Jurnal Elemen, 12(2), 554-576, April 2026
https:/doi.org/10.29408/jel.v12i2.33851

e- ISSN: 2442 - 4226

Generative AI in mathematics education: Considerations for academic integrity and assessment strategies

Kunti Robiatul Mahmudah [1], Nur Robiah Nofikusumawati Peni [1 *], Faida Musa'ad [2], Soth Chea [3], Sommay Shingphachanh [4]

[1] Master Program of Mathematics Education, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[2] Study Program of Mathematics Education, Universitas Muhammadiyah Sorong, Papua Barat Daya, Indonesia
[3] Phnom Penh Teacher Education College, Cambodia
[4] Khangkhay Teacher Training College, Laos

[*] Correspondence: nur.peni@mpmat.uad.ac.id

Abstract

The rapid advancement of generative artificial intelligence (GenAI), particularly tools like ChatGPT, has introduced both opportunities and challenges for academic assessment in higher education. This systematic review explores how GenAI has influenced academic integrity concerns and highlights the assessment redesign strategies proposed or implemented in response. Drawing from 18 peer-reviewed articles published between 2022 and 2025, the review identifies seven key thematic areas: integration of GenAI in educational settings, pedagogical opportunities, integrity-related challenges, impacts on critical thinking and originality, educator and student perspectives, practical implementation outcomes, and strategic recommendations. While GenAI offers personalized feedback, improved access, and scaffolding for learning, it also raises critical issues, including plagiarism, superficial engagement, and the erosion of authorship. The review further reveals a lack of institutional policy, inconsistent ethical guidelines, and disparities in GenAI access among students. In response, researchers advocate for AI-resilient assessment models, ethical literacy, and adaptive institutional frameworks. Although the reviewed studies are general, these issues are critical in mathematics education, where assessment emphasizes reasoning and problem-solving. GenAI may bypass key cognitive processes, undermining assessment validity. The findings suggest proactive, pedagogically informed assessment redesign that leverages GenAI while safeguarding academic integrity, particularly in mathematics learning contexts.

Keywords: academic integrity; assessment redesign; ChatGPT; educational assessment; generative AI

How to cite: Mahmudah, K. R., Peni, N. R. N., Musa'ad, F., Chea, S., & Shingphachanh, S. (2026). Generative AI in mathematics education: Considerations for academic integrity and assessment strategies. *Jurnal Elemen, 12*(2), 554-576. https://doi.org/10.29408/jel.v12i2.33851

Received: 6 January 2026 | Revised: 30 January 2026
Accepted: 4 May 2026 | Published: 7 May 2026

Introduction

The emergence of generative artificial intelligence (AI), particularly large language models like OpenAI's ChatGPT, is changing the face of higher education. Since ChatGPT became widely accessible in late 2022, these tools have found a place in academic spaces and learning processes. They now assist students in drafting essays, summarizing materials, generating data insights, and even creating feedback. With over 100 million users in just two months, ChatGPT became the fastest-growing consumer application in history (Makridakis et al., 2023), signaling how quickly GenAI has been embraced by both students and educators. In many ways, these tools offer exciting potential to support learning, improve productivity and enhance access to information (Mao et al., 2024; Tenakwah et al., 2023). However, their rapid uptake has outpaced the development of pedagogical and assessment frameworks that clearly define appropriate academic use.

At the same time, this rapid integration has raised serious questions about how academic institutions uphold learning integrity. As students gain access to increasingly advanced AI tools, concerns about originality, authorship and fairness have grown. GenAI makes it easier to produce polished assignments with little effort, and that reality has intensified fears about plagiarism and academic dishonesty (Farag et al., 2024). Traditional plagiarism detection systems are often ineffective in this context because AI-generated content does not replicate exact sources. Instead, it produces new but unearned material, making academic misconduct harder to identify (Jiang et al., 2024; Johnson et al., 2024). As a result, assessment has become a central site of tension between technological capability and academic standards. This development puts pressure on educators to rethink how assessments are designed and how student learning is evaluated.

The stakes are not only about cheating. There are broader implications for student development. When learners lean heavily on AI tools, they risk missing opportunities to strengthen critical thinking, creativity and independent problem-solving (Kouam & Muchowe, 2024; Usher, 2025). Students from under-resourced environments may also find themselves disadvantaged if they lack access to quality AI platforms, raising questions about fairness and equity in assessment (Ateeq et al., 2024). Additionally, AI systems are trained on data that may carry embedded biases, which can influence the quality or objectivity of their outputs in educational contexts (Mao et al., 2024). These concerns suggest that the consequences of GenAI use extend beyond misconduct to the core purposes of assessment itself. In the context of mathematics education, these challenges take on additional significance due to the discipline's strong emphasis on problem-solving processes, logical reasoning, and the demonstration of conceptual understanding. Mathematical learning is not only concerned with obtaining correct answers but also with how students construct arguments, justify solutions, and engage in higher-order thinking. The increasing use of GenAI tools in solving mathematical problems raises concerns that students may bypass essential cognitive processes such as formulating strategies, interpreting results, and reflecting on errors. Recent studies suggest that while AI can support procedural fluency and provide step-by-step solutions, it may also reduce opportunities for productive struggle and deep conceptual engagement if used uncritically (Zhai, 2023). In

response, mathematics educators have emphasized the importance of assesssments designs that prioritize reasoning, explanation, and process over final answers, such as open-ended problem-solving tasks, mathematical modelling, and oral justification (Niss & Højgaard, 2019). These approaches align closely with the broader shift toward authentic and process-based assessment identified in this review where mathematics education provides a particularly relevant contxt for examining AI-resilient assessment strategies.

In response to these concerns, educators are exploring new approaches to assessment that preserve academic integrity while recognizing the presence of AI in the learning process. One of the most promising strategies has been the move toward authentic assessments. These tasks require students to apply their knowledge to practical, often real-world challenges, making them less susceptible to AI misuse. Projects, case analyses and reflective tasks ask for personal insight and contextual understanding, which AI tools struggle to replicate effectively (Schultz et al., 2022; Vlachopoulos & Makri, 2024). Such approaches reposition assessment from product verification toward learning process visibility.

Educators are also experimenting with formative assessment strategies that involve feedback during the learning process. AI can play a helpful role here, offering students immediate responses and guidance that help clarify misunderstandings before final evaluations (Bellido-García et al., 2024; Mao et al., 2024; Teng et al., 2024). When feedback is embedded throughout a course rather than concentrated at the end, students are encouraged to stay engaged, reflect on their progress and take greater responsibility for their learning.

Alongside changes in assessment design, there is a growing push to develop students' AI literacy. This includes helping them understand how generative tools work, where they can fall short and how to use them ethically in academic contexts (Ali et al., 2024; Farag et al., 2024). Some institutions are beginning to include critical reflection on AI use as part of their curriculum. Others are asking students to compare their own writing with AI-generated samples or critique the quality of AI responses, building both awareness and judgment (Gander & Harris, 2024; Pratiwi et al., 2025). While the challenges are significant, the rise of GenAI also presents an opportunity to reimagine assessment in ways that are more meaningful and resilient. Educators are moving beyond reactive measures and working toward strategies that support originality, learning depth and fairness. Nevertheless, existing studies remain fragmented across disciplines, assessment types, and research designs, making it difficult to derive coherent guidance for practice.

This review explores how institutions and instructors respond to the academic challenges posed by GenAI, with particular attention to assessment-related implications in higher education. It adopts an assessment-focused lens to examine how integrity concerns are addressed through different redesign strategies. It is guided by the following research question "How has GenAI influenced academic integrity concerns, and what assessment redesign strategies have been proposed or implemented in education to address these challenges?". The purpose of this study is to systematically review recent literature published between 2022 and 2025 that addresses the intersection of GenAI, academic integrity and assessment redesign in education. This timeframe was deliberately selected to capture the most current and relevant body of research following the rapid emergence and widespread adoption of generative artificial

intelligence tools beginning in late 2022. By synthesizing key findings across empirical studies and conceptual papers, this review aims to map the current state of research, distinguish between different types of assessment responses, identify effective practices and inform educators, policymakers and curriculum designers in developing more robust and ethical approaches to student assessment in the AI era.

Methods

This systematic review was conducted using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021) to ensure transparency, methodological rigor, and replicability. The aim of this review was to explore how GenAI has influenced academic integrity in education and to examine the assessment redesign strategies that have been proposed or implemented to address these emerging challenges.

Databases and search strategy

A comprehensive literature search was performed across five major academic databases: Scopus, Web of Science, ERIC, ScienceDirect, and Google Scholar. These databases were selected to capture a broad and interdisciplinary range of relevant literature in the fields of education, technology, and ethics. The search was limited to publications between January 2022 and July 2025 to reflect the recent and rapid development of GenAI tools such as ChatGPT, Bard, and Claude within educational settings. Searches were conducted in English only.

The following Boolean keyword string was used and adapted slightly for each database's syntax: ("academic integrity") AND ("assessment redesign" OR "authentic assessment") AND ("generative AI" OR ChatGPT) AND ("education"). Search results were imported into Zotero reference management software, where duplicate entries were removed automatically. The remaining titles and abstracts were screened independently by reviewers according to predefined eligibility criteria. Articles that passed this initial screening were then retrieved in full and assessed for inclusion. Any discrepancies between reviewers during the selection process were resolved through discussion and consensus.

Inclusion and exclusion criteria

Articles were included in this review if they met the following criteria: (1) focused on education settings, (2) discussed the use or impact of GenAI tools such as ChatGPT, (3) addressed academic integrity concerns or proposed assessment redesign strategies, (4) were based on empirical research, conceptual frameworks, or institutional case studies, and (5) were published in English as peer-reviewed journal articles or full-text conference papers. Studies were excluded if they met one or more of the following conditions: (1) focused on primary education, vocational training, or non-academic contexts, (2) did not explicitly discuss assessment or academic integrity, (3) were editorial, opinion, or commentary pieces lacking empirical or conceptual depth, or (4) were not published in English or did not provide full-text access, (5) not peer-reviewed.

Screening and selection process

The initial database search yielded a total of 4,048 records. After removing duplicates, 450 unique records remained for title and abstract screening. Based on the inclusion and exclusion criteria, 64 full-text articles were reviewed for eligibility. Following 50 articles for full-text screening, 18 articles were retained for final inclusion in the review. The article selection process is visually summarized in the PRISMA flow diagram (Figure 1).

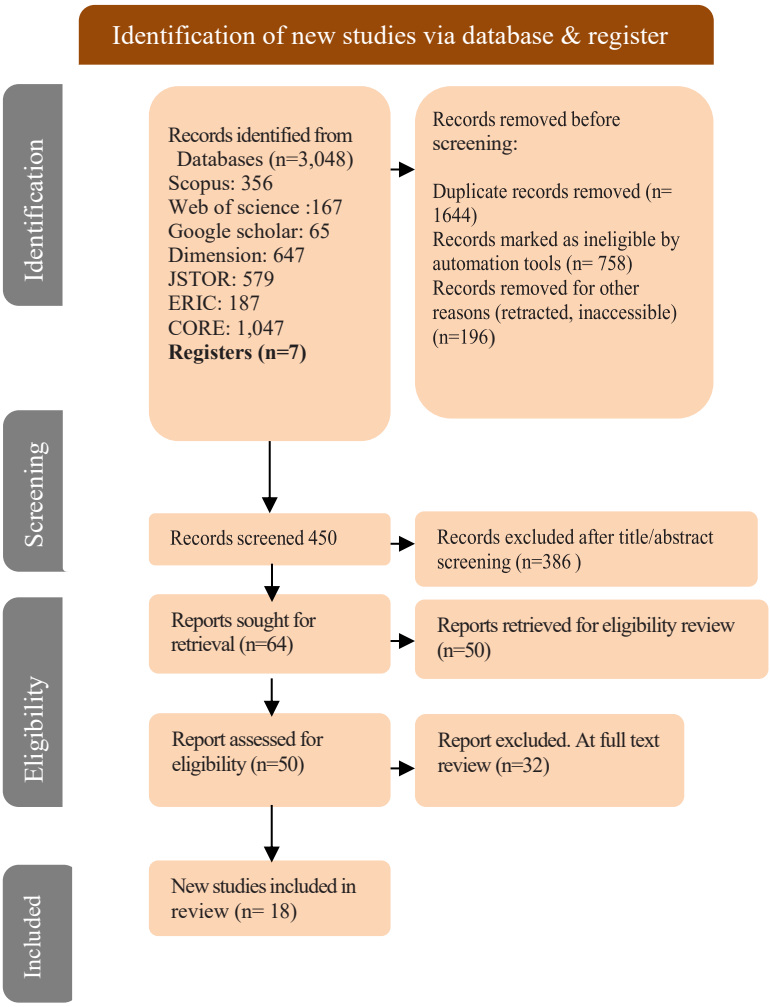

Figure 1. PRISMA 2020 flow diagram template for systematic reviews (adapted from flow diagrams proposed by (Page et al., 2021).

To enhance methodological transparency, the screening process was conducted in two stages (title/abstract screening and full-text review) by two reviewers working independently. Disagreements were resolved through discussion until consensus was reached.

Quality appraisal

Given the inclusion of heterogeneous study designs (empirical, conceptual, and design-based research), a formal risk-of-bias assessment using a single standardized tool was not feasible. Instead, studies were categorized according to methodological type (empirical, conceptual, and design-based) to support differentiated interpretation of findings. Empirical studies were

examined for clarity of research aims, data collection procedures, and analytical transparency, while conceptual and design-based studies were assessed for coherence of argumentation and relevance to assessment and integrity issues. This approach aligns with practices used in emerging research domains where standardized appraisal tools are not yet fully applicable.

Data extraction and synthesis

All 18 included studies were reviewed in full and subjected to a standardized data extraction process. Key information such as author(s), year of publication, research aims, methods, AI tools discussed, and assessment-related findings were collected. Studies were first grouped according to methodological approach (empirical, conceptual, and design-based research) to avoid treating all evidence as methodologically equivalent.

Thematic analysis was then applied to identify recurrent concepts and categories across the studies. This process involved both deductive coding, guided by the review question, and inductive coding, based on patterns emerging from the data. The analysis followed a six-stage procedure: familiarization with the data, initial coding, theme development, theme review, theme definition, and reporting. Two reviewers independently coded all included studies using an open coding approach. Discrepancies in coding were discussed and resolved through consensus, and themes were refined iteratively to ensure internal coherence and alignment with the research objectives.

Two primary themes were identified: (1) threats to academic integrity arising from the use of GenAI, and (2) assessment redesign strategies aimed at addressing these threats in education contexts. These themes are discussed in detail in the Findings section. These themes were further examined using an assessment-focused analytical lens, distinguishing between strategies related to traditional assessment formats and those associated with authentic or process-based assessment approaches. These themes are discussed in detail in the Findings section. While the reviewed studies span multiple educational contexts, the analysis is interpreted with particular attention to implications for mathematics education, especially in relation to assessment and problem-solving processes.

Results

The selected 18 studies represent a diverse mix of geographic origins, disciplinary perspectives, and methodological designs, which reflect the rapidly evolving and interdisciplinary nature of research on GenAI in education assessment and academic integrity. The reviewed studies are situated within the context of education, where emphasize the sector’s immediate concern with how GenAI impacts assessment design, learning processes, and institutional policy (Gruenhagen et al., 2024; Liu, 2025; Lukianenko & Kornieva, 2024; Mao et al., 2024; Saher et al., 2022). This focus underscores the urgency with which universities are grappling with AI-related challenges and transformations. While most studies were centered on general education settings, a few were embedded within specific disciplinary domains such as teacher education (Gruenhagen et al., 2024), health sciences (Saher et al., 2022), language and writing instruction (Lukianenko & Kornieva, 2024), and digital pedagogy (Carbonel et al., 2025). Although the findings are drawn from general educational contexts, they are highly relevant to mathematics

education, where assessment emphasizes reasoning, problem-solving, and conceptual understanding.

Methodologically, the studies span qualitative, quantitative, mixed methods, conceptual, and design science research (DSR) approaches. Several researchers employed qualitative designs to capture student and educator experiences and perceptions (Beynen, 2024; Gruenhagen et al., 2024; Liu, 2025; Lukianenko & Kornieva, 2024; Nikolic et al., 2024; Saher et al., 2022; Stack, 2023) which offer nuanced insights into how GenAI is being used, perceived, and negotiated within the learning environment. Quantitative studies typically focused on measuring trends in AI use, academic performance impacts, or attitudinal differences across demographic groups(Bernal, 2024; Lehane & Wright, 2024; Lukianenko & Kornieva, 2024; Saher et al., 2022; Stack, 2023). Mixed-methods studies provided a comprehensive view by triangulating perspectives and behaviors with data on assessment outcomes or institutional responses (Carbonel et al., 2025; Kofinas et al., 2025; Lukianenko & Kornieva, 2024; Saher et al., 2022). Conceptual research is a non-empirical method that focuses on developing, analyzing, or refining theories and frameworks through logical reasoning and critical examination of existing literature (Almpanis et al., 2025; Chaudhry et al., 2023; Ilieva et al., 2025; Khan et al., 2023; Khlaif et al., 2025; Perkins et al., 2024). Hsiao et al., (2023) adopted a DSR approach to develop a framework that helps educators integrate GenAI ethically into assessment design.

Thematically, a strong and recurring focus across literature is the role of assessment in either enabling or mitigating academic integrity challenges brought about by GenAI (Gruenhagen et al., 2024; Lukianenko & Kornieva, 2024). Many studies argue that traditional forms of assessment, particularly closed-book, product-oriented examinations are increasingly vulnerable to misuse or automation via AI tools. In response, researchers advocate for the design and implementation of authentic assessments that emphasize process, reflection, critical thinking, and originality (Gruenhagen et al., 2024; Saher et al., 2022). These include dialogic assessments, iterative submissions, oral components, and scaffolded tasks that make AI-generated content less applicable or less effective.

Beyond assessment strategies, a recurring concern in the literature is the lack of clear and consistent institutional policies regarding acceptable and unacceptable use of GenAI (Beynen, 2024; Chaudhry et al., 2023; Gruenhagen et al., 2024; Khlaif et al., 2025; Kofinas et al., 2025; Liu, 2025; Perkins et al., 2024). Many students reported uncertainty or confusion about whether and how they were permitted to use tools like ChatGPT, revealing a pressing need for universities to develop transparent guidelines and educational campaigns around AI ethics. Similarly, educators often expressed a lack of confidence in how to adapt their teaching and assessment practices, signaling a parallel need for professional development (Liu, 2025).

The reviewed studies suggest that GenAI poses both risks and opportunities for education. While it challenges longstanding norms around authorship, originality, and effort, it also invites educators and institutions to revisit the purposes and practices of assessment itself. Rather than viewing GenAI solely as a threat to academic integrity, literature points to its potential as a catalyst for more meaningful, skill-based, and student-centered approaches to learning and evaluation (Gruenhagen et al., 2024).

This review of 18 studies identifies seven key thematic areas that shape current discourse on GenAI in higher education assessment. These include: the integration of AI tools into educational settings; emerging opportunities and challenges in academic assessment; the impact of GenAI on critical thinking, writing, and originality; strategies to uphold academic integrity; educator perspectives on teaching and evaluation; student perspectives on AI's role in learning; and practical implementations of GenAI in assessment design. Together, these themes offer a comprehensive understanding of GenAI's transformative yet complex influence on assessment practices. The review not only documents empirical applications of GenAI across diverse contexts but also underscores the urgent need for ethical frameworks, institutional support, and future-oriented strategies to ensure integrity, inclusivity, and pedagogical value in AI-enhanced education.

Integration of AI technologies in educational settings

Exploring current classroom practices involving AI tools and applications.

The integration of GenAI tools, particularly ChatGPT, is reshaping classroom practices in education. Across the 18 reviewed studies, many researchers highlight how ChatGPT has begun to influence both instructional strategies and student learning experiences within formal learning environments. Ilieva et al. (2025) describe the implementation of ChatGPT as part of a larger digital transformation initiative. They report that the use of GenAI facilitated interactive learning, encouraged critical thinking, and allowed students to explore topics with greater autonomy. Instructors used ChatGPT to enhance their digital pedagogy, particularly by prompting students to analyze, reflect, and construct arguments based on AI outputs. They reflect on how AI tools can serve as both teaching aids and sources of concern. While educators recognize the convenience of using ChatGPT to prepare teaching materials and support students with diverse needs, they also call for clear institutional policies to guide its ethical use in classrooms.

Beynen (2024) adopts an autoethnographic lens to reflect on the incorporation of ChatGPT in writing courses. The paper argues that instead of banning GenAI tools, educators should embrace them critically. Beynen's experience shows how AI can be used as a drafting companion and a tool for modeling writing, especially when paired with class discussions about authorship, originality, and voice. Gruenhagen et al. (2024) explore students' actual practices and perceptions regarding ChatGPT. Their study reveals that a significant number of students use ChatGPT to help complete assignments, especially for generating initial drafts, overcoming writer's block, or checking grammar. While the article focuses largely on academic integrity, it provides valuable insight into how AI is already being used informally within learning processes. Student experiences with ChatGPT, indicating that the tool is often used to simplify difficult concepts or to assist with problem-solving tasks. Although students appreciate its utility, the study also notes concerns overreliance and the potential reduction in critical engagement. Saher et al., (2022) discuss authentic assessments as a response to increasing technological involvement in education. While not focused exclusively on AI, their findings

imply that classroom practices are evolving toward more task-based, real-world applications where GenAI can play a role as a supportive or scaffolding tool for students.

Opportunities and challenges for academic assessment

Opportunities: personalized feedback, improved access, scaffolding learning

GenAI offers notable opportunities for education, particularly in delivering personalized feedback, improving access to learning, and scaffolding students' learning processes. Several studies highlight that GenAI tools like ChatGPT and GPT-4 can provide immediate, tailored feedback on students' work, which enables learners to iterate and refine their understanding in real time. Bernal (2024) demonstrated that GPT-driven platforms can dynamically generate multiple-choice questions and detailed coding feedback, fostering adaptive and individualized learning experiences that surpass static content delivery. Similarly, Perkins et al. (2024) emphasized that when integrated ethically, AI tools can be leveraged to provide structured and reflective feedback aligned with different levels of assessment complexity, promoting critical engagement and self-improvement.

Improved access is another key benefit. Tools like GenAI reduce barriers to learning by offering students on-demand academic support, especially for those who might otherwise lack access to human tutors or immediate feedback. Liu (2025) noted that educators see potential for AI to supplement feedback and guidance, particularly in large or diverse classrooms where individual attention is limited. This aligns with Almpanis et al. (2025) who found that students engaged more deeply when AI tools were incorporated as part of critical reflection tasks, allowing them to access new perspectives and refine their work iteratively.

Scaffolding learning through GenAI is increasingly explored as a strategy for building integrity and authentic engagement. Lehane & Wright (2024) designed authentic assessments with scaffolded tasks, peer feedback loops, and co-created rubrics, which significantly enhanced student motivation, engagement, and ethical participation. Likewise, Beynen (2024) stressed that students benefit from scaffolded designs that clarify expectations and provide step-by-step support, helping them navigate the ethical and practical challenges of AI use. Hsiao et al. (2023) also emphasized that well-structured, process-based assessment stages such as drafts and reflective logs supported by AI feedback can improve critical thinking and reduce over-reliance on automated solutions.

Challenges: plagiarism, superficial learning, bypassing cognitive effort

Despite the pedagogical potential of generative AI, its adoption in education introduces significant challenges to academic integrity, notably plagiarism, superficial learning, and the bypassing of cognitive effort. One of the most frequently cited concerns is the risk of undetected plagiarism. Chaudhry et al. (2023) showed that ChatGPT-generated content often evades traditional detection tools like Turnitin and GPTZero, achieving scores equal to or than top-performing student submissions with minimal plagiarism flags. This finding underscores the inadequacy of current academic integrity systems to detect AI-generated texts, raising alarms about inflated grades and misrepresentation of students' actual capabilities. Similarly,

Gruenhagen et al. (2024) stated that students often use GenAI tools to complete substantial portions of their assessments, with institutions failing to provide clear guidelines, leading to unintentional misconduct.

Beyond plagiarism, the risk of superficial learning is a core challenge. Lukianenko & Kornieva (2024) reported that educators were concerned about a decline in students' critical thinking and authentic voice, especially when relying on AI for idea generation and drafting. This was echoed by Mao et al. (2024), whose qualitative study revealed that many instructors feared that AI could enable students to produce polished but shallow work without genuine understanding or reflection.

Furthermore, bypassing cognitive effort has become a pressing issue in assessment contexts. Kofinas et al. (2025) demonstrated that GenAI could generate convincing responses even for authentic tasks, making it difficult for markers to detect misuse. This undermines not only authorship but also engagement with the learning process, as students may be tempted to submit AI-generated answers rather than investing effort in problem-solving or analysis. Stack (2023) noted that this challenge was effectively mitigated by incorporating oral and video-based components into assessments, which forced students to demonstrate comprehension and ownership of their work.

Moreover, the consequentialist mindset of students was identified by Beynen (2024) and Kofinas et al. (2025) as a troubling trend that reflects a lack of ethical clarity around GenAI use, exacerbated by vague institutional messaging. These challenges highlight the urgent need for institutions to rethink assessment practices and integrity policies. Without deliberate redesign and clarity, GenAI may encourage shortcuts that erode academic standards and devalue authentic learning.

Impact on critical thinking, writing skills, and originality

The integration of GenAI in education has sparked intense debate over its impact on students' critical thinking, writing skills, and originality which are core competencies that underpin academic integrity and intellectual development. A growing body of evidence suggests that GenAI use can undermine critical thinking when it substitutes rather than supports cognitive effort. Liu (2025) reported that educators expressed deep concern about the increasing reliance on GenAI outputs, which often results in surface-level understanding and a diminished ability to engage with complex or abstract ideas. Similarly, Lukianenko & Kornieva (2024) found that English teachers were worried about the erosion of analytical thinking and individual interpretation in student essays, as students leaned on AI-generated content instead of formulating original arguments.

This concern extends to writing skills, particularly in disciplines where language development and structure are central to learning outcomes. According to Beynen (2024), Hsiao et al. (2023), Liu (2025), amd Lukianenko & Kornieva (2024) students themselves reported uncertainty about how to appropriately integrate AI tools without compromising their own voice or understanding. The presence of AI-generated grammar-perfect texts often tempted students to bypass the drafting and revision process, leading to a reduction in authentic writing practice. Teachers in the study by Gruenhagen et al. (2024) also noted that students commonly

used GenAI for composing or polishing entire paragraphs, contributing to a decline in personal writing development.

The loss of originality both in thought and expression is another major consequence. Chaudhry et al. (2023) and Stack, 2023) observed that GenAI can easily produce highly plausible academic responses that mirror student work, making it difficult to distinguish between genuine and AI-generated submissions. This poses a threat not only to authorship verification but also to creativity and independent reasoning. Hsiao et al. (2023) explicitly addressed this issue by developing a framework to help educators redesign writing assignments in ways that re-center student voice and discourage uncritical AI dependence.

However, several researchers also proposed ways to leverage GenAI to enhance critical thinking, rather than diminish it. Ilieva et al. (2025) advocated for embedding GenAI analysis and critique directly into assessments such as asking students to identify logical fallacies, bias, or factual errors in AI-generated texts. This pedagogical approach transforms AI from a shortcut into a learning object, promoting meta-cognition and deeper engagement with academic content. Almpanis et al. (2025) reported similar benefits when students were asked to critique GenAI outputs in psychology and HRM modules.

Ensuring academic integrity amidst generative AI innovations

Addressing challenges and developing strategies for upholding ethical standards in educational assessments within a GenAI environment

The rapid proliferation of GenAI tools has created an urgent need for redefining academic integrity in education. Traditional safeguards such as plagiarism detection software and authorship verification are increasingly inadequate in an era where AI-generated outputs can convincingly mimic student work. To uphold ethical standards, institutions and educators must develop proactive, pedagogically sound strategies that move beyond detection and toward integrity-by-design approaches.

Several studies in the review emphasize the limitations of existing integrity mechanisms. Chaudhry et al. (2023) found that GenAI-generated content could bypass standard detection tools like Turnitin and GPTZero with ease, scoring highly across various assessment types. (Gruenhagen et al., 2024) similarly highlighted how blurred boundaries between acceptable and inappropriate AI use have left students confused, often leading to inadvertent breaches of integrity.

To address this, researchers argue for a shift from punitive frameworks to ethical pedagogy. Lehane & Wright (2024) demonstrated that authentic, scaffolded assessments which feature personalization, peer feedback, and iterative submissions can deter misconduct by embedding accountability and transparency into the assessment process. Their framework allowed educators to better understand individual student trajectories, making contract cheating and AI misuse more difficult to execute undetected.

The Artificial Intelligence Assessment Scale (AIAS) proposed by Perkins et al. (2024) offers a structured approach to GenAI integration, ranging from complete prohibition to mandated use, depending on task type and learning outcomes. This continuum helps institutions

define clear expectations and set ethical boundaries while acknowledging the evolving role of AI in learning environments. Moreover, many authors underscore the importance of educating students and staff in AI literacy and integrity. Beynen (2024) and Liu (2025) stressed that without clear institutional policies and instructional support, both students and educators may inadvertently engage in unethical practices. Accordingly, Khlaif et al. (2025) and Hsiao et al. (2023) proposed frameworks that emphasize teacher training, transparency in AI usage, and the co-creation of assessment policies with students to foster a shared culture of responsibility.

Researchers suggest that multi-modal, process-based assessments including oral defenses (Stack, 2023), reflective writing (Ilieva et al., 2025), and collaborative projects can serve as practical integrity measures in a GenAI-enabled world. These tasks not only challenge students to demonstrate their individual thinking but also provide layers of visibility into their learning journey.

Educator perspectives on generative AI's influence on teaching and evaluation

The integration of GenAI in education has sparked a dual narrative among educators, highlighting both its transformative potential and the serious ethical challenges it introduces to academic assessment. While many instructors recognize opportunities to improve engagement, feedback, and instructional design, there remains widespread concern about academic misconduct, insufficient institutional support, and the need for strategic adaptation.

Pedagogical opportunities and assessment innovation

Educators across multiple studies acknowledged that GenAI holds promise for enhancing assessment practices, particularly through personalization, feedback, and engagement. Bernal (2024) demonstrated how GPT-4 was used to generate real-time, individualized feedback and dynamic multiple-choice questions within an eLearning platform. This allowed for more responsive instruction and sustained learner motivation. Similarly, Almpanis et al. (2025) found that integrating AI into psychology and HR modules such as through critical commentary on AI-generated responses encouraged students' metacognitive reflection and fostered deeper learning.

Several researchers proposed structured frameworks to guide responsible use. Perkins et al. (2024) introduced the AI Assessment Scale (AIAS), a five-level model helping educators determine when and how GenAI tools could be ethically and pedagogically integrated from complete prohibition to full collaboration. In a design-focused study, Lehane & Wright (2024) showed how authentic, scaffolded tasks co-created with students could reduce opportunities for misconduct while building assessment literacy. These practices enhanced engagement and provided educators with a clearer window into students' thinking processes, making misuse more difficult to conceal. Educators also began experimenting with oral, video, and process-based assessments (Hsiao et al., 2023; Stack, 2023) to ensure authorship and embed integrity directly into task structure.

Concerns about misuse and institutional gaps

Despite these innovations, concerns about unethical student practices dominate much of the discourse. A major fear is that students may use GenAI to bypass cognitive effort, leading to superficial learning and a loss of critical thinking and writing skills (Liu, 2025; Lukianenko & Kornieva, 2024; Nikolic et al., 2024). Educators observed that students often submitted AI-generated work undetected by standard tools, leading to inflated grades and compromised assessment validity (Chaudhry et al., 2023; Gruenhagen et al., 2024).

Teachers also pointed to the ethical ambiguity surrounding AI use. Many students, particularly in Beynen's (2024) and Gruenhagen et al. (2024) studies, were unsure when GenAI use crossed the line into misconduct which reflecting a broader lack of clear institutional policy. Instructors themselves frequently cited a gap in institutional guidance, as reported by Liu (2025) and Khlaif et al. (2025) noting the absence of consistent training or frameworks for addressing AI's presence in classrooms. Moreover, faculty in studies by Hsiao et al. (2023) and (Ilieva et al., 2025) emphasized that redesigning assessments for AI resilience requires professional development and institutional support. Without shared policy, training, or ethical norms, both educators and students are left to navigate these challenges in isolation, increasing the risk of unintentional or undetected misconduct.

Student perspectives on generative AI's impact on learning and evaluation

Students' experiences and perceptions of GenAI in assessment contexts reveal a nuanced balance between practical benefits and ethical uncertainties, as well as concerns around access and equity. While some students see GenAI tools as valuable support for learning, others express confusion, anxiety, or skepticism, particularly in the absence of institutional clarity and equitable access.

Perceived benefits and practical use

Many students view GenAI tools as useful for idea generation, language support, and productivity enhancement. In the study by Gruenhagen et al. (2024), students reported using tools like ChatGPT for drafting, paraphrasing, or brainstorming assessment responses. Some appreciated the speed, clarity, and grammatical accuracy GenAI offered, especially when managing workloads or overcoming writer's block. Similarly, in Bernal's (2024) study, students using a GPT-powered eLearning platform benefited from real-time, personalized feedback on coding and short-answer questions. This adaptive feedback loop contributed to increased motivation and engagement which is often missing from traditional assessment methods. In Beynen's (2024) study, students expressed appreciation for assessments that were scaffolded, interactive, and personally relevant, noting that such designs clarify expectations and reduce anxiety.

Ethical ambiguity and lack of guidance

Despite these advantages, ethical ambiguity remains a significant concern. Across several studies (Gruenhagen et al., 2024; Beynen, 2024; Liu, 2025), students voiced uncertainty about when GenAI use becomes misconduct. Many believed that if AI was used responsibly for

proofreading or brainstorming, it was acceptable, yet few could clearly define those boundaries, largely due to inconsistent messaging from instructors and institutions.

This ambiguity often caused students to either over-rely on GenAI or avoid it entirely out of fear of being accused of cheating. Some students described a desire for clearer rules, examples of permitted vs. prohibited use, and more open classroom discussions about GenAI ethics (Beynen, 2024; Stack, 2023).

Equity and access disparities

Another concern raised by student data involves disparities in GenAI access and usage skills. As noted in Khlaif et al. (2024), not all students have equal access to premium AI tools or the digital literacy needed to use them effectively. This digital divide introduces inequity in assessment conditions in which students with better AI access may gain unfair advantages in speed, quality, or insight generation. Furthermore, Liu (2025) observed that institutional responses often assumed a homogeneous student population, overlooking linguistic diversity, differing tech access, or variation in academic preparedness. This lack of support particularly affected students from under-resourced backgrounds, who might benefit most from GenAI but have the least ability to access or ethically navigate it

Practical implementation and outcomes of generative AI in assessment

The practical integration of GenAI into education assessment is still in its experimental stages, with varying levels of adoption, outcomes, and institutional readiness. While early implementations show promise, particularly in formative feedback, task redesign, and AI-augmented learning, several studies highlight critical limitations, risks, and operational constraints when using GenAI for grading or evaluating student work.

Embedding GenAI in assessment design

Educators have begun embedding GenAI into assessment workflows in both controlled and exploratory formats. Bernal (2024) presented a robust application through the Learnix platform, which integrates GPT-4 to dynamically generate multiple-choice questions and personalized feedback on short answers and coding tasks. This allowed assessments to be both adaptive and scalable, especially in high-enrollment or technical courses.

Similarly, Lehane and Wright (2024) implemented scaffolded, authentic assessments that aligned well with GenAI-supported learning. While GenAI was not used for grading, students were permitted to use it for idea generation and iterative writing, under the condition of disclosure and critical reflection. The design emphasis was on process visibility where students submitted drafts, received peer feedback, and co-developed rubrics, which discouraged overreliance on GenAI.

Ilieva et al. (2024) recommended the intentional use of GenAI within assessments to encourage students to critically engage with AI-generated texts for example, by identifying inaccuracies or bias. This effort positioned GenAI not as a grading surrogate but as a pedagogical object to promote critical analysis and academic dialogue.

Efficacy and limitations in grading

While the potential for GenAI to support grading exists, actual use of GenAI as an autonomous grading agent remains highly limited and problematic. Bernal (2024) compared GPT-4 to other models (Claude-3-Opus, Gemma-7b) and found that while it provided high-quality feedback in structured technical tasks, its response variability and occasional lack of precision raised concerns for summative grading purposes. Moreover, Stack (2023) and Kofinas et al. (2024) emphasized that GenAI-generated responses can be indistinguishable from high-achieving student submissions. This blurs the boundary between genuine effort and machine output which make grading automation ethically and pedagogically problematic. Educators risk endorsing work that bypasses the student's own thinking if grading is delegated to AI without human oversight.

No study in this review endorsed full automation of grading via GenAI, and most authors expressed caution about its current reliability, fairness, and transparency especially in open-ended or reflective assessments.

Challenges and complications in evaluation

Implementing GenAI in assessment also introduces complications, including inconsistent output quality, opacity in reasoning, and ethical concerns over authorship. Chaudhry et al. (2023) and Liu (2025) both observed that while GenAI could produce grammatically flawless responses, it often lacked depth, nuance, or disciplinary specificity that are essential for fair and meaningful assessment.

There is also the risk of bias and hallucination in AI-generated responses. Several authors, including Almpanis et al. (2024), reported challenges when students submitted AI-generated answers without critical review which promote inaccuracies and superficial work. As a result, evaluation tasks must now include mechanisms to verify processes such as oral defense and revision tracking rather than rely solely on written outputs. Furthermore, using GenAI without clear institutional policy or ethical guidelines complicates assessment validity. Instructors may disagree on what constitutes acceptable use, leading to inconsistent grading and possible unfair penalization (Beynen, 2024; Gruenhagen et al., 2024)

Strategic recommendations for future assessment development

The evolving presence of GenAI in education necessitates a strategic overhaul of assessment design, underpinned by proactive guidance, educator training, and inclusive student support. The reviewed studies consistently call for comprehensive frameworks, professional learning programs, and AI-literate pedagogical practices to build ethically sound and resilient assessment ecosystems.

Frameworks and professional development for educators

Several studies emphasize the need to equip educators with clear, adaptable frameworks that support the ethical and pedagogically sound integration of GenAI. Perkins et al. (2024) proposed the Artificial Intelligence Assessment Scale (AIAS), a five-level model offering

graduated pathways from GenAI prohibition to mandated use, depending on learning outcomes and task types. This model provides structure for institutions in navigating GenAI inclusion and promotes transparency in AI-related assessment policies.

Additionally, Lehane and Wright (2024) highlighted the value of co-designed, scaffolded assessment models that not only promote engagement and integrity but also serve as vehicles for educator development. Hsiao et al. (2023) further recommended the inclusion of workshop-based training, where faculty apply GenAI-integrated frameworks to real assignments and reflect on pedagogical implications.

Student guidance and AI literacy

Students need structured educational resources to ethically and effectively engage with GenAI. Multiple studies (Beynen, 2024; Gruenhagen et al., 2024; Liu, 2025) found that students were often confused about acceptable AI use which results in either overuse or complete avoidance. The solution lies in explicit instruction on AI literacy, academic integrity, and the boundaries of ethical use.

Khlaif et al. (2024) advocated for institution-wide initiatives that support equitable access to AI tools and critical thinking skills. Student-facing resources should include clear use-case examples, interactive AI policy briefings, and reflection-based assignments to help students internalize the role of GenAI in learning rather than misuse it for production-only purposes.

Innovative assessment model development

Several articles recommend a shift toward process-based, multi-modal, and AI-inclusive assessments. Ilieva et al. (2024) and Almpanis et al. (2024) promoted the experimental implementation of tasks that embed GenAI as a subject of critique (e.g., evaluating AI-generated outputs), or that integrate human expression components such as oral defenses and peer interactions.

Stack (2023) demonstrated that requiring students to submit personalized video responses on AI-generated reports significantly reduced misconduct while improving engagement and comprehension. These innovations signal a movement toward authentic assessments that are resistant to GenAI misuse and tailored to diverse student learning needs. The recommended assessment redesign reviewed from these articles are summarized in Table 1 below.

Table 1. Strategic recommendations for assessment practices in the GenAI era

No	Strategy Cluster	Key Assessment Redesign Focus	Underlying Integrity Mechanism	Representative Studies
1	Process-based assessment	Shift from output-based to process-based tasks (drafts, journals, staged submissions, learning logs)	Increases visibility of student thinking and learning progression, reducing feasibility of full AI substitution	Kofinas et al. (2024); Hsiao et al. (2023); Almpanis et al. (2025)
2	Process-based assessment	Scaffolded, progressive, and iterative task design with peer feedback	Embeds accountability and transparency throughout the assessment process	Lehane & Wright (2024); Carbonel et al. (2025)

No	Strategy Cluster	Key Assessment Redesign Focus	Underlying Integrity Mechanism	Representative Studies
3	Multimodal assessment	Combination of written, oral, video, or presentation-based components	Requires demonstration of understanding beyond text-only outputs	Stack (2023); Gruenhagen et al. (2024); Nikolic et al. (2024)
4	Authentic assessment	Real-world, scenario-based, and work-integrated learning tasks	Limits applicability of generic AI-generated responses and promotes contextual reasoning	Saher et al. (2022); Schultz et al. (2022); Vlachopoulos & Makri (2024)
5	Ethical GenAI integration	Explicit permission, limitation, or requirement of GenAI use depending on task goals	Aligns AI use with transparency and ethical reasoning rather than prohibition	Perkins et al. (2024); Khlaif et al. (2024)
6	Ethical GenAI integration	AI critique, comparison, and reflection tasks (human vs. AI outputs)	Repositions GenAI as an object of learning that fosters critical evaluation	Ilieva et al. (2025); Almpanis et al. (2025)
7	Assessment literacy and policy support	Integration of AI literacy, institutional policy clarity, and staff development	Reduces unintentional misconduct and supports consistent assessment practices	Beynen (2024); Liu (2025)
8	Hybrid assessment models	Combining traditional assessment formats with reflective or authentic components	Balances reliability of conventional assessment with integrity resilience	Saher et al. (2022); Lukianenko & Kornieva (2024)
9	Formative assessment and feedback	Use of GenAI-supported formative feedback and adaptive questioning	Supports learning without replacing student cognitive effort	Bernal (2024); Mao et al. (2024); Teng et al. (2024)
10	Integrity-by-design frameworks	Framework-driven assessment redesign (e.g., analyze–redesign–justify models)	Embeds academic integrity principles directly into assessment planning	Ilieva et al. (2025); Khan et al. (2023)
11	Dialogic and reflective assessment	Dialogic tasks requiring justification, reflection, or revision history	Encourages ownership of ideas and discourages superficial AI reliance	Liu (2025); Lukianenko & Kornieva (2024)
12	Collaborative assessment	Collaborative and interactive assessment tasks with shared accountability	Makes individual contribution visible and discourages concealed AI use	Chaudhry et al. (2023); Gruenhagen et al. (2024)

Although the redesign of assessment in response to GenAI has garnered growing attention, several important areas remain insufficiently explored. Chaudhry et al. (2023), Bernal (2024), Maulana et al. (2025) point to a lack of longitudinal studies that evaluate the long-term

impact of GenAI on student learning outcomes, particularly within discipline-specific contexts where cognitive demands and assessment conventions vary widely. Liu (2025) further emphasizes the limited understanding of how GenAI influences assessment equity across diverse student populations, especially with regard to linguistic diversity, cultural differences, and socioeconomic disparities. Additional to this Acopio (2025) mention the usage of technology matter regarding teacher professional development where the preservice teacher must creatively develop their lesson in this era of AI.

Future research should critically examine how GenAI functions within non-textual and practice-based disciplines, such as art, design, and laboratory sciences, where output evaluation relies on skills beyond written expression. There is also a pressing need to evaluate the scalability of AI-integrated formative assessments, especially in large or resource-constrained learning environments. Furthermore, the co-creation of ethical frameworks with students could enhance relevance, transparency, and accountability in AI use. Finally, more comparative, cross-cultural studies are needed to understand variations in students' perceptions, access, and usage patterns of GenAI across different educational systems. Addressing these gaps is essential for developing inclusive, pedagogically sound, and ethically grounded assessment practices in an AI-enhanced academic future.

Discussion

The reviewed studies suggest that GenAI poses both risks and opportunities for education. While it challenges longstanding norms around authorship, originality, and effort, it also invites educators and institutions to revisit the purposes and practices of assessment itself. Rather than viewing GenAI solely as a threat to academic integrity, literature points to its potential as a catalyst for more meaningful, skill-based, and student-centered approaches to learning and evaluation (Gruenhagen et al., 2024; Kasneci et al., 2023).

From a Mathematics Education perspective, these findings are particularly consequential, as the discipline places central importance on reasoning, problem-solving processes, and the justification of solutions rather than merely correct answers (Findell et al., 2001; Niss & Højgaard, 2019). The increasing use of GenAI tools risks externalizing core cognitive processes such as strategy formulation, representation, and reflective verification, thereby potentially weakening students' opportunities to develop deep mathematical understanding (Schoenfeld, 2016; Zhai, 2023). This concern aligns with established views that mathematical proficiency involves the integration of conceptual understanding, procedural fluency, and strategic competence (Findell et al., 2001).

In this light, the shift toward authentic, process-based, and multi-modal assessments identified in this review is consistent with established practices in mathematics education, including mathematical modeling and explanation-centered tasks (Lithner, 2008; Stylianides et al., 2017; Weinhandl et al., 2020). Such approaches increase the cognitive and epistemic demand of assessment tasks by requiring students to articulate, justify, and reflect on their reasoning processes, thereby limiting reliance on superficial or AI-generated responses that do not demonstrate genuine understanding (Darling-Hammond et al., 2020; Wang, T, 2023).

At the same time, emerging research suggests that generative AI can support learning when embedded within structured and reflective assessment environments, rather than used as a substitute for thinking (Kasneci et al., 2023; Zhai, 2023). Therefore, the challenge for mathematics educators is not simply to restrict AI use, but to design assessment tasks that re-center reasoning, transparency, and intellectual accountability.

Conclusion

GenAI is reshaping the landscape of education assessment in profound ways. Based on the current body of literature reviewed, this systematic review demonstrates that while the risks to academic integrity are real, particularly regarding authorship, plagiarism, and superficial learning, there is also significant potential for GenAI to enhance feedback, personalization, and engagement when implemented thoughtfully.

The key to addressing GenAI's challenges lies not in prohibiting its use, but in transforming assessment design. Process-based, interactive, and contextually grounded tasks can discourage misuse while reinforcing meaningful learning. However, it is important to note that much of the evidence supporting these approaches is derived from conceptual, exploratory, or small-scale empirical studies. At the same time, clear policies, ethical guidance, and AI literacy programs are essential to equip both educators and students for responsible engagement with these technologies.

Therefore, institutions must invest in empirical research, cross-sector collaboration, and sustain professional development. Given the relatively small number of included studies (n = 18) and the predominance of non-experimental research designs, the conclusions drawn from this review should be interpreted cautiously. As AI continues to advance, the standards and approaches that uphold academic integrity and educational excellence must also adapt and evolve.

Future research should prioritize robust empirical and longitudinal investigations to validate the effectiveness of proposed assessment redesign strategies across diverse disciplinary and institutional contexts. The future of assessment will depend not only on how we respond to the capabilities of GenAI, but on how we reimagine learning considering its presence..

Acknowledgment

The authors acknowledge Diktilitbang PP Muhammadiyah for funding this research under the fundamental research scheme RisetMU 2024 Batch VIII. This support has been instrumental in enabling the successful completion of this study.

Declaration

Conflict of interest : The authors declare no conflict of interest regarding the publication of this manuscript. In addition, the authors have completed the ethical issues, including plagiarism, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies.

Generative AI Statement	:	Generative AI tools, such as Grammarly and Microsoft Copilot, were employed solely for language editing and minor phrasing enhancements. All conceptualization, analysis, and scholarly content were independently developed and verified by the authors.
Funding statement		This work was supported by the Muhammadiyah National Research Grant Batch VIII, 2024 Number: 0258.028/I.3/D/2025
Author contribution		**Kunti Robiatul Mahmudah:** Conceptualization, writing - original draft, methodology, and formal analysis; **Nur Robiah Nofikusumawati Peni:** Writing-review, editing, investigation, and visualization; **Faida Musa'ad:** Validation, writing-review, and editing; **Soth Chea:** Writing-review, editing, and visualization; **Sommay Shingphachanh:** Writing-review, editing, and supervision.

References

Acopio, M. K. M. G. (2025). Technological proficiency and online resource utilization in mathematics education: A study of higher education instructors in the Philippines. *Jurnal Elemen, 11*(4), 845-859. https://doi.org/10.29408/jel.v11i4.32134

Ali, O., Murray, P. A., Momin, M., Dwivedi, Y. K., & Malik, T. (2024). The effects of artificial intelligence applications in educational settings: Challenges and strategies. *Technological Forecasting and Social Change*, *199*, 123076. https://doi.org/10.1016/j.techfore.2023.123076

Almpanis, T., Conroy, D., & Joseph-Richard, P. (2025). Practical implications of generative AI on assessment: Snapshot of early reactions to assessment redesign in an HRM and a psychology course. *Electronic Journal of E-Learning*, *23*(3), 19–29. https://doi.org/10.34190/ejel.23.3.3971

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*. https://doi.org/10.3389/feduc.2024.1470979

Bellido-García, R. S., Venturo-Orbegoso, C. O., Cruzata-Martínez, A., Sarmiento-Villanueva, E. B., Corro-Quispe, J., & Rejas-Borjas, L. G. (2024). Involvement of the student in their learning: Effects of formative assessment on competency development. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(5), em2440. https://doi.org/10.29333/ejmste/14453

Bernal, M. E. (2024). Revolutionizing elearning assessments: The role of GPT in crafting dynamic content and feedback. *Journal of Artificial Intelligence and Technology*, *4*(3), 188–199. https://doi.org/10.37965/jait.2024.0513

Beynen, T. (2024). The role of students. In *Assessment Literacies in Navigating University Assessment, GenAI, and Academic Integrity A journal of educational research and practice* (Vol. 33, Issue 3). https://journals.library.brocku.ca/brocked

Carbonel, H., Belardi, A., Ross, J., & Jullien, J. M. (2025). Integrity and motivation in remote assessment. *Online Learning Journal*, *29*(2), 25–46. https://doi.org/10.24059/olj.v29i2.4309

Chaudhry, I. S., Sarwary, S. A. M., El Refae, G. A., & Chabchoub, H. (2023). Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT — A case study. *Cogent Education*, *10*(1). https://doi.org/10.1080/2331186X.2023.2210461

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied developmental science*, 24(2), 97-140. https://doi.org/10.1080/10888691.2018.1537791

Farag, W. A., Nadeem, M., & Helal, M. (2024). Assessment transformation in the age of AI: Moving beyond the influence of generative tools. *2024 Mediterranean Smart Cities Conference (MSCC)*, 1–6. https://doi.org/10.1109/MSCC62288.2024.10697011

Findell, B., Swafford, J., & Kilpatrick, J. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. National Academies Press. https://doi.org/10.17226/9822

Gander, T., & Harris, G. (2024). Understanding AI literacy for higher education students: Implications for assessment. *He Rourou*, 8. https://doi.org/10.54474/herourou.v1i1.10579

Gruenhagen, J. H., Sinclair, P. M., Carroll, J. A., Baker, P. R. A., Wilson, A., & Demant, D. (2024). The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments. *Computers and Education: Artificial Intelligence*, *7*. https://doi.org/10.1016/j.caeai.2024.100273

Hsiao, Y. P., Klijn, N., & Chiu, M. S. (2023). Developing a framework to re-design writing assignment assessment for the era of Large Language Models. *Learning: Research and Practice*, *9*(2), 148–158. https://doi.org/10.1080/23735082.2023.2257234

Ilieva, G., Yankova, T., Ruseva, M., & Kabaivanov, S. (2025). A framework for generative AI-driven assessment in higher education. *Information*, *16*(6), 472. https://doi.org/10.3390/info16060472

Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? *Computers & Education*, *217*, 105070. https://doi.org/10.1016/j.compedu.2024.105070

Johnson, S., Owens, E., Menendez, H., & Kim, D. (2024). Using ChatGPT-generated essays in library instruction. *The Journal of Academic Librarianship*, *50*(2), 102863. https://doi.org/10.1016/j.acalib.2024.102863

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Khan, M. M., Dong, Y., & Manesh, N. A. (2023). Authentic assessment design for meeting the challenges of generative artificial intelligence. *Proceedings - Frontiers in Education Conference, FIE*. https://doi.org/10.1109/FIE58773.2023.10343376

Khlaif, Z. N., Alkouk, W. A., Salama, N., & Abu Eideh, B. (2025). Redesigning assessments for AI-enhanced learning: A framework for educators in the generative AI era. *Education Sciences*, *15*(2). https://doi.org/10.3390/educsci15020174

Kofinas, A. K., Tsay, C. H. H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13585

Kouam, A. W. F., & Muchowe, R. M. (2024). Exploring graduate students' perception and adoption of AI chatbots in Zimbabwe: Balancing pedagogical innovation and development of higher-order cognitive skills. *Journal of Applied Learning & Teaching*, *7*(1). https://doi.org/10.37074/jalt.2024.7.1.12

Lehane, S., & Wright, A. (2024). Designing authentic assessment to improve academic integrity. *International Conference on Higher Education Advances*, 564–571. https://doi.org/10.4995/HEAd24.2024.17136

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in mathematics*, 67(3), 255-276. https://doi.org/10.1007/s10649-007-9104-2

Liu, X. (2025). Navigating uncharted waters: Teachers' perceptions of and reactions to AI-induced challenges to assessment. *Asia-Pacific Education Researcher*, *34*(2), 711–722. https://doi.org/10.1007/s40299-024-00890-x

Lukianenko, V., & Kornieva, Z. (2024). Generative AI in student essays: English teachers' perspectives on effective assessment methods. *XLinguae*, *17*(4), 235–250. https://doi.org/10.18355/XL.2024.17.04.14

Makridakis, S., Petropoulos, F., & Kang, Y. (2023). Large language models: Their success and impact. *Forecasting*, *5*(3), 536–549. https://doi.org/10.3390/forecast5030030

Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends*, *68*(1), 58–66. https://doi.org/10.1007/s11528-023-00911-4

Maulana, A., Murtafiah, W.,Handhika, J.,&, Alvares, J. I. (2025). Integrating augmented reality with the e-IM3 structured thinking model to enhance problem-solving skills and learning interest in elementary spatial geometry. *Jurnal Elemen, 11*(4), 1030-1049. https://doi.org/10.29408/jel.v11i4.32139

Nikolic, S., Sandison, C., Haque, R., Daniel, S., Grundy, S., Belkina, M., Lyden, S., Hassan, G. M., & Neal, P. (2024). ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian Journal of Engineering Education*, *29*(2), 126–153. https://doi.org/10.1080/22054952.2024.2372154

Niss, M., & Højgaard, T. (2019). Mathematical competencies revisited. *Educational studies in mathematics, 102*(1), 9-28. https://doi.org/10.1007/s10649-019-09903-9

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, n160. https://doi.org/10.1136/bmj.n160

Perkins, C., Furze, M., Roe, L., & Macvaugh, J. (2024). The artificial intelligence assessment scale (AIAS): A framework for ethical integration of generative AI in educational assessment. In *Journal of University Teaching and Learning Practice* (Issue 6).

Pratiwi, H., Suherman, Hasruddin, & Ridha, M. (2025). Between shortcut and ethics: Navigating the use of artificial intelligence in academic writing among Indonesian doctoral students. *European Journal of Education*, *60*(2). https://doi.org/10.1111/ejed.70083

Saher, A. S., Ali, A. M. J., Amani, D., & Najwan, F. (2022). Traditional versus authentic assessments in higher education. *Pegem Egitim ve Ogretim Dergisi*, *12*(1), 283–291. https://doi.org/10.47750/pegegog.12.01.29

Schoenfeld, A. H. (2016). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics (Reprint). *Journal of education*, 196(2), 1-38. https://doi.org/10.1177/00220574161960

Schultz, M., Young, K., K. Gunning, T., & Harvey, M. L. (2022). Defining and measuring authentic assessment: a case study in the context of tertiary science. *Assessment &*

Evaluation in Higher Education, *47*(1), 77–94. https://doi.org/10.1080/02602938.2021.1887811

Stack, M. (2023). Investigating an assessment design that prevents students from using ChatGPT as the sole basis to pass assessment at the tertiary level. *E-Journal of Humanities, Arts and Social Sciences*, 64–77. https://doi.org/10.38159/ehass.20234127

Stylianides, G. J. (2009). Reasoning-and-proving in school mathematics textbooks. *Mathematical thinking and learning*, 11(4), 258-288. https://doi.org/10.1080/10986060903253954

Tenakwah, E. S., Boadu, G., Tenakwah, E. J., Parzakonis, M., Brady, M., Kansiime, P., Said, S., Ayilu, R., Radavoi, C., & Berman, A. (2023). *Generative AI and higher education assessments: A competency-based analysis*. https://doi.org/10.21203/rs.3.rs-2968456/v2

Teng, M. F., Mizumoto, A., & Takeuchi, O. (2024). Understanding growth mindset, self-regulated vocabulary learning, and vocabulary knowledge. *System*, *122*, 103255. https://doi.org/10.1016/j.system.2024.103255

Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education*, 1–16. https://doi.org/10.1080/02602938.2025.2487495

Vlachopoulos, D., & Makri, A. (2024). A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations. *Studies in Educational Evaluation*, *83*, 101425. https://doi.org/10.1016/j.stueduc.2024.101425

Wang, T. (2023, August). Navigating generative AI (ChatGPT) in higher education: Opportunities and challenges. *In International Conference on Smart Learning Environments* (pp. 215-225). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-5961-7_28

Weinhandl, R., Baldinger, S., & Riegler, V. (2025). Design characteristics for discovery learning within digital mathematics learning environments from students' perspectives. *International Journal of Science and Mathematics Education*, 1-29. https://doi.org/10.1007/s10763-025-10619-x

Zhai, X. (2023). ChatGPT for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students, 29*(3), 42-46. https://doi.org/10.1145/3589649