards. Together, these dimensions provide a structured measurement framework for understanding when AI-assisted assessment can generate valid, reliable, fair, and educationally meaningful evidence of student learning.

Assessment Design and Construct Representation

AI-assisted assessment necessitates precise alignment of the targeted learning concept, assessment task, scoring criteria, and the function allocated to AI. Generative AI can help with assessment development, feedback, automated scoring, and other assessment operations, but its usage may alter what is actually measured. When students utilize AI to produce, organize, translate, or improve their replies, their performance may be influenced by a mix of discipline knowledge, AI literacy, prompting skills, and access to AI, rather than just the intended learning construct (Kaldaras et al., 2024; Xia et al., 2024). This is a possible source of concept-irrelevant variance when AI-related abilities are not included in the planned assessment construct.

Thus, while designing assessments, AI literacy is crucial. Research shows that university students' knowledge and comprehension of AI vary widely, and these variations have also been documented across educational and national settings (Hornberger et al., 2023; Hornberger et al., 2025a, 2025b; Mansoor et al., 2024; Rožman et al., 2025). Increased AI literacy may put

students in a better position to provide efficient prompts, analyze AI results, and incorporate AI into evaluation assignments. Therefore, whether AI literacy is a component of the targeted construct or whether its effect should be regulated should be made clear in assessment designs that include the use of AI.

Assessment approaches that emphasize how students exhibit learning rather than just final deliverables are becoming more and more popular in the literature. In addition to offering more proof of individual learning, oral exams, reflective narratives, staged submissions, process evidence, and other authentic assessment methods can lessen an over-reliance on AI-generated results (Ateeq et al., 2024; Kaldaras et al., 2024; Moorhouse et al., 2023; Perkins et al., 2024). Because the success of AI-assisted assessment ultimately depends on whether the assessment captures the desired learning outcomes, assessment design and construct representation constitute the first component of the measurement framework.

Algorithmic Consistency and Scoring Invariance

In AI-assisted evaluation, reliability goes beyond whether an AI system consistently generates scores in several applications. The model used, prompt phrasing, rubric design, task features, scoring circumstances, and assessment context can all have an impact on AI-generated results. Large language models can generate helpful assessments that demonstrate relationships with human ratings, according to research on automated essay scoring; however, their performance may differ depending on prompting strategies and scoring conditions (Kim, 2025; Kostic et al., 2024; Pack et al., 2024; Yavuz et al., 2025).

Important issues regarding scoring consistency and invariance are brought up by this variability. The measurement may be sensitive to the assessment technology rather than just student performance if two AI models, distinct prompts, or variations of the same model yield significantly different scores for the same student response. The significance of analyzing the consistency and comparability of AI-generated scores with human judgments is also demonstrated by evidence from automated essay assessment (Kim, 2025; Pack et al., 2024). The effects of raters, tasks, and scoring systems on assessment outcomes can be demonstrated by more

sophisticated measurement techniques, such as many-facet models (Uto & Aramaki, 2024).

Therefore, rather than being a one-time feature of an AI model, reliability should be seen as a continuous attribute of the evaluation system. When models, prompts, rubrics, or assessment circumstances change, institutions should keep an eye on scoring consistency. Unusual or unstable AI judgments can also be found via human moderation and comparison with established scoring systems. As a result, this dimension highlights algorithmic consistency, scoring invariance, rapid sensitivity, and ongoing monitoring as crucial elements of reliable AI-assisted assessment.

Sociotechnical Equity and Bias

AI-assisted evaluation functions within larger social, linguistic, cultural, and technological contexts and is not only a technical measuring procedure. Even when an evaluation has a high scoring consistency, some groups may nonetheless receive unjust results. Differences in AI literacy, socioeconomic status, language background, digital access, technical expertise, and familiarity with AI tools are some potential causes of inequality (Beckman et al., 2025; Hornberger et al., 2025a; Mansoor et al., 2024). The emerging evidence on the “GenAI divide” suggests that students do not have equal opportunities to access, understand, or benefit from generative AI technologies (Beckman et al., 2025). Such differences may influence students' ability to use AI effectively and may therefore become embedded in assessment outcomes. Similarly, AI systems may respond differently to linguistic and cultural variations in student writing, raising concerns about potential bias in automated assessment. These issues are particularly important in multilingual and international higher-education contexts where students may use English or other languages as additional languages.

Fairness should therefore be considered alongside validity and reliability rather than treated as a separate post-assessment concern. Research on fairness, accountability, transparency, and ethics in higher education emphasizes the importance of examining how AI systems affect different groups and whether assessment decisions can be explained and justified (Memarian & Doleck, 2023). Future assessment research should examine potential differences in AI scoring across language, cultural, demographic,

and educational groups and investigate whether such differences represent meaningful construct differences or measurement bias. Psychometric approaches such as differential item functioning and measurement invariance can contribute to this evaluation where appropriate.

Governance, Academic Integrity, and Human-in-the-Loop Safeguards

The fourth dimension concerns the governance structures required to ensure that AI-assisted assessment remains educationally and ethically defensible. Generative AI has complicated traditional understandings of academic integrity because students may use AI to generate ideas, improve language, structure arguments, or produce substantial portions of assessed work. Consequently, academic integrity cannot be addressed effectively through detection technologies alone. Studies have raised concerns about the limitations of AI-detection tools and the difficulty of reliably distinguishing AI-generated from human-generated text (Fleckenstein et al., 2024; Perkins et al., 2024; Weber-Wulff et al., 2023).

Clear institutional policies are therefore essential. Universities need to specify when and how AI may be used, what forms of AI assistance must be disclosed, and which assessment activities should demonstrate unaided individual performance. Evidence from university guidelines and recent assessment scholarship supports moving toward transparent AI-use policies, redesigned assessment tasks, and approaches that emphasize authentic evidence of learning rather than relying primarily on AI detection (Moorhouse et al., 2023; Ncube et al., 2026; Perkins et al., 2024).

Human oversight remains central to this process. AI-generated scores, feedback, or recommendations should not automatically be treated as final assessment judgements, particularly when decisions have significant consequences for students. Human-AI collaboration can combine the efficiency and consistency of automated systems with professional judgement and contextual understanding. This approach is consistent with broader calls for greater ethics, collaboration, transparency, and methodological rigour in AI research in higher education (Bond et al., 2024). Academic integrity, therefore, should be understood as part of a broader governance framework involving transparent policies, student and staff AI literacy, responsible AI use, human moderation, and institutional accountability (Ateeq et al., 2024; Karkoulou et al., 2025).

Interrelationships Among the Four Dimensions

The four dimensions are interconnected rather than sequential. Assessment design and construct representation determine what evidence is collected and what the AI system is expected to measure. Algorithmic consistency and scoring invariance determine whether that evidence can be scored consistently across models, prompts, tasks, and occasions. Sociotechnical equity and bias determine whether the resulting assessment process operates fairly across different student groups. Governance, academic integrity, and human oversight provide the institutional mechanisms needed to ensure that AI-assisted assessment is used responsibly and that assessment decisions remain educationally defensible.

The framework therefore conceptualizes trustworthy AI-assisted assessment as a product of interaction among the assessment task, student, lecturer, AI system, scoring criteria, technological infrastructure, and institutional policies. AI literacy cuts across the framework because both students and educators require sufficient knowledge to use, evaluate, and critically interpret AI-generated outputs (Hornberger et al., 2023; Kong et al., 2025). Similarly, human oversight operates across all four dimensions by supporting construct interpretation, monitoring scoring consistency, identifying potential bias, and ensuring compliance with academic-integrity requirements.

The Human-AI Measurement Ecosystem

The synthesis indicates that trustworthy AI-assisted assessment cannot be achieved through technological capability alone. It requires alignment between assessment design, psychometric quality, algorithmic consistency, equity, governance, and human judgement. The proposed framework consequently shifts the discussion from whether AI is simply “good” or “bad” for assessment toward the conditions under which AI can provide valid, reliable, fair, and educationally meaningful evidence of learning. This human-AI measurement ecosystem recognizes that assessment quality depends not only on the AI model but also on the intended construct, assessment task, rubric, prompt, user competence, access conditions, institutional policy, and mechanisms for continuous evaluation (Bond et al., 2024; Kasneci et al., 2023; Xia et al., 2024).

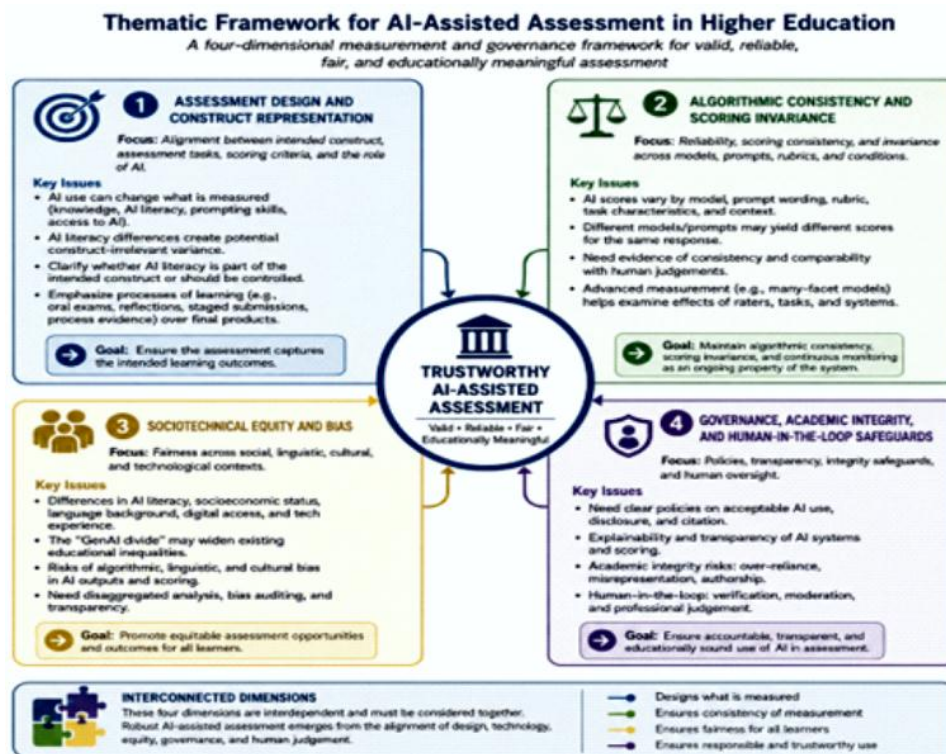


Figure 1. Thematic Framework for Trustworthy AI-Assisted Assessment in Higher Education

Figure 1 presents the four interconnected dimensions of trustworthy AI-assisted assessment in higher education: assessment design, algorithmic consistency, sociotechnical equity, and governance. Assessment design ensures that AI-assisted tasks measure intended learning outcomes, while algorithmic consistency promotes reliable and comparable scoring. Sociotechnical equity addresses potential biases arising from differences in AI literacy, language, socioeconomic background, and digital access. Governance emphasizes academic integrity, transparent policies, and human oversight in AI-supported assessment decisions. Together, these dimensions provide an integrated framework for promoting valid, reliable, fair, and educationally meaningful AI-assisted assessment.

Research Gaps

Despite the rising volume of research on artificial intelligence in higher education, significant gaps still exist in assessment design, psychometric quality, equity, and governance. Longitudinal study is required to investigate how changes in AI models impact scoring stability, validity, reliability, and fairness over time. Evidence from African higher education, particularly in Nigeria, remains insufficient in terms of AI literacy, internet access, linguistic variety, and institutional preparation. Furthermore, many research focus on correlations between AI-generated and human ratings without thoroughly investigating measurement error, scoring invariance, and differences in performance across student populations. The relationship between AI literacy and assessment fairness has to be investigated further to see if disparities in AI knowledge, prompting abilities, and technology access contribute to measurement bias and educational inequity. Cross-disciplinary research are required to determine the applicability of AI-assisted evaluation in various academic domains. Furthermore, comparing human-only, AI-only, and human-AI collaborative assessment systems could shed light on how human oversight affects assessment quality. Finally, additional research should look into the effects of AI-assisted assessment on student learning, motivation, progression, trust, and educational possibilities. Addressing these deficiencies will help to create legitimate, reliable, fair, and educationally relevant AI-assisted evaluation systems.

Implications**Implications for Students**

Students require AI literacy as part of their overall digital competency. They should grasp how AI systems work, be aware of their limitations, analyze AI-generated data, uncover flaws and bias, and utilize AI responsibly. Effective AI implementation necessitates critical involvement rather than merely accepting machine-generated solutions. The ability to examine, verify, and appropriately utilize AI-generated knowledge is so critical.

Implications for Lecturers

Lecturers require professional development that goes beyond just using AI tools. Training should cover assessment redesign, AI literacy, interpreting

AI-generated scores, academic integrity, fairness, prejudice, data privacy, and fundamental psychometric principles. Lecturers should also recognize that AI-generated outputs necessitate professional judgment and should serve as a supplement rather than an undeniable assessment authority.

Implications for Higher-Education Institutions

Higher-education institutions need coherent frameworks for AI-assisted assessment covering acceptable AI use, assessment design, data privacy, academic integrity, human oversight, fairness, transparency and validation. Such frameworks should also provide mechanisms for students to question or appeal AI-supported decisions. Policies should remain flexible enough to accommodate rapid developments in AI technologies.

Implications for Assessment Practice

The findings suggest greater emphasis on authentic and process-based evidence of learning. Oral examinations, portfolios, demonstrations, reflective activities, project work and staged assignments can provide richer evidence of students' reasoning and development. The aim should not simply be to make assessment "AI-proof" but to ensure that AI use supports the intended learning outcomes without compromising the construct being measured.

Implications for Educational Measurement

The increasing use of AI in assessment creates an important role for educational measurement specialists and psychometricians. Psychometric principles should be applied to determine whether AI-supported scores are valid, reliable and fair, with attention to measurement error, model effects, prompt effects, rater effects and consequential validity. This requires collaboration among educational technologists, AI specialists, lecturers, assessment professionals and psychometricians.

Conclusion

Artificial intelligence-assisted assessment is reshaping higher education through applications such as automated scoring, feedback, assessment generation and personalised learning. However, the thematic analysis of the 30 reviewed studies shows that technological capability alone does not guarantee trustworthy assessment. Six interconnected themes emerged: AI literacy, AI-assisted assessment, validity, reliability, fairness and bias, and

academic integrity. The findings indicate that AI-generated scores can vary across models, prompts and assessment conditions, raising important concerns about measurement validity and reliability. Differences in students' AI literacy may also influence how effectively they use AI during assessment and may introduce construct-irrelevant variance when AI competence is not part of the intended learning outcome. Fairness is another major concern, as algorithmic bias, unequal access to AI and limitations in AI-detection tools may disadvantage particular groups of students. The review therefore concludes that AI-assisted assessment should be understood as a human-AI measurement ecosystem rather than a replacement for conventional assessment. Trustworthy implementation requires appropriate assessment design, AI literacy, strong psychometric evidence, reliable procedures, fairness safeguards, transparent institutional policies and meaningful human oversight. The future of AI-assisted assessment should consequently focus not on whether AI should replace human assessment, but on how it can be responsibly integrated to strengthen evidence of student learning while protecting validity, reliability, fairness and academic integrity.

Recommendations

Based on the thematic findings, the following recommendations are proposed:

1. Higher education institutions should incorporate AI literacy into digital literacy programs to ensure that students comprehend AI systems, evaluate AI-generated content, and use AI responsibly.
2. Universities should provide continuous training covering AI-assisted assessment, assessment redesign, interpretation of AI outputs, academic integrity, fairness, bias and basic psychometric principles.
3. AI-generated scores should not be used independently for high-stakes decisions until adequate evidence of validity, reliability, fairness and consequential validity has been established.
4. AI should support rather than replace professional assessment judgement. Qualified assessors should retain responsibility for consequential decisions concerning grades, progression and academic misconduct.
5. Institutions should increase the use of process-based and authentic approaches, including oral examinations, portfolios, demonstrations,

reflective activities, project work and staged assignments where appropriate.

6. Universities should clearly define when AI use is permitted, restricted or prohibited, explain acceptable levels of assistance and provide fair mechanisms for questioning or appealing AI-supported decisions.
7. AI-detection outputs should be treated as supplementary evidence rather than definitive proof. Suspected cases should be examined using additional evidence and qualified human judgement.

References

- Ateeq, A., Alzoraiki, M., Milhem, M., & Ateeq, R. A. (2024). Artificial intelligence in education: Implications for academic integrity and the shift toward holistic assessment. *Frontiers in Education*, 9, Article 1470979. <https://doi.org/10.3389/feduc.2024.1470979>
- Beckman, K., Apps, T., Howard, S. K., Rogerson, C., Rogerson, A., & Tondeur, J. (2025). The GenAI divide among university students: A call for action. *The Internet and Higher Education*, 67, Article 101036. <https://doi.org/10.1016/j.iheduc.2025.101036>
- Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta-systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21, Article 4. <https://doi.org/10.1186/s41239-023-00436-z>
- Evangelista, E. D. L. (2025). Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. *Contemporary Educational Technology*, 17(1), Article ep559.
- Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence*, 6, Article 100209. <https://doi.org/10.1016/j.caeai.2024.100209>

- Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about artificial intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence*, 5, Article 100165. <https://doi.org/10.1016/j.caeai.2023.100165>
- Hornberger, M., Bewersdorff, A., Schiff, D. S., & Nerdel, C. (2025a). A multinational assessment of AI literacy among university students in Germany, the UK, and the US. *Computers in Human Behavior: Artificial Humans*, 4, Article 100132.
- Hornberger, M., Bewersdorff, A., Schiff, D. S., & Nerdel, C. (2025b). Development and validation of a short AI literacy test for university students. *Computers in Human Behavior: Artificial Humans*, 5, Article 100176.
- Kaldaras, L., Akaze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education*, 9, Article 1399377. <https://doi.org/10.3389/educ.2024.1399377>
- Karkoulis, S., Sayegh, N., & Sayegh, N. (2025). ChatGPT unveiled: Understanding perceptions of academic integrity in higher education—A qualitative approach. *Journal of Academic Ethics*, 23, 1171–1188.
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kim, Y. (2025). Automated essay scoring with GPT-4 for a local placement test: Investigating prompting strategies, intra-rater reliability, and alignment with human scores. *TESOL Quarterly*. <https://doi.org/10.1002/tesq.3405>
- Kong, J., Liu, J., Chen, G., & Shang, W. (2025). Assessing AI literacy in college students: The mediating role of self-efficacy in motivational

- commitment pathways. *Education and Information Technologies*, 30, 23957–23979
- Kostic, M., Witschel, H. F., Hinkelmann, K., & Spahic-Bogdanovic, M. (2024). LLMs in automated essay evaluation: A case study. *Proceedings of the AAAI Symposium Series*, 3(1), 143–147.
- Lee, Y.-J., Oh, J., & Hong, C. (2024). Exploratory research on understanding university students' artificial intelligence literacy in a Korean university. *Online Journal of Communication and Media Technologies*, 14(3), Article e202440.
- Mansoor, H. M. H., Bawazir, A., Alsabri, M. A., Alharbi, A., & Okela, A. H. (2024). Artificial intelligence literacy among university students—A comparative transnational survey. *Frontiers in Communication*, 9, Article 1478476. <https://doi.org/10.3389/fcomm.2024.1478476>
- Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5, Article 100152.
- Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, 5, Article 100151.
- Ncube, P. D. N., Dzvapatsva, G. P., Matobobo, C., & Ranga, M. M. (2026). Redefining student assessment in AI-infused learning environments: A systematic review of challenges and strategies for academic integrity. *AI and Ethics*, 6, Article 68.
- Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024). A systematic review of generative AI for teaching and learning practice. *Education Sciences*, 14(6), Article 636.
- Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6, Article 100234.
- Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical

- integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, 21(6).
- Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics*, 22, 89–113.
- Rožman, M., Oreški, D., Elias, A. A., Pynadath, M. F., & Tominc, P. (2025). AI literacy among university students: A comparative study of three countries—Slovenia, Croatia, and India. *IEEE Access*, 13, 110671–110688.
- Tian, J., & Zhang, R. (2025). Learners' AI dependence and critical thinking: The psychological mechanism of fatigue and the social buffering role of AI literacy. *Acta Psychologica*, 260, Article 105725. <https://doi.org/10.1016/j.actpsy.2025.105725>
- Uto, M., & Aramaki, K. (2024). Linking essay-writing tests using many-facet models and neural automated essay scoring. *Behavior Research Methods*, 56, 8450–8479.
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, Article 26.
- Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, 21, Article 40. <https://doi.org/10.1186/s41239-024-00468-z>
- Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, 56, 150–166.
- Zhao, C. (2024). AI-assisted assessment in higher education: A systematic review. *Journal of Educational Technology and Innovation*, 6(4).