

What should assessment measure when AI can produce the output? A conceptual framework for human-centered assessment in online education

Mesut Aydemir^{a*} 

^a Anadolu University, Türkiye

Suggested citation: Aydemir, M. (2026). What should assessment measure when AI can produce the output? A conceptual framework for human-centered assessment in online education. *Journal of Educational Technology and Online Learning*, 9(3), 537-551.

Highlights

- Generative AI undermines the construct validity of output-focused assessment, necessitating a paradigm shift toward evaluating distinctly human capacities.
- Four human capacities resistant to AI replication are identified: contextual judgment, lived-experience originality, dialogic meaning-making, and reflective action under uncertainty.
- A four-principle framework (processuality, dialogicity, situatedness, reflexivity) provides principled foundations for human-centered assessment design in online education.

Abstract

Background: Generative artificial intelligence (AI) systems can now produce written texts, solve complex problems, and create academic artifacts previously treated as reliable indicators of student learning. This capability undermines the construct validity of output-focused assessment, a challenge especially acute in online and distance learning (ODL) environments where AI agents can complete assessments on students' behalf. Purpose: This paper identifies the distinctly human capacities that educational assessment should prioritize and proposes conceptual principles for an assessment paradigm adequate to cultivating and evaluating those capacities across educational levels and delivery modes. Method: A systematic conceptual review was conducted across ERIC, PsycINFO, Scopus, and Web of Science (November 2022 to March 2026). From 874 initial records, 90 sources (64 journal articles, 17 monographs, 9 policy reports) were selected through a PRISMA-adapted screening process and synthesized using an integrative conceptual approach. Findings: The review identifies four distinctly human capacities resistant to AI replication (contextual judgment, lived-experience originality, dialogic meaning-making, and reflective action under uncertainty) and derives a four-principle assessment framework: processuality, dialogicity, situatedness, and reflexivity. These principles are organized along temporal/contextual and individual/relational axes. Implications: The framework offers assessment designers, particularly in online and technology-enhanced learning contexts, principled foundations for creating tasks that evidence genuine human learning rather than automatable output production. Implications for educational policy, teacher education, equity, and future empirical research are discussed.

Article Info: Research Article

Keywords: *generative artificial intelligence; educational assessment; human-centered assessment; online learning; assessment design framework*

1. Introduction

The rapid integration of generative artificial intelligence (AI) into educational settings poses a fundamental challenge to established conceptions of teaching, learning, and assessment. Contemporary AI systems can produce written essays, solve complex problems, generate multimedia presentations, and simulate forms of reasoning that were previously regarded as reliable indicators of student achievement. This capability undermines the construct validity of assessment practices that treat submitted products as transparent evidence of learning. Assessment can no longer rely primarily on evaluating polished outputs when those outputs may be wholly or partially generated by AI at any educational level, from early childhood through post-secondary education. There is accordingly a need for assessment models that foreground distinctly

* Corresponding author. Open Education Faculty, Anadolu University, Türkiye.
e-mail address: mesutaydemir@anadolu.edu.tr

Doi: <http://doi.org/10.31681/jetol.1954582>

Received 19 May 2026; Accepted 23 May 2026

ISSN: 2618-6586. This is an open Access article under the CC BY license.



human capacities, including contextual judgment, originality rooted in lived experience, meaning-making through dialogue, and reflective action under uncertainty, which AI systems cannot consistently replicate. From this perspective, generative AI should be understood not merely as a threat to academic integrity but as a catalyst for redefining what education values and what assessment should authentically measure.

This paper addresses two interrelated questions: (1) What are the distinctly human capacities that educational assessment should prioritize in the age of generative AI? (2) What conceptual principles should organize an assessment paradigm adequate to the cultivation and evaluation of those capacities across educational levels?

The crisis is especially pronounced in open and distance learning (ODL) environments, which expanded significantly following the COVID-19 pandemic and now constitute a substantial proportion of global tertiary enrollment (Zawacki-Richter et al., 2019). In online assessment environments, academic integrity may be compromised both by contract cheating, in which students outsource assessed work to third parties, and by AI-enabled writing technologies, which can generate human-like text with minimal task-specific training and produce outputs that may be difficult to distinguish from human writing, thereby raising concerns about academic integrity (Cotton et al., 2024). This challenge is especially acute in ODL contexts, where agentic AI systems can now log into learning management systems with valid student credentials and complete assessments on students' behalf (Bozkurt et al., 2026). Although the ODL context sharpens the challenge, the epistemological problem extends beyond higher education: the structural question of whether assessment measures genuine human learning or merely the capacity to produce valued artifacts applies across all educational levels and delivery modes.

Educational assessment has long been shaped by prior theories of learning and measurement, and large-scale testing practices have typically relied on standardized conditions and tasks that can be scored reliably and efficiently (Pellegrino et al., 2001). The dominant logic was one of credentialing: submitted products such as essays, examination scripts, and problem solutions were treated as transparent windows onto student competence. Contemporary generative AI systems are increasingly capable of producing student-like academic writing; in a blinded evaluation of short-form physics essays, GPT-4-generated responses were found to be comparable in quality to student essays and difficult for human markers to distinguish reliably from human-authored work (Yeadon et al., 2024). The arrival of these systems does not merely create new opportunities for dishonesty; it exposes the structural fragility of an assessment paradigm that never adequately addressed the question of what student-produced artifacts were designed to evidence in the first place.

Existing scholarship offers foundational resources for a principled response. Biesta's (2010) tripartite framework distinguishes qualification, socialization, and subjectification as the irreducible purposes of education, arguing that assessment has become disproportionately focused on the first at the expense of the third. Generative AI intensifies this imbalance by automating the measurable products of qualification, thereby revealing the poverty of an assessment system that equates learning with output production. Boud and Falchikov's (2006) concept of sustainable assessment similarly argues that assessment must develop students' capacities for judgment, self-regulation, and lifelong learning, capacities that no AI system can substitute. The educational institution of the future must accordingly engage in a principled examination of which distinctly human capacities ought to be nurtured, evaluated, and safeguarded in an era of intelligent machines (Biesta, 2013; Holmes et al., 2019; Luckin, 2018; Selwyn, 2019a).

This paper contributes a systematic, philosophically grounded conceptual framework for human-centered assessment that addresses the epistemological challenge generative AI poses to assessment in online and technology-enhanced learning environments. Unlike responses that concentrate on detection technologies or incremental task redesign, the proposed framework of processuality, dialogicity, situatedness, and reflexivity offers a principled reconceptualization of assessment purpose, design, and practice. By synthesizing educational philosophy, assessment theory, and AI scholarship, the paper reframes the generative AI challenge as an occasion to articulate what distinctly human capacities educational assessment must cultivate and evidence. The framework carries practical implications for assessment task

design in online learning, institutional policy, teacher education, and the empirical research agenda required to evaluate human-centered assessment in practice.

2. Literature

This section reviews the literature that informs the present study, organized around five thematic areas: the evolving landscape of artificial intelligence in education, foundational theories of educational assessment, philosophical perspectives on distinctly human capacities and educational purpose, sociocultural and dialogic frameworks for learning and assessment, and emerging approaches to assessment design in online and distance learning environments. Together, these bodies of scholarship establish the conceptual terrain from which the proposed framework for human-centered assessment is derived.

2.1. Artificial Intelligence in Educational Contexts

Research on artificial intelligence in education has evolved substantially over the past decade, shifting from explorations of adaptive tutoring systems and learning analytics toward confrontation with the implications of generative AI for teaching, learning, and assessment. Zawacki-Richter et al. (2019), in a systematic review of AI applications in higher education, identified a significant gap: the majority of studies focused on technical system development and performance prediction, with little attention to educators' perspectives or the pedagogical implications of AI integration. Holmes et al. (2019) offered a broader view, arguing that AI in education raises not only technical questions but also curricular, pedagogical, and ethical ones, including implications for assessment design, the evolving role of educators, and the preparation of students to navigate complexity with adaptability and wisdom. Selwyn (2019a) advanced a critical perspective, questioning whether AI-driven educational technologies genuinely serve educational purposes or primarily serve the logics of efficiency and scalability that may undermine the relational and developmental dimensions of education.

The release of large language models capable of generating human-like academic text has intensified these concerns. Cotton et al. (2024) documented the challenge that AI-enabled writing technologies pose to academic integrity, noting that such systems can produce text that may be difficult to distinguish from human writing. Yeadon et al. (2024), in a blinded evaluation of physics essays, found that GPT-4-generated responses were comparable in quality to student essays and that human markers could not reliably differentiate between human-authored and AI-generated work. The challenge extends beyond text generation: Bozkurt et al. (2026) identified the emergence of agentic AI systems capable of logging into learning management systems with valid student credentials and completing assessments on students' behalf, a development that transforms academic dishonesty from an individual act of plagiarism into a structural vulnerability of the assessment system itself. Sadasivan et al. (2023) further demonstrated that current AI-text detection tools are unreliable in practical settings, as simple paraphrasing can substantially reduce their effectiveness, rendering detection-based responses structurally inadequate.

Institutional and policy responses to generative AI in education have begun to take shape, though they remain uneven. Chan (2023) proposed a comprehensive AI policy framework for university teaching and learning encompassing assessment redesign, institutional investment in training and infrastructure, and transparent AI governance. At the international level, UNESCO (2023) issued recommendations on the ethics of AI in education, while the European Parliament and Council (2024) adopted the EU AI Act, which establishes harmonized rules on artificial intelligence with implications for educational applications. Bearman and Ajjawi (2023) argued for a pedagogical rather than purely regulatory response, advocating that students develop evaluative judgment and epistemic doubt through meaningful interaction with AI systems rather than seeking to eliminate AI from educational processes entirely.

2.2. Foundations of Educational Assessment Theory

The theoretical foundations of educational assessment provide essential context for understanding both the structural vulnerabilities that generative AI exposes and the conceptual resources available for a principled response. The distinction between formative and summative assessment, developed by Black and Wiliam (1998) in their seminal review, established that assessment practices oriented toward improving learning

during instruction yield substantial gains in student achievement. The Assessment Reform Group (1999, 2002) formalized this distinction into principles of assessment for learning, which emphasized the role of feedback, self-assessment, and learner agency in the assessment process. Bennett (2011) subsequently offered a critical review of formative assessment, cautioning that the term had been applied so broadly as to risk losing its theoretical coherence, while Wiliam (2011) argued for embedding formative assessment within everyday instructional practice rather than treating it as a separate activity.

Pellegrino et al. (2001) provided a foundational framework for understanding assessment as a reasoning system built upon models of cognition, observation, and interpretation, arguing that assessment design must be grounded in explicit theories of what students know and how they learn. Brown (2022), in a transdisciplinary historical review, traced the evolution of educational assessment from early psychometric testing to contemporary competency-based approaches, noting that large-scale testing practices have typically privileged standardized conditions and reliable scoring over the richer but less measurable dimensions of learning. Boud and Falchikov (2006) advanced the concept of sustainable assessment, contending that assessment must develop students' capacities for judgment, self-regulation, and lifelong learning rather than merely certifying achievement at a single point in time. Nicol and Macfarlane-Dick (2006) complemented this perspective by articulating seven principles of good feedback practice grounded in self-regulated learning theory, emphasizing that effective assessment supports students in monitoring and closing the gap between their current and desired performance. Collectively, these works establish that assessment theory has long recognized the limitations of product-focused, summative evaluation, a critique that generative AI renders both more visible and more urgent.

2.3. Philosophical Perspectives on Human Capacities and Educational Purpose

A central concern of this study is the identification of distinctly human capacities that educational assessment should prioritize, a question that draws on a rich tradition in philosophy of education. Biesta's (2010) tripartite framework distinguishes qualification, socialization, and subjectification as the irreducible purposes of education, arguing that contemporary assessment systems have become disproportionately focused on qualification at the expense of subjectification, the process through which students come to exist as subjects of initiative and responsibility. In subsequent work, Biesta (2013) characterized education as an inherently risky endeavor in which outcomes cannot be fully predetermined, and Biesta (2020) revisited the relationship among the three purposes, emphasizing that subjectification requires assessment practices that treat students as agents rather than objects of measurement. More recently, Biesta (2022) developed a world-centred conception of education that critiques the reduction of education to learning and competence acquisition, a perspective with direct implications for assessment design in an era when AI can efficiently produce the outputs that competence-based systems are designed to measure.

Complementary philosophical traditions reinforce the centrality of human capacities in educational purpose. Nussbaum (1997) argued that liberal education must cultivate critical self-examination, cosmopolitan empathy, and narrative imagination, capacities that require the active engagement of a particular human self with the world. In later work, Nussbaum (2010) extended this argument to contend that democracy itself depends on educational practices that develop the humanities-oriented capacities that standardized testing tends to neglect. Kristjánsson (2015), drawing on Aristotelian virtue ethics, argued that education must attend to the formation of character, including the practical wisdom (phronesis) necessary to judge wisely in particular situations and the moral courage to act on that judgment. Dewey's (1910) foundational account of reflective thought, which holds that genuine thinking originates in a felt difficulty or state of perplexity that suspends immediate judgment and initiates inquiry, provides the epistemological grounding for understanding reflection as a distinctly human cognitive process that AI systems, architecturally constituted to reduce uncertainty by selecting high-probability responses, cannot replicate. Luckin (2018) brought these philosophical concerns into dialogue with AI scholarship, arguing that the future of education requires clarity about the relationship between machine learning and human intelligence, and that educational systems must prioritize the development of uniquely human capacities that intelligent machines cannot substitute.

2.4. Sociocultural and Dialogic Frameworks

Sociocultural theory provides a further critical lens for understanding what is at stake when AI systems mediate or replace the processes through which learning occurs. Vygotsky's (1978) concept of the zone of proximal development established that cognitive development is fundamentally social, occurring through interaction with more capable others in culturally situated contexts. Bakhtin's (1981) theory of dialogism extended this insight by emphasizing that meaning is constituted through the responsive encounter between different voices and perspectives, a process that cannot be reduced to information transmission or pattern recognition. Mercer (2000) developed the concept of interthinking to describe how collaborative verbal reasoning enables forms of understanding that individual cognition cannot achieve independently, while Mercer et al. (2019) explored the educational implications of dialogic interaction in technology-mediated classrooms, arguing that digital technologies can support but should not supplant the genuine intersubjective encounters through which understanding is constructed.

Wegerif (2007) articulated a dialogic theory of education in which the central aim is to open and sustain dialogic space, enabling learners to be drawn beyond themselves by engaging with the essential otherness of others. This perspective has particular relevance for assessment, as it implies that certain forms of understanding can only be evidenced through genuine participation in dialogue and cannot be adequately captured by individually produced artifacts. Recent empirical work has begun to explore how dialogic principles can be operationalized in technology-enhanced assessment contexts. Tang et al. (2024) demonstrated in a proof-of-concept study that students engaged in dialogic interaction with generative AI exhibited follow-up questioning indicative of critical thinking and creativity. Zhang and Liu (2025), in a design-based study, developed a dialogic feedback triangle comprising cognitive, socio-affective, and structural dimensions, showing that structured human–AI dialogue in writing tasks can cultivate genuine meaning-making rather than mere content generation. These emerging findings suggest that dialogic approaches to assessment offer a promising pathway for evidencing distinctly human capacities even in technology-mediated learning environments.

2.5. Cognitive and Metacognitive Dimensions of Learning in the Age of AI

A growing body of research examines how interaction with generative AI affects the cognitive and metacognitive processes that assessment is designed to develop and evaluate. Bjork and Bjork (2020) established that desirable difficulties, conditions that introduce challenge, retrieval effort, and generation demands, enhance long-term learning and transfer even when they reduce immediate performance. Kapur (2015) extended this insight through the concept of productive failure, demonstrating that initial struggle with complex problems, prior to direct instruction, produces deeper conceptual understanding than more efficient instructional sequences. These findings have direct implications for assessment in the age of AI: when students use generative AI to bypass the effortful processes that promote learning, assessment may fail to capture, or indeed to cultivate, the cognitive capacities it is intended to evaluate.

Recent empirical evidence supports these concerns. Gerlich (2025) found that greater reliance on AI tools is associated with cognitive offloading and diminished critical thinking, while Stadler et al. (2024) demonstrated that students using large language models in scientific inquiry tasks exhibited lower mental effort but also lower-quality reasoning and justification. Teng and Yue (2023) established through structural equation modeling that metacognitive writing strategies and critical thinking skills are positively associated with academic writing performance, suggesting that overreliance on AI-generated text may reduce opportunities to develop both writing competence and the reflective processes that underpin it. Wang (2025) provided further evidence from both native and non-native English speakers that AI-assisted writing processes alter the cognitive demands of composition in ways that may attenuate the development of independent writing ability. On the metacognitive dimension, Tanner (2012) emphasized the importance of structured opportunities for students to monitor their own understanding, while Ryan and Ryan (2013) proposed a model for teaching and assessing reflective learning that moves beyond surface-level reporting toward the reconstruction of understanding and future practice. These perspectives underscore the need for assessment designs that protect and cultivate the cognitive and metacognitive processes that generative AI may otherwise allow students to circumvent.

2.6. *Authentic and Alternative Assessment in Online and Distance Learning*

The challenge of designing assessment that resists AI substitution has prompted renewed interest in authentic assessment approaches, particularly in online and distance learning contexts where the structural vulnerability to AI-mediated proxy assessment is greatest. Gulikers et al. (2004) proposed a five-dimensional framework for authentic assessment that emphasizes the alignment of task, physical context, social context, assessment result, and criteria with professional practice. Villarroel et al. (2018) developed a blueprint for authentic assessment design organized around cognitive challenge, evaluative judgment, and realism. However, recent empirical evidence suggests that product-focused authentic assessments alone may be insufficient in the age of generative AI. Kofinas et al. (2025), in a controlled experimental study, demonstrated that generative AI could produce submissions across assessments of varying authenticity levels that passed the scrutiny of experienced academic markers, leading them to conclude that the focus must shift from assessing output to assessing process and interpersonal engagement.

Emerging practitioner-oriented frameworks have begun to address this challenge. Khlaif and Awadallah Alkouk (2024) developed a Process-Product Assessment Approach through faculty training workshops in the Global South, foregrounding critical thinking, creativity, and collaboration as assessable process indicators alongside final products. Elshall and Badir (2025) proposed the FACT assessment framework, which strategically differentiates AI-free components that verify core competencies from AI-assisted applied projects that embed students in authentic professional contexts of human–AI collaboration. Selwyn (2019b), in his critique of learning analytics and datafication, warned that when education is judged primarily through measurable outputs, assessment risks privileging proxies and artifacts over learning itself, thereby creating conditions in which externally produced work can more easily pass as evidence of achievement. These diverse strands of scholarship collectively point toward the need for a comprehensive conceptual framework that moves beyond incremental task redesign to reconceptualize what assessment should measure and how it should be structured in the age of generative AI, particularly in the online and distance learning environments where the challenge is most acute.

3. Methodology

This study employs a systematic conceptual review methodology that emphasizes the role of the conceptual framework in linking theoretical perspectives, research design, and the interpretation of evidence (Leshem & Trafford, 2007). Rather than generating new empirical data, it constructs and justifies a theoretical argument through the systematic synthesis of existing scholarship. Conceptual research of this kind is recognized as a necessary form of scholarly inquiry in educational science (Gilson & Goldberg, 2015), particularly when emergent phenomena, such as the rapid integration of generative AI into educational systems, have outpaced empirical investigation.

The author approaches this inquiry from a position informed by critical educational philosophy and sociocultural theory, with a commitment to assessment practices that prioritize human agency and educational purpose over measurement efficiency. This orientation shapes the selection and interpretation of sources and should be acknowledged as a framing influence on the resulting framework.

The literature search was conducted across four databases: ERIC, PsycINFO (via APA PsycNet), Scopus, and Web of Science, supplemented by targeted searches in Google Scholar for grey literature and policy reports. The search period was defined as November 2022 to March 2026, corresponding to the emergence, widespread diffusion, and institutional consolidation of large language models following OpenAI's public release of ChatGPT in November 2022, an event widely recognized as marking generative AI's transition from a research context to a mass-accessible educational technology. The extended endpoint of March 2026 was adopted to encompass the full arc of the initial LLM proliferation phase, including the rapid successive releases of GPT-4, Gemini, Claude, and open-source Llama series models throughout 2023 to 2025, as well as emerging institutional and regulatory responses such as the UNESCO Recommendation on the Ethics of AI in Education (2023) and the EU AI Act (2024). Foundational pre-2022 works in educational philosophy and theory (Dewey, 1910; Vygotsky, 1978), assessment theory (Boud & Falchikov, 2006), and dialogic pedagogy (Bakhtin, 1981; Mercer, 2000) were included where theoretically essential and remain widely

cited in contemporary literature; Biesta (2010) was also included for its influential critique of measurement-led education and its emphasis on educational purpose. The initial search yielded 874 records. After automated deduplication, 631 unique records remained. Following title and abstract screening, 157 records were retained for full-text review. Full-text assessment against the inclusion and exclusion criteria resulted in a final corpus of 64 peer-reviewed journal articles, 17 scholarly monographs, and 9 institutional or policy reports (N = 90 sources). The systematic search and selection procedure is illustrated in Figure 1 (adapted from PRISMA 2020). It should be noted that this study follows an integrative review logic rather than a full systematic review protocol (Snyder, 2019). As Snyder (2019) notes, integrative reviews are particularly suited to emergent topics that span multiple research traditions, enabling the generation of new conceptual frameworks rather than the aggregation of comparable empirical findings. The adapted PRISMA 2020 flow diagram (Page et al., 2021) is employed not because the study follows a full systematic review protocol, but to provide a transparent, reproducible account of the source identification, screening, and selection process; this methodological transparency strengthens the auditability of an integrative review without altering its fundamentally conceptual rather than meta-analytic synthesis logic, in which sources are analyzed for their theoretical contributions rather than aggregated statistically.

The search strategy employed a structured combination of keywords connected through Boolean operators to maximize both precision and recall across the four databases. The AND operator was used to intersect thematic domains (e.g., artificial intelligence with educational assessment), ensuring that retrieved records addressed the intersection of relevant fields rather than either domain in isolation. The OR operator was used to capture terminological variation within each domain (e.g., “generative AI” OR “large language models”), thereby broadening retrieval to include studies using different but conceptually equivalent terminology. Truncation (*) was applied where appropriate to capture morphological variants (e.g., “assess*” to retrieve assessment, assessing, and assessments). Quotation marks were used to search for exact multi-word phrases. Table 1 presents the complete search strategy with Boolean operators as applied across all databases.

Table 1.

Boolean Search Strategy Applied Across Databases

	Search Component	Search Terms and Boolean Operators
S1	Artificial Intelligence	“generative AI” OR “generative artificial intelligence” OR “large language model*” OR “ChatGPT” OR “GPT-4”
S2	Educational Assessment	“educational assess*” OR “authentic assess*” OR “formative assess*” OR “summative assess*” OR “academic integrity”
S3	Educational Context	“higher education” OR “online learning” OR “distance learning” OR “open and distance learning” OR “blended learning”
S4	Dialogic and Process-Oriented	“dialogic assess*” OR “dialogic education” OR “process-oriented assess*” OR “reflective assess*”
S5	Philosophy of Education	“philosophy of education” OR “educational philosophy” OR “human capacit*” OR “critical thinking” OR “metacognition”
S6	Combined Search	S1 AND (S2 OR S3) AND (S4 OR S5)
S7	Supplementary Search	“AI AND education AND technology” (Google Scholar, targeted grey literature)

Note. The asterisk (*) denotes truncation to capture variant word endings. Quotation marks indicate exact phrase searching. The combined search (S6) represents the primary retrieval strategy applied across ERIC, PsycINFO, Scopus, and Web of Science. The supplementary search (S7) was conducted in Google Scholar to identify grey literature and policy reports.

Inclusion criteria were: (a) peer-reviewed empirical, theoretical, or systematic review articles addressing AI, assessment, or educational philosophy in post-secondary or secondary education; (b) scholarly monographs providing foundational theoretical frameworks; and (c) institutional reports from recognized bodies (OECD, UNESCO, European Commission) addressing AI in education. Exclusion criteria were articles focused exclusively on technical AI model development without educational application;

journalistic or opinion pieces lacking peer review; and studies in languages other than English. The conceptual synthesis proceeded through three stages: (1) mapping longstanding critiques of output-focused assessment that generative AI now intensifies; (2) synthesizing accounts of distinctively human cognitive and moral capacities resistant to AI substitution; and (3) generating a principled conceptual framework for human-centered assessment. The systematic search and selection procedure is illustrated in Figure 1 below, which provides a schematic representation of the four-stage source identification, screening, and inclusion process. The diagram illustrates the four-stage procedure used to identify and select the final source corpus. The procedure is adapted from the PRISMA 2020 reporting guidelines (Page et al., 2021).

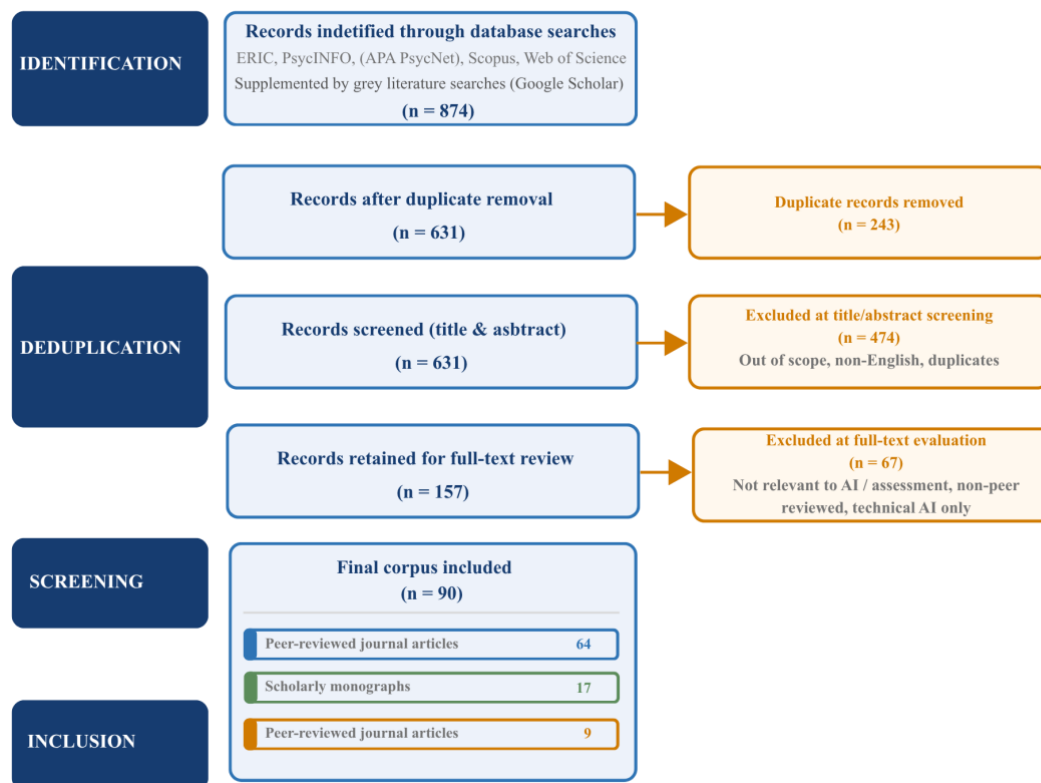


Fig. 1. PRISMA-Adapted Flow Diagram of the Systematic Literature Search (November 2022 to March 2026).

4. Findings and Discussion

This study set out to address two interrelated questions: (1) What are the distinctly human capacities that educational assessment should prioritize in the age of generative AI? (2) What conceptual principles should organize an assessment paradigm adequate to the cultivation and evaluation of those capacities across educational levels? Through a systematic conceptual review of 90 sources drawn from ERIC, PsycINFO, Scopus, and Web of Science (November 2022 to March 2026), the study synthesized educational philosophy, assessment theory, and AI scholarship to construct a principled response to the epistemological challenge that generative AI poses to assessment, particularly in online and distance learning (ODL) environments.

The review first revealed two structural limitations of conventional assessment that generative AI exposes. The attribution problem, wherein AI-text detectors are unreliable in practical settings (Sadasivan et al., 2023) and agentic AI systems can now complete assessments on students' behalf (Bozkurt et al., 2026), renders detection-based responses structurally inadequate. The validity problem is equally serious: when students use AI to produce assessed work, they may bypass the cognitively effortful processes that promote learning, including the desirable difficulties (Bjork & Bjork, 2020) and productive failure (Kapur, 2015) that deepen understanding, as well as the metacognitive and critical thinking processes positively associated with academic writing performance (Teng & Yue, 2023). Additionally, assessment systems built around individual, decontextualized tasks suppress the irreducibly social character of cognition established by

sociocultural theory (Vygotsky, 1978; Bakhtin, 1981), a suppression made more consequential when AI can simulate interactional reasoning.

In response to these structural vulnerabilities, the review identified four distinctly human capacities that are both educationally central and resistant to AI replication: (a) contextual judgment, the situated ability to perceive the morally and practically significant features of particular human situations (Nussbaum, 2010); (b) originality grounded in lived experience, distinct from the statistical novelty AI generates; (c) dialogic meaning-making, the construction of understanding through sustained, responsive dialogue with other persons (Mercer, 2000; Wegerif, 2007); and (d) reflective thinking under uncertainty, grounded in Dewey's (1910) account of reflective thought and extended by Biesta's (2013) characterization of education as an inherently risky endeavor.

Each of these four human capacities maps onto a corresponding assessment principle that translates the capacity into a design orientation. Contextual judgment, which requires perceiving the morally and practically salient features of particular situations, is addressed by the principle of situatedness, which embeds assessment tasks in professional or real-world contexts requiring such judgment. Originality grounded in lived experience, which cannot be reduced to statistical recombination, is addressed by processuality, which shifts the locus of assessment from final products to the temporal processes of thinking, drafting, and revising through which original insight develops. Dialogic meaning-making, constituted through responsive encounter with other voices, is addressed by dialogicity, which centers assessment in real-time conversational and relational contexts. Finally, reflective thinking under uncertainty is addressed by reflexivity, which develops students' capacity to critically examine their own thinking, assumptions, and learning processes. From these one-to-one correspondences, the study derived a four-principle conceptual framework for human-centered assessment: processuality, dialogicity, situatedness, and reflexivity.

Figure 2 presents a schematic representation of these four capacities and their systemic relationship. The dashed outer ring indicates the boundary beyond which AI cannot currently operate reliably, demarcating four distinctly human capacities that AI systems can simulate linguistically but cannot genuinely perform.

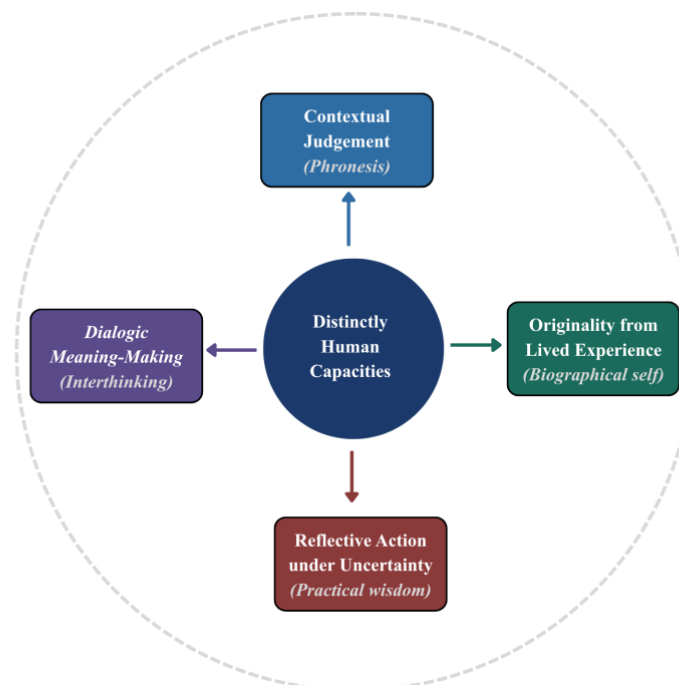


Fig. 2. Four Distinctly Human Capacities in the Age of Generative AI

These four principles are organized along two axes: the horizontal axis distinguishes temporal/process-oriented from contextual/situated assessment, while the vertical axis distinguishes individual-focused from collective/relational assessment. The bidirectional arrows connecting each adjacent pair of principles indicate active mutual reinforcement. Figure 3 illustrates the framework.

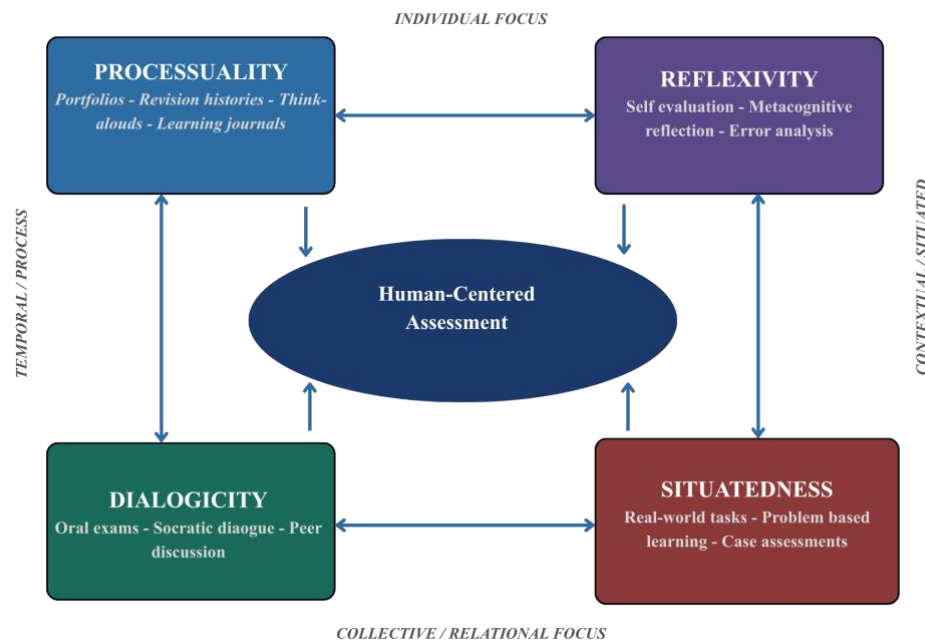


Fig. 3. Human-Centered Assessment Framework: Four Interrelated Principles

The framework carries significant implications for assessment practice across educational levels. For online and blended learning environments, the principles can be operationalized through e-portfolio platforms, synchronous oral assessment, workplace-integrated projects, and structured metacognitive journals. Emerging empirical evidence provides preliminary support: process-oriented approaches have shown promise in faculty development contexts (Khlaif & Awadallah Alkhouk, 2024), dialogic assessment designs leveraging AI as a structured interlocutor have demonstrated capacity to evidence critical thinking (Tang et al., 2024; Zhang & Liu, 2025), and frameworks that strategically differentiate AI-free and AI-assisted components (Elshall & Badir, 2025) illustrate concrete pathways for operationalizing these principles. At the same time, product-focused authentic assessments remain vulnerable to AI-generated submissions (Kofinas et al., 2025), underscoring the urgency of the paradigmatic shift this framework proposes. For educational policy, the dominant framing of AI as primarily a tool for personalization and scalability risks being epistemically inadequate; applying Biesta's (2022) framework, policy responses should distinguish between uses that preserve the conditions for subjectification and uses that reduce education to the efficient production of outcomes. For teacher education, the framework implies that educators need not only technical knowledge of AI tools but the philosophical literacy to understand what is educationally at stake in assessment design (Holmes et al., 2019).

Table 2.

Assessment Design Exemplars for Online Learning Environments

Principle	Assessment Task	Technology Platform	What It Evidences	How It Resists AI Substitution
Processuality	Iterative e-portfolio with timestamped drafts: semester-long portfolio with weekly drafts and reflective annotations on revisions and conceptual shifts.	LMS e-portfolio (e.g., Mahara, PebblePad) with version history and timestamped logs.	Cognitive growth over time; self-regulated revision; feedback integration.	Requires temporal thinking record across weeks in the student's voice. AI cannot retrospectively generate an authentic developmental arc or genuine responses to specific feedback.

Principle	Assessment Task	Technology Platform	What It Evidences	How It Resists AI Substitution
Processuality	Think-aloud screencasting: students record narrated screen-capture of their reasoning during complex problem solving, including false starts and self-corrections.	Screen recording tools (e.g., Loom) with LMS submission.	Real-time reasoning; metacognitive monitoring; productive struggle and strategy adaptation.	Demands spontaneous verbal reasoning synchronized with on-screen actions. Temporal correspondence between speech and behavior cannot be AI-generated.
Dialogicity	Synchronous oral defense: live video session where the examiner poses adaptive follow-up questions probing reasoning depth and conceptual boundaries.	Video conferencing (e.g., Zoom, BigBlueButton) with recording for moderation.	Reasoning under pressure; responsiveness to challenge; depth of understanding beyond surface articulation.	Requires real-time responsiveness to unpredictable examiner challenges in a verifiable synchronous interaction.
Dialogicity	Collaborative peer argumentation: small groups in synchronous discussion where students build upon, challenge, or synthesize peers' contributions.	Synchronous breakout rooms (e.g., Zoom) with annotation tools (e.g., Hypothes.is, Perusall).	Interthinking; responsiveness to others' reasoning; sustained collaborative meaning-making.	Captures between-person reasoning from actual interaction, which a single AI system cannot generate.
Situatedness	Workplace-integrated project: authentic problem solving in professional context with documented stakeholder interactions and ethical navigation.	E-portfolios, shared cloud folders, and LMS logs with stakeholder verification protocols.	Contextual judgment in authentic settings; ethical navigation with real stakes; disciplinary-practice integration.	Demands verifiable engagement with a real professional context. Evidence includes documented stakeholder interactions that AI cannot fabricate.
Situatedness	Real-time case analysis: students make and justify decisions under time pressure as the instructor reveals new information incrementally.	Branching scenario tools (e.g., H5P, Articulate Storyline) with synchronous instructor-led sessions.	Evaluative judgment under uncertainty; decisive action with incomplete information; context-specific justification.	Time-pressured decisions with unpredictable instructor-introduced developments require a genuinely present human agent.
Reflexivity	Analytics-informed metacognitive journal: students examine their learning processes and set developmental goals using LMS analytics data.	LMS analytics dashboards (e.g., Canvas, Moodle) with structured journal platforms.	Metacognitive awareness; honest self-assessment; identification of conceptual gaps.	Requires genuine self-knowledge from a subject at stake in their own development. Analytics data anchors reflection in actual behavioral patterns.
Reflexivity	Critical incident analysis: students identify a threshold learning moment, analyze its personal significance, and articulate how it reshaped their understanding.	LMS reflective writing with structured prompts and audio/video tools (e.g., Padlet, Loom).	Genuine self-examination; integration of learning with personal biography; professional identity development.	Demands biographical-academic integration requiring a particular human self. AI cannot fabricate authentic critical incidents from actual experience.

Note. Each exemplar is designed for online or blended delivery modes. Technology platforms listed are representative; equivalent tools may be substituted based on institutional context. The resistance to AI substitution derives from the structural features of each task rather than from any single technological safeguard.

5. Conclusion and Suggestions

Several limitations warrant acknowledgment. First, the claim that contextual judgment, originality, and dialogic engagement are currently beyond AI's reliable capacity will be contested as AI capabilities evolve (LeCun, 2022); the framework should be understood as a theoretically grounded response to the current state of AI, requiring revision as that state changes. Second, the tension between the proposed assessment forms and the credentialing demands of labor markets and university admissions systems remains unresolved. Third, the reliance on a conceptual rather than empirical methodology means that the framework's practical effects on student learning, equity, and institutional culture remain to be examined through implementation studies. Fourth, the review's restriction to English-language sources may have excluded relevant perspectives from non-Anglophone educational traditions, particularly those in the Global South where assessment reform faces distinct structural constraints. It should be acknowledged that substantial bodies of ODL assessment scholarship exist in non-Anglophone contexts that this review could not capture. For instance, Türkiye has a well-established open and distance learning tradition through institutions such as Anadolu University, one of the world's largest open universities, and Turkish researchers have contributed significantly to assessment design, technology-enhanced learning, and academic integrity in ODL contexts (e.g., Zawacki-Richter et al., 2019). Similarly, research communities in East Asia, Latin America, and Francophone Africa have developed context-specific approaches to assessment that may offer valuable perspectives on the challenges addressed here. Future iterations of this framework should incorporate multilingual source corpora to strengthen its cross-cultural applicability. The equity dimension also deserves emphasis: the framework's dialogic and processual components are resource-intensive, and a just transition toward human-centered assessment must be accompanied by targeted policy support and equitable resource allocation.

In conclusion, educational institutions best positioned to thrive in the age of generative AI will not be those that have integrated more technology but those that have achieved greater clarity about what human education is for and how assessment can serve that purpose across online and face-to-face contexts alike. Based on the findings of this study, the following suggestions for future research are offered: (a) empirical testing of specific assessment instruments aligned with the proposed framework, with particular attention to feasibility, equity implications, and learning outcomes across diverse educational contexts, including ODL environments; (b) investigation of how initial teacher education and professional development programs can equip educators with the philosophical literacy and practical competences required to implement human-centered assessment at scale; (c) comparative studies examining implementation of processual, dialogic, situated, and reflexive assessment across different institutional contexts, resource levels, and cultural settings, with special attention to the Global South; (d) longitudinal research on the effects of human-centered assessment on student development of contextual judgment, dialogic capacity, and reflective thinking; and (e) policy-oriented research connecting classroom assessment reform to changes in credentialing and qualification systems, the broader institutional environment without which assessment reform cannot realize its full transformative potential.

Declaration of generative AI and AI-assisted technologies in the writing process

Claude Sonnet 4.6 was used to assist with the literature search process and language refinement of the manuscript. All content was critically reviewed, verified, and finalized by the author. The conceptual framework and intellectual contribution remain solely the author's own.

References

- Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box*. University of Cambridge School of Education.
- Assessment Reform Group. (2002). *Assessment for learning: 10 principles — Research-based principles to guide classroom practice*.

- Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (C. Emerson & M. Holquist, Trans.). University of Texas Press.
- Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, 54, 1160–1173. <https://doi.org/10.1111/bjet.13337>
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. <https://doi.org/10.1080/0969594x.2010.513678>
- Biesta, G. (2010). *Good education in an age of measurement: Ethics, politics, democracy*. Paradigm Publishers.
- Biesta, G. J. J. (2013). *Beautiful risk of education*. Routledge. <https://doi.org/10.4324/9781315635866>
- Biesta, G. J. J. (2020). Risking ourselves in education: Qualification, socialization, and subjectification revisited. *Educational Theory*, 70(1), 89–104. <https://doi.org/10.1111/edth.12411>
- Biesta, G. (2022). *World-centred education: A view for the present*. Routledge.
- Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4), 475–479. <https://doi.org/10.1016/j.jarmac.2020.09.003>
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Bozkurt, A., Crompton, H., & Fell Kurban, C. (2026). The devil is in the det[ai]ls: AI agents, ghost students, and the crisis of verified presence in an agentic AI world. *Open Praxis*, 18(1), 1–12. <https://doi.org/10.55982/openpraxis.18.1.1145>
- Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, 31(4), 399–413. <https://doi.org/10.1080/02602930600679050>
- Brown, G. T. (2022). The past, present and future of educational assessment: A transdisciplinary perspective. *Frontiers in Education*, 7. <https://doi.org/10.3389/feduc.2022.1060633>
- Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20(1), 38. <https://doi.org/10.1186/s41239-023-00408-3>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Dewey, J. (1910). *How we think*. D. C. Heath & Co.
- Elshall, A. S., & Badir, A. (2025). Balancing AI-assisted learning and traditional assessment: The FACT assessment in environmental data science education. *Frontiers in Education*, 10, Article 1596462. <https://doi.org/10.3389/feduc.2025.1596462>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
- Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15 (1), Article 6. <https://doi.org/10.3390/soc15010006>
- Gilson, L. L., & Goldberg, C. B. (2015). Editors' comment: So, what is a conceptual paper? *Group & Organization Management*, 40(2), 99–100. <https://doi.org/10.1177/1059601115576425>

- Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, 52(3), 67–86. <https://doi.org/10.1007/BF02504676>
- Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign.
- Kapur, M. (2015). Learning from productive failure. *Learning: Research and Practice*, 1 (1), 51–65. <https://doi.org/10.1080/23735082.2015.1002195>
- Khlaif, Z. N., & Awadallah Alkhouk, W. (2024). AI-resistant assessments in higher education: Practical insights from faculty training workshops. *Frontiers in Education*, 9, 1499495. <https://doi.org/10.3389/feduc.2024.1499495>
- Kofinas, A., 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*, 56(6), 2522–2549. <https://doi.org/10.1111/bjet.13585>
- Kristjánsson, K. (2015). Aristotelian character education. Routledge.
- LeCun, Y. (2022). A path towards autonomous machine intelligence (Version 0.9.2) [Position paper]. OpenReview. <https://openreview.net/pdf?id=BZ5a1r-kVsf>
- Leshem, S., & Trafford, V. (2007). Overlooking the conceptual framework. *Innovations in Education and Teaching International*, 44(1), 93–105. <https://doi.org/10.1080/14703290601081407>
- Luckin, R. (2018). Machine learning and human intelligence: The future of education for the 21st century. UCL IOE Press.
- Mercer, N. (2000). Words and minds: How we use language to think together. Routledge.
- Mercer, N., Hennessy, S., & Warwick, P. (2019). Dialogue, thinking together and digital technology in the classroom: Some educational implications of a continuing line of inquiry. *International Journal of Educational Research*, 97, 187–199. <https://doi.org/10.1016/j.ijer.2017.08.007>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- Nussbaum, M. C. (1997). Cultivating humanity: A classical defense of reform in liberal education. Harvard University Press.
- Nussbaum, M. C. (2010). Not for profit: Why democracy needs the humanities. Princeton University Press.
- Page, M. J., McKenzie, J. E., 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., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press.
- Ryan, M., & Ryan, M. (2013). Theorising a model for teaching and assessing reflective learning in higher education. *Higher Education Research & Development*, 32(2), 244–257. <https://doi.org/10.1080/07294360.2012.661704>
- Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, Z., & Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv. <https://doi.org/10.48550/arXiv.2303.11156>
- Selwyn, N. (2019a). Should robots replace teachers? AI and the future of education. Polity.

- Selwyn, N. (2019b). What's the problem with learning analytics? *Journal of Learning Analytics*, 6(3), 11–19. <https://doi.org/10.18608/jla.2019.63.3>
- Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
- Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160, Article 108386. <https://doi.org/10.1016/j.chb.2024.108386>
- Tang, K. S., Cooper, G., Rappa, N., Cooper, M., Sims, C., & Nonis, K. (2024). A dialogic approach to transform teaching, learning and assessment with generative AI in secondary education: A proof of concept. *Pedagogies: An International Journal*, 19(3), 493–503. <https://doi.org/10.1080/1554480X.2024.2379774>
- Tanner, K. D. (2012). Promoting student metacognition. *CBE—Life Sciences Education*, 11(2), 113–120. <https://doi.org/10.1187/cbe.12-03-0033>
- Teng, M. F., & Yue, M. (2023). Metacognitive writing strategies, critical thinking skills, and academic writing performance: A structural equation modeling approach. *Metacognition and Learning*, 18(1), 237–260. <https://doi.org/10.1007/s11409-022-09328-5>
- UNESCO. (2023). Recommendation on the ethics of artificial intelligence in education. UNESCO.
- Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, 43(5), 840–854. <https://doi.org/10.1080/02602938.2017.1412396>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Wang, C. (2025). Exploring Students' Generative AI-Assisted Writing Processes: Perceptions and Experiences from Native and Nonnative English Speakers. *Technology, Knowledge and Learning*, 30, 1825–1846. <https://doi.org/10.1007/s10758-024-09744-3>
- Wegerif, R. (2007). *Dialogic education and technology: Expanding the space of learning*. Springer.
- Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.
- Yeadon, W., Agra, E., Inyang, O., Mackay, P., & Mizouri, A. (2024). Evaluating AI and human authorship quality in academic writing through physics essays. *European Journal of Physics*, 45(5), 055703. <https://doi.org/10.1088/1361-6404/ad669d>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhang, Y., & Liu, Y. (2025). Designing ChatGPT-mediated feedback activities in EFL writing: A design-based study of the dialogic feedback triangle. *Assessment & Evaluation in Higher Education*. Advance online publication. <https://doi.org/10.1080/02602938.2025.2571846>