



Preparing Teachers to Assess in the Age of Artificial Intelligence: A Critical Integrative Review and Framework for Assessment-AI Knowledge

Konstantinos T. Kotsis ^{1*} 

¹ Department of Primary Education, University of Ioannina, GREECE

* Correspondence: kkotsis@uoi.gr

CITATION: Kotsis, K. T. (2026). Preparing Teachers to Assess in the Age of Artificial Intelligence: A Critical Integrative Review and Framework for Assessment-AI Knowledge. *Educational Point*, 3(2), e167.

<https://doi.org/10.71176/edup/19619>

ARTICLE INFO

Received: 23 August 2026

Accepted: 28 September 2026

OPEN ACCESS

ABSTRACT

Artificial intelligence (AI), particularly generative AI, changes assessment by altering the evidential relationship among student work, learning, and professional judgment. Teachers must design tasks when students can readily obtain machine-generated responses, interpret automated feedback and analytics whose validity is conditional, and justify consequential decisions under uncertainty. This critical integrative review synthesizes scholarship on AI in education, assessment, teacher learning, and emerging professional-development interventions. It proposes Assessment-AI Knowledge (A-AIK) as an assessment-specific configuration of professional knowledge rather than a replacement for assessment literacy, AI literacy, or technological pedagogical content knowledge. A-AIK comprises three interdependent capacities: design competence, which establishes conditions for eliciting meaningful evidence; interpretive competence, which calibrates the meaning, uncertainty, and evidential weight of AI-mediated information; and justificatory competence, which supports transparent, fair, and reviewable decisions. Disciplinary authenticity, human agency, equity, transparency and data governance, and epistemic humility operate as cross-cutting commitments. The reviewed literature suggests that promising interventions use scenarios, authentic assessment redesign, structured critique, peer deliberation, and iterative classroom trials. Recurrent limitations include small samples, reliance on self-report, short duration, and measures of confidence or acceptance rather than demonstrated assessment performance. A four-phase professional-learning model—calibrate, redesign, enact, and justify, is proposed as a synthesis-informed design hypothesis requiring empirical evaluation. The article concludes with provisional performance indicators, institutional conditions, and a research agenda for responsible AI-mediated assessment.

Keywords: artificial intelligence, assessment literacy, generative AI, teacher education, professional development

INTRODUCTION

The educational debate about artificial intelligence has moved with unusual speed from speculation to ordinary professional practice. Generative systems can draft essays, solve familiar quantitative problems, produce computer code, explain scientific phenomena, summarize readings, generate images, and revise text in response to feedback. At the same time, automated scoring, learning analytics, recommender systems, and AI-assisted feedback increasingly present teachers with apparently precise claims about students. These developments create genuine opportunities: teachers can generate task variants, explore misconceptions, differentiate scaffolds, obtain preliminary feedback, and reduce some routine workload. They also create a more fundamental problem. Assessment depends on an evidential inference from what a learner does to what the learner knows, understands, or can do. When a powerful, probabilistic system participates in producing the observed work or in interpreting it, the inference is altered. The central professional question is therefore not simply whether teachers can use AI. It is whether they can sustain valid, fair, and educationally useful assessment in its presence.

This distinction matters because the dominant response in teacher education and professional development has been framed as AI literacy. AI-literacy frameworks appropriately emphasize awareness, use, evaluation, and ethics, and recent international guidance has made an important contribution by placing human agency and responsibility at the center of teachers' competence (Ng et al., 2021; UNESCO, 2024). However, general literacy does not by itself specify the judgments required in assessment. A teacher may know how a large language model generates text, recognize that outputs can be biased or fabricated, and write effective prompts, yet still be unable to determine whether an assignment provides defensible evidence of learning. Conversely, a teacher may have strong conventional assessment literacy but lack the technical understanding needed to appraise the provenance, uncertainty, or systematic limitations of an AI-generated score. The new situation is located at the intersection of assessment knowledge, disciplinary knowledge, pedagogical judgment, AI literacy, and professional ethics.

The urgency is intensified by the limitations of familiar responses. Attempts to prohibit AI are difficult to enforce when access is ubiquitous and often invisible. Attempts to identify AI-generated text through automated detectors are unreliable and vulnerable to paraphrasing, translation, and ordinary variation in human writing; detector outputs are therefore inadequate as stand-alone evidence in misconduct proceedings (Weber-Wulff et al., 2023). At the opposite extreme, unrestricted adoption can weaken the interpretability of student work if the assessed construct is never distinguished from the assistance used to produce the response. Neither prohibition nor uncritical incorporation resolves the evidential problem. Teachers need a principled way to decide when AI use is irrelevant, when it threatens validity, when it should be constrained, and when it is part of the capability that assessment ought to measure.

Recent scholarship has started to articulate the broader contours of this challenge. Work on assessment in the age of AI emphasizes the need to reconsider what assessment claims mean when learning and performance are algorithmically mediated (Swiecki et al., 2022). Research on feedback generated by generative AI shows that fluent output is not equivalent to high-quality judgment and that human and machine feedback differ in accuracy, specificity, relational quality, and pedagogical usefulness (Steiss et al., 2024). Studies of teachers and faculty reveal strong interest in redesigning assessment, but also considerable uncertainty about integrity, policy, and the appropriate division of labor between human and machine (Bower et al., 2024). Emerging interventions in teacher preparation and professional development offer promising examples, including scenario-based learning, critique of AI-generated lesson plans, workshops on AI-resilient assessment, microlearning communities, and the use of generative AI to expose and develop pedagogical content knowledge (Awadallah Alkhouk & Khlaif, 2024; Blonder et al., 2025; Meegan & Young, 2025; Park et al., 2026).

Nevertheless, the evidence base remains young. Much of it measures attitudes, acceptance, confidence, or intended use rather than teachers' demonstrated capacity to make sound assessment decisions.

This article argues that preparing teachers to assess in AI-mediated environments requires an assessment-specific configuration of professional knowledge, termed Assessment-AI Knowledge (A-AIK). A-AIK is not proposed as a freestanding body of knowledge or as a replacement for teacher assessment literacy, AI literacy, or technological pedagogical content knowledge (TPACK). Rather, it describes the integrated capacity to make defensible assessment judgments when AI alters the provenance, interpretability, or evidential weight of evidence of learning. Its distinctiveness is therefore functional and relational rather than additive: it concerns how teachers coordinate assessment, technological, pedagogical, disciplinary, and ethical knowledge across the learner-task-AI-evidence-decision relationship. Three interdependent capacities structure the framework. Design competence concerns the conditions under which meaningful evidence is elicited; interpretive competence concerns the meaning, uncertainty, and evidential weight assigned to AI-mediated information; and justificatory competence concerns the justification, documentation, communication, and review of consequential decisions. Disciplinary authenticity, human agency, equity, transparency and data governance, and epistemic humility operate as cross-cutting commitments across all three capacities.

WHY AI CHANGES THE LOGIC OF ASSESSMENT

Assessment as an Evidential Argument

Assessment is often discussed in practical terms: teachers write questions, collect work, assign scores, and provide feedback. Beneath these activities is an argument. Observed performance is treated as evidence for an interpretation about a learner, and that interpretation is used for some purpose. Kane's (2013) argument-based approach to validation makes this structure explicit: the defensibility of an assessment depends on the warrants connecting observation, scoring, generalization, extrapolation, and decision. Those warrants are never automatic. They depend on task design, response conditions, scoring quality, construct representation, and the consequences of use. AI affects every link. At the elicitation stage, generative systems can contribute ideas, language, procedures, calculations, representations, and complete products. The observed response may therefore reflect an unknown distribution of student and machine work. At the scoring stage, automated systems can classify, rank, comment, or predict, but their output is generated through models that may be opaque, unstable across versions, sensitive to prompt wording, and uneven across groups or content domains. At the generalization stage, apparent performance on one AI-supported task may not transfer to an unaided context or to a different tool. At the extrapolation stage, a polished product may overstate the learner's independent conceptual understanding. At the decision stage, uncertainty can be obscured by a single score or a detector probability presented with unwarranted precision.

The appropriate conclusion is not that AI invalidates assessment. Rather, it changes what must be controlled, documented, and interpreted. If the intended construct is unaided recall, unrestricted AI assistance compromises the claim. If the intended construct is the capacity to use disciplinary knowledge while critically orchestrating AI, banning the tool would underrepresent the construct. If the purpose is formative, AI may provide useful provisional feedback, provided learners and teachers understand its fallibility. Validity therefore depends on alignment among purpose, construct, assistance conditions, evidence, and decision. The teacher's work becomes the design and defense of that alignment.

This evidential perspective also clarifies why the search for universally "AI-proof" tasks is misguided. Any fixed task format may eventually be approximated by increasingly capable systems, and superficial modifications may merely create a temporary technological contest. A more durable goal is AI-resilient assessment: an assessment system whose claims remain interpretable because permitted assistance is explicit, evidence comes from multiple moments and modes, and important judgments are anchored in observable processes

or interaction. Resilience does not require every task to be invulnerable. It requires the overall evidence system to remain meaningful when AI is available.

Generative AI as Participant, Object, and Infrastructure

The role of AI in assessment is not singular. It can be a participant in producing student work, an object of critical evaluation, or infrastructure used by institutions to score and monitor. These roles impose different obligations. When AI participates in production, teachers must define the boundary between legitimate support and construct substitution. When AI is the object of assessment, students may be asked to verify, correct, compare, or improve its output; the ability to critique becomes part of the construct. When AI functions as infrastructure, teachers must decide how much evidential weight to assign to its classifications and whether governance arrangements protect students' rights. STEM subjects make this dependence especially visible.

Large language models can succeed on many conventional physics problems while failing through subtle sign errors, inappropriate assumptions, invented data, or reasoning that is verbally coherent but physically inconsistent. Kortemeyer (2023) demonstrated that an AI agent could perform strongly across components of an introductory physics course, which directly challenges the evidential value of familiar take-home tasks. Yet this does not imply that physics understanding has become unassessable. It implies that assessment must make modeling choices, representations, experimental evidence, uncertainty, and oral or interactive reasoning more visible. Similar issues arise in mathematics, computing, chemistry, and engineering, where correctness of a final answer may reveal less about the learner than the quality of assumptions, validation, debugging, or explanation.

Opportunity and Risk are Properties of Designs, not Tools Alone

Educational AI is often discussed through lists of benefits and risks, yet the same capability can support or undermine learning depending on its use. Feedback may support revision or encourage dependence; automated scoring may reduce workload or obscure uncertainty; and personalization may improve access while unequal access to high-quality tools deepens inequity. The appropriate unit of analysis is therefore the sociotechnical assessment design: learner, teacher, task, model, data, policy, and decision. Bearman and Ajjawi (2023) describe this as learning to work with the “black box,” requiring neither blind trust nor blanket rejection. Model outputs should remain inspectable, contestable, and subordinate to educational purposes.

European and international guidance likewise emphasizes human oversight, transparency, privacy, and agency (European Commission, 2022; European Parliament & Council of the European Union, 2024; UNESCO, 2024). Kotsis and Stylos (2026) similarly argue that AI governance in STEM education must preserve teacher agency as assessment is redesigned. The implication for teacher preparation is clear. Demonstrating how to generate a rubric with AI develops functionality; requiring teachers to evaluate its alignment, identify inappropriate criteria, test it against student work, and determine when human moderation is necessary develops professional judgment. This distinction separates AI use from A-AIK.

CONCEPTUAL FOUNDATIONS FOR ASSESSMENT-AI KNOWLEDGE

Teacher Assessment Literacy as Situated Practice

Assessment literacy was once treated primarily as knowledge of measurement concepts and technical procedures. Contemporary accounts view it as situated professional practice shaped by beliefs, identity,

context, and institutional conditions. Xu and Brown (2016) conceptualize teacher assessment literacy as a dynamic interaction among a knowledge base, teachers' conceptions, institutional and sociocultural environments, and enacted practice. Pastore and Andrade (2019) similarly argues for a multidimensional model that includes conceptual, praxeological, and socio-emotional dimensions. These perspectives are important because AI-mediated assessment is not solved by adding a technical module to an existing knowledge list. Teachers must reconcile validity, pedagogy, policy, ethics, workload, student expectations, and the realities of their disciplines.

AI Literacy and the Limits of Generic Competence

AI literacy provides a necessary foundation. Ng et al. (2021) synthesize the construct around understanding, use, evaluation, and ethical engagement. UNESCO's (2024) AI Competency Framework for Teachers expands the field through 15 competencies in five dimensions: a human-centered mindset, ethics of AI, AI foundations and applications, AI pedagogy, and AI for professional learning. This framework rightly rejects purely technical preparation. It also recognizes developmental progression and the need to align competence with professional values.

However, generic competency frameworks operate at a broad level, whereas assessment requires purpose- and consequence-sensitive judgment. Transparency, acceptable error, and human oversight have different meanings in low-stakes formative feedback than in grading, placement, or misconduct decisions. In disciplinary contexts, this distinction becomes sharper. A teacher evaluating an AI-generated explanation of force and motion must determine not only whether it is factually correct, but whether it represents the intended physical construct, conceals misconceptions, substitutes for student reasoning, or deserves evidential weight alongside other work. This need for deeper preparation is reinforced by evidence that pre-service teachers may themselves hold misconceptions about artificial intelligence, suggesting that familiarity with AI tools or positive attitudes toward their use cannot be assumed to constitute adequate AI understanding (Kotsis, 2025). Generic AI literacy is therefore necessary but insufficient for defensible assessment decisions.

Teacher assessment literacy provides a second foundation by focusing on constructs, evidence, validity, scoring, feedback, fairness, and consequences. AI-mediated assessment nevertheless introduces new uncertainty because observed performance may combine student and machine contributions, while scores or feedback may come from probabilistic systems whose validity varies across tasks and populations. Teachers must therefore judge how AI participation changes the meaning and sufficiency of evidence itself.

TPACK offers a parallel conceptual lesson. Mishra and Koehler's (2006) framework shows that effective technology integration depends on relationships among technological, pedagogical, and content knowledge rather than isolated technical expertise. Generative AI intensifies this relational requirement, and recent teacher-development literature increasingly emphasizes ethical reasoning and critical evaluation (Brandão et al., 2024; Shenoy & Saarela, 2026). Assessment adds a further requirement: teachers must determine how technology, pedagogy, disciplinary content, and assessment conditions interact to support a defensible inference about learning.

The distinctiveness of A-AIK therefore lies in its object of professional judgment: the relationship among learner, task, AI participation, evidence, and educational decision. It does not replace AI literacy, teacher assessment literacy, or TPACK, but configures these knowledge resources around assessment. A-AIK is assessment-centered because competence is demonstrated through defensible claims about learning; performance-oriented because it becomes visible in design and interpretation; and accountability-oriented because consequential decisions must be explainable, documented, and open to review.

This accountability dimension is especially distinctive. Recognizing that an AI detector can produce false positives reflects AI literacy; deciding that its output is insufficient for a misconduct finding is an assessment

judgment. Seeking additional evidence, documenting the reasoning, and enabling review extend that judgment into professional accountability. A-AIK therefore captures the integration of design, interpretation, and defensible decision-making when AI alters the provenance, meaning, or consequences of assessment evidence.

The Three Capacities of A-AIK

A-AIK is defined as the integrated professional capacity to design, interpret, and justify assessment when AI participates in the production, analysis, or governance of evidence. The three verbs are deliberate. “Design” directs attention to the conditions under which evidence is elicited. “Interpret” emphasizes that machine outputs are claims requiring appraisal rather than facts. “Justify” refers to the professional responsibility to demonstrate how an assessment decision follows from the available evidence, acknowledge uncertainty and plausible alternative explanations, ensure procedural fairness, and make the reasoning open to appropriate challenge or review. Justificatory competence therefore concerns the accountability and defensibility of professional judgment, particularly when assessment decisions have substantial consequences for learners.

Design competence includes construct definition, task selection, specification of permitted AI assistance, evidence triangulation, rubric construction, and the orchestration of human and machine contributions. It also includes the capacity to choose whether AI should be excluded, bounded, disclosed, or deliberately integrated. Interpretive competence includes knowledge of model limitations, domain-dependent error, uncertainty, bias, data provenance, and the distinction between a proxy and the educational construct it is taken to represent. It requires comparison with other evidence and attention to the consequences of false-positive and false-negative judgments.

Justificatory competence includes the capacity to document the evidence and reasoning underlying a decision, distinguish machine-generated indicators from conclusions, communicate uncertainty proportionately, seek corroboration where necessary, and ensure that consequential judgments are subject to appropriate human review. It also includes procedural fairness, clear communication, professional accountability, and meaningful opportunities for explanation, reconsideration, or appeal. A teacher demonstrating justificatory competence should therefore be able to explain not only what decision was reached, but why the available evidence justifies that decision, what uncertainties remain, and how the decision may be reviewed if challenged.

The capacities are analytically distinguishable but practically interdependent. A poorly designed task cannot be rescued by sophisticated interpretation. A well-designed task can still produce injustice if detector scores are treated as proof. A carefully calibrated interpretation can lose legitimacy if the reasoning, data use, or route of appeal is invisible. The capacities are analytically distinguishable but operate recursively rather than as a linear sequence. Design decisions establish the conditions under which evidence is elicited and therefore shape its subsequent interpretation. Interpretive judgments determine the evidential weight assigned to human- and AI-generated information and influence the justification of educational decisions. Justification, review, disagreement, and evidence from enactment can, in turn, reveal weaknesses in the original task or interpretation and initiate a new cycle of design and appraisal. Although the cycle is presented in a principal direction, from design to interpretation to justification, each capacity can prompt reconsideration of the others.

Figure 1 presents A-AIK as a recursive professional-judgment cycle rather than a linear progression. Design competence establishes the conditions for producing interpretable evidence; interpretive competence evaluates the meaning, uncertainty, and evidential weight of that evidence; and justificatory competence supports transparent, fair, and reviewable decisions. The return connection from justification to design shows that decisions, appeals, implementation evidence, and unexpected consequences can require tasks, AI-use conditions, or interpretive criteria to be reconsidered. The additional reciprocal connections indicate that difficulties identified at any point may prompt re-examination of the other capacities.

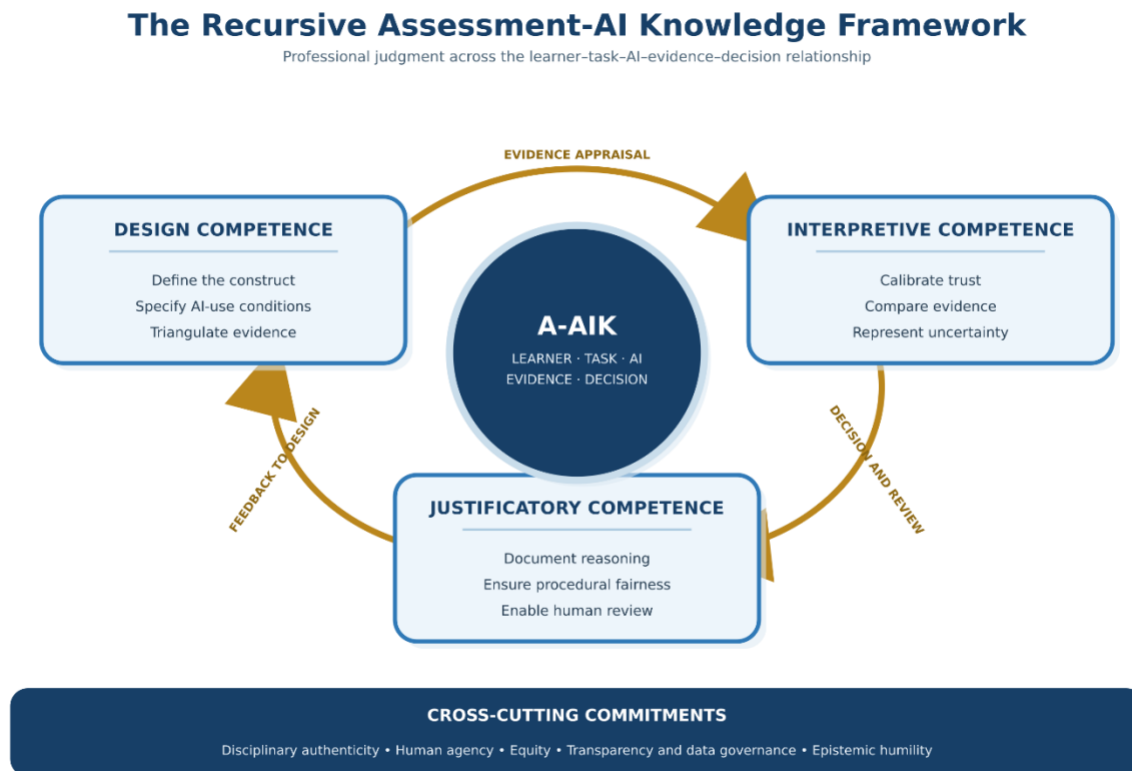


Figure 1. The recursive Assessment-AI Knowledge framework. Design, interpretive, and justificatory competence operate as an iterative cycle across the learner-task-AI-evidence-decision relationship. Design establishes the conditions under which evidence is elicited, interpretation determines its meaning and evidential weight, and justification makes consequential decisions transparent and reviewable. Feedback from decisions, implementation, disagreement, and review informs subsequent design and interpretation. Disciplinary authenticity, human agency, equity, transparency and data governance, and epistemic humility constrain the complete cycle.

Cross-Cutting Commitments

Five commitments operate across A-AIK. The first is disciplinary authenticity. The role of AI must be judged relative to what knowledgeable practice looks like in a domain. In some contexts, responsible AI orchestration may be an authentic capability; in others, independent fluency remains essential. The second is human agency. AI can advise, generate, or flag, but responsibility for educational decisions cannot be delegated to a system that does not understand the learner or bear the consequences. This emphasis is particularly important in physics and STEAM education, where AI should support rather than displace teachers' disciplinary expertise, professional judgment, and agency in making pedagogical and assessment decisions (Kotsis, 2026). The third is equity. Access, language, disability, cultural representation, and differential error must be considered before assessment conditions are declared fair. The fourth is transparency and data governance. Participants should know when AI is used, what data are processed, how outputs influence decisions, and what alternatives or appeals exist. The fifth is epistemic humility. Teachers need confidence to act, but also a disciplined awareness of uncertainty and the limitations of available evidence.

Table 1 distinguishes A-AIK from generic AI literacy without treating the two as competitors. AI literacy is a foundation; A-AIK is a situated professional enactment focused on assessment.

Table 1. From Generic AI literacy to Assessment-AI Knowledge

Dimension	Generic AI literacy emphasis	A-AIK assessment enactment
Purpose	Understand, use, and evaluate AI	Sustain defensible claims about learning
Unit of judgment	Tool, output, or interaction	Learner-task-AI-evidence-decision system
Design	Select tools and create content	Define constructs, assistance conditions, and evidence
Interpretation	Recognize error, bias, and limits	Calibrate evidential weight for a specific purpose and stakes
Ethics	Apply broad responsible-use principles	Examine equity, privacy, due process, and consequences in cases
Evidence of competence	Knowledge, confidence, or reported use	Quality of tasks, calibrated judgments, and justified decisions

WHAT THE EMERGING INTERVENTION LITERATURE SHOWS

Scope and Interpretive Approach

This study adopts a critical integrative review approach designed to bring together research from several intersecting areas: artificial intelligence in education, teacher education and professional development, assessment and feedback, academic integrity, and teacher AI literacy. The evidence base includes empirical interventions involving pre-service teachers, in-service teachers, teacher educators, and higher-education faculty, together with recent review studies and authoritative policy and competency frameworks. This breadth was considered necessary because research specifically examining teachers' assessment competence in AI-mediated environments remains emergent and is distributed across several related research traditions. At the same time, these professional populations should not be treated as interchangeable. Pre-service teachers are developing assessment judgment within supervised preparation contexts, in-service teachers operate within established school curricula and institutional policies, teacher educators work simultaneously with pedagogical and professional-learning responsibilities, and higher-education faculty often exercise greater autonomy over assessment design while working within different academic-integrity and governance structures.

Accordingly, evidence from one professional population is used in this review to illuminate potentially transferable principles rather than to imply direct equivalence across contexts. Findings from higher-education faculty development, for example, can inform teacher preparation when they concern general processes such as assessment redesign, calibration of trust in AI-generated outputs, or professional deliberation, but their applicability to school teaching must be considered in relation to differences in learner age, curricular prescription, assessment stakes, institutional authority, safeguarding responsibilities, and local policy. Similarly, interventions with pre-service teachers provide evidence about the development of emerging professional competence but cannot by themselves demonstrate how such competence is enacted under the workload, accountability, and organizational constraints of established classroom practice. The synthesis therefore attends not only to what an intervention appears to support, but also to the professional population and context in which that evidence was generated.

Sources were identified through structured iterative searching and backward and forward citation chaining. Searches were conducted in Scopus, Web of Science Core Collection, ERIC, and Google Scholar between January and July 2026, with the final search update completed on 31 July 2026. The search combined terms from three concept groups: artificial intelligence (“artificial intelligence” OR “generative AI” OR GenAI OR “large language model*”), teacher learning (“teacher education” OR “preservice teacher*” OR “pre-service teacher*” OR “in-service teacher*” OR “teacher professional development” OR “faculty development”), and assessment (“assessment” OR “assessment literacy” OR feedback OR scoring OR “academic integrity”). Additional combinations included “AI literacy,” “assessment redesign,” “automated feedback,” “AI-supported assessment,” and “teacher competence.” Search terms were adapted to the syntax and functionality of each database. Reference lists of relevant reviews and empirical studies were examined, and forward citation searching was used to identify additional publications.

The search process retrieved approximately 286 records before duplicate removal. After duplicates were removed, 231 titles and abstracts were screened for relevance, and 68 publications were examined in full text. Sources were included when they: (a) examined teacher, teacher-candidate, teacher-educator, or faculty learning in relation to AI; (b) addressed AI-supported assessment, feedback, assessment redesign, academic integrity, or professional AI competence; (c) reported an intervention, professional-development activity, teacher-education experience, or findings relevant to professional assessment judgment; or (d) provided a theoretical, ethical, policy, or competency framework directly relevant to assessment validity, AI literacy, teacher agency, fairness, or governance. Sources were excluded when they focused exclusively on students’ general attitudes toward AI, described technical model performance without implications for educational assessment, provided only promotional or non-scholarly commentary, were not available in English, or lacked sufficient information to determine their relevance. Earlier publications were retained only when they supplied foundational perspectives on assessment validity, teacher assessment literacy, TPACK, or effective professional development.

The final analytical corpus comprised 32 sources: five empirical intervention studies, five additional empirical studies addressing AI-mediated assessment or teacher engagement, 11 review or conceptual publications, eight foundational theoretical works, and three policy, ethical, or regulatory documents. These categories distinguished evidence about implemented professional learning from contextual, theoretical, and normative sources. They were not treated as equivalent forms of evidence. Empirical studies informed claims about intervention characteristics, teacher responses, and observable outcomes, whereas conceptual and policy sources informed the interpretation of validity, professional responsibility, equity, transparency, and governance.

An interpretive thematic synthesis was undertaken using an extraction framework covering professional population, educational context, intervention design, duration, AI role, targeted competence, learning activity, outcome measure, and evidence of demonstrated assessment judgment. One author conducted the initial extraction and thematic categorization, and a second author independently checked the inclusion decisions, source classifications, and interpretation of the principal themes. Differences were resolved through discussion and re-examination of the relevant publications. Because this was an interpretive integrative review rather than a formal systematic review, inter-rater reliability coefficients were not calculated.

Recurrent patterns were identified through repeated comparison across sources, with particular attention to the distinction between technical engagement or self-reported confidence and observable assessment performance. The three A-AIK capacities were not established beforehand as validated categories, nor were they statistically extracted from the studies. They were developed abductively through an iterative interaction between argument-based assessment theory and recurrent professional demands identified in the literature. Activities and judgments concerning task conditions, construct alignment, and the elicitation of meaningful evidence informed design competence. Demands concerning the appraisal of AI-generated scores, feedback,

classifications, uncertainty, and bias informed interpretive competence. Demands concerning documentation, explanation, fairness, accountability, and review informed justificatory competence. The five cross-cutting commitments were similarly developed by comparing recurring normative requirements across the empirical, theoretical, ethical, and policy sources.

This procedure does not make the review systematic or exhaustive. No preregistered protocol was used, and the heterogeneity of populations, settings, interventions, and outcome measures precluded quantitative synthesis. The retrieval and screening figures are approximate because the search process was iterative and was not originally maintained as a PRISMA-style systematic-review record. Nevertheless, the expanded account makes the scope, selection principles, composition, and interpretive procedure of the review substantially more transparent.

A Recurrent Pattern: Tool Engagement before Demonstrated Assessment Judgment

Early teacher-development responses understandably emphasized orientation and familiarization with generative AI through prompting, material generation, ethical discussion, and exploration of classroom applications. These activities are valuable, but they should be distinguished from the development of defensible assessment judgment.

Recent reviews illustrate this distinction. Brandão et al. (2024) emphasize the need for professional development suited to environments in which generative AI is widely available, while Shenoy and Saarela (2026), reviewing 20 studies, identify functional, cognitive, social, and reflective affordances. Across the literature considered here, a recurrent pattern is the concentration of reported outcomes on technical familiarity, attitudes, self-efficacy, perceived usefulness, acceptance, confidence, and intended use. Research with pre-service teachers similarly shows that their engagement with educational AI is strongly shaped by their perceptions, preparedness, and views about its potential integration into teaching, highlighting the importance of addressing these dimensions systematically within teacher education (Panagou et al., 2025). Such measures indicate readiness to engage with AI, but not necessarily the capacity to design valid assessments, evaluate machine-generated judgments, or justify consequential decisions.

This limitation is visible across individual interventions. Karakaya Özyer and Aydın (2026) involved 94 pre-service teachers in developing performance tasks with text-, image-, and video-generative tools, but the reported outcomes focus primarily on attitudes and acceptance rather than the validity of the resulting assessments or the quality of AI-use decisions. Meegan and Young (2025) similarly report gains in self-efficacy and AI literacy through scenario-based work, while Park et al. (2026) demonstrate the potential of microlearning for sustained faculty engagement. In each case, positive engagement should be distinguished from demonstrated assessment competence.

Other interventions move closer to performance-oriented evidence. Blonder et al. (2025) use generative AI to elicit and develop pre-service science teachers' pedagogical content knowledge, requiring participants to critique generated material and thereby engaging interpretive judgment. Awadallah Alkhouk and Khlaif (2024) focus on the redesign of authentic assessment tasks, providing a stronger basis for examining alignment, validity, equity, and AI-use conditions.

Taken together, the literature suggests a progression from orientation and acceptance toward critique, redesign, and situated professional judgment. Nevertheless, even authentic interventions are frequently evaluated through self-report rather than the quality of teachers' resulting tasks, interpretations, or decisions. Technical fluency, confidence, and acceptance should therefore be treated as enabling conditions rather than proxies for A-AIK. Demonstrating A-AIK requires observable evidence that teachers can preserve meaningful

learning evidence, calibrate trust in AI-generated claims, recognize uncertainty and differential risk, and justify decisions transparently and fairly.

Promising Design Features: Scenarios, Critique, Production, and Community

Several interventions move beyond tool demonstration by engaging participants in authentic professional tasks. Meegan and Young (2025) used scenario-based learning with pre-service teachers and reported gains in self-efficacy and AI literacy. The value of scenarios lies in situating AI decisions within realistic classroom dilemmas that require participants to identify missing evidence, weigh competing considerations, and defend a professional judgment.

Blonder et al. (2025) offer a particularly relevant science-education example, using generative AI to elicit and develop pre-service science teachers' pedagogical content knowledge. Treating AI-generated material as an object of critique requires participants to identify conceptual problems, anticipate learner difficulties, and revise explanations. This approach aligns closely with interpretive competence and disciplinary authenticity and is especially promising for STEM teacher education.

Assessment-redesign workshops provide a complementary approach. Awadallah Alkouk and Khlaif (2024) moved faculty from general concerns about academic integrity toward the redesign of authentic assessment tasks. Analyzing an existing assignment, clarifying its construct, testing its vulnerability, and developing alternatives connects professional learning directly with practice. Although "AI-resistant" assessment is appropriate when independent performance is intended, "AI-resilient" better captures the broader aim of maintaining interpretable evidence while allowing purposeful AI integration.

These approaches are consistent with established principles of effective professional development, including content focus, active learning, coherence, duration, and collective participation (Darling-Hammond et al., 2017; Desimone, 2009; Sims & Fletcher-Wood, 2021). Generative-AI interventions often achieve relevance and active engagement, but sustained implementation and evidence of effects on classroom practice remain limited. Short workshops may initiate development, but calibrated professional judgment requires repeated opportunities for application, critique, and revision.

Table 2 summarizes representative interventions discussed in this review and makes visible an important distinction between the professional activity included in an intervention and the evidence used to demonstrate learning. Several studies involve authentic or potentially assessment-relevant activities, yet their reported outcomes remain concentrated on attitudes, confidence, AI literacy, acceptance, or participation. Other studies move closer to observable professional performance by requiring participants to critique generated material or redesign actual assessment tasks. This distinction supports the argument that future research should assess not only whether teachers engage productively with generative AI, but also the quality of the assessment judgments and artifacts they produce. Taken together, the studies show a movement from orientation and acceptance toward critique, redesign, and situated professional judgment, but the evidential progression is incomplete. The strongest test of A-AIK would require participants not only to engage in authentic activities but also to produce assessable evidence of design quality, calibrated interpretation, enactment, and justified decision-making.

The Measurement Problem

One important limitation identified across several of the studies reviewed is a mismatch between the competence claimed and the outcome measured. Satisfaction indicates whether participants found a program acceptable. Intention to use indicates possible adoption. Self-efficacy indicates belief in capability.

Table 2. Representative Professional-Development Interventions and Their Relationship to A-AIK

Study	Participants-context	Intervention focus	Evidence-outcomes emphasized	Relevance to A-AIK
Karakaya-Özyer and Aydın (2026)	94 pre-service teachers	Development of performance tasks using text-, image-, and video-generative tools	Attitudes and acceptance	Primarily design-related exposure
Meegan and Young (2025)	Small cohorts of primary and post-primary pre-service teachers	Scenario-based generative-AI intervention	Self-efficacy and AI literacy	Potentially supports interpretive and justificatory reasoning through classroom dilemmas
Blonder et al. (2025)	Pre-service science teachers	Use of generative AI to uncover and develop pedagogical content knowledge	Participants' engagement with and critique of generated material	Strong relevance to interpretive competence and disciplinary authenticity
Awadallah Alkouk and Khlaif (2024)	Higher-education faculty	Workshops on redesigning assessment in response to generative AI	Redesign of authentic assessment tasks and practical workshop insights	Strong relevance to design competence and AI-resilient assessment
Park et al. (2026)	Higher-education faculty	Microlearning model for professional development in generative AI	Ongoing engagement with generative-AI professional learning	Relevant to sustained professional-development architecture

None demonstrates capability itself. If the learning objective is to evaluate AI-generated feedback, participants should evaluate a set of outputs of known quality and explain their decisions. If the objective is to design resilient assessment, they should produce and revise a task whose alignment, evidential coverage, equity, and AI-use conditions can be scored. If the objective is to handle an integrity allegation fairly, they should reason through a case file, identify the insufficiency of detector evidence, conduct an appropriate inquiry, and write a defensible decision.

Shenoy and Saarela (2026) similarly conclude that performance-based assessment of teacher AI competence is underdeveloped and that self-report dominates. The problem is magnified by novelty effects. Participants may report rapid gains because they have moved from unfamiliarity to successful initial use. Such gains are meaningful but fragile. Tool interfaces and model capabilities change, and confidence obtained through one successful interaction may not transfer to ambiguous, high-stakes decisions. Longitudinal assessment is needed to determine whether teachers retain judgment, adapt to model change, and alter classroom practice.

Interpretation and Feedback as a Neglected Professional Task

Many interventions teach teachers to generate feedback, questions, or rubrics; fewer require systematic evaluation of machine output. This imbalance is problematic because AI systems are persuasive precisely when they are wrong in plausible ways. Steiss et al. (2024), comparing human and ChatGPT feedback on student writing, found that human evaluators generally produced higher-quality feedback across most categories, although AI performed comparably in some criterion-focused respects. Corbin et al. (2025) argue that the value of feedback generated by generative AI is better understood through recognition: learners and

educators must be able to recognize what is useful, credible, and actionable. The professional capacity lies not in receiving feedback, but in judging and using it.

Justificatory Competence and Procedural Fairness: An Underdeveloped Area

Justificatory competence and procedural fairness receive comparatively limited attention in the teacher-preparation literature examined in this review. The rapid emergence of generative AI has intensified longstanding concerns about assessment security and academic integrity, while also complicating conventional assumptions about authorship, originality, and acceptable assistance (Dawson, 2021; Cotton et al., 2024; Perkins, 2023). These concerns are further complicated by the difficulty of distinguishing legitimate AI-supported academic work from forms of assistance that may undermine authorship, originality, or academic integrity, making transparent and context-sensitive criteria for acceptable AI use increasingly important (Cotton et al., 2024; Perkins, 2023). Academic-integrity discussions often emphasize policy awareness, prevention, or detection strategies, but less often require teachers to conduct a fair, evidence-based process when AI use is suspected. This is a serious omission because the availability of detection technologies can create an appearance of evidential certainty that their technical performance does not justify.

AI-detector results have uncertain accuracy and can be affected by editing, translation, genre, language background, and the characteristics of the text being evaluated. Weber-Wulff et al. (2023) found that available detection systems were neither sufficiently accurate nor reliable and could be easily undermined. A detector probability should therefore be treated as one potentially relevant but fallible indicator rather than as direct evidence of authorship, intention, or misconduct. Treating such an output as proof creates risks of false accusation, procedural unfairness, and inequitable consequences.

Teachers need preparation in evidential reasoning: what the detector result does and does not show; what corroborating evidence is relevant; how to invite a student's account without presuming guilt; how to compare drafts, process records, oral explanation, and prior work; and how to document the final judgment. They also need language for communicating uncertainty. "The detector indicates 82% AI" is not a conclusion; it is an output from an imperfect classifier under unspecified conditions. A defensible decision must explain the policy, evidence, interpretation, alternative explanations, and route of appeal.

Justificatory competence also applies beyond misconduct. If automated scoring contributes to a grade, the teacher must be able to explain the rubric, moderation, known limitations, and human review. If analytics label a learner "at risk," the teacher must distinguish prediction from destiny and avoid interventions that stigmatize. If an adaptive system changes the tasks presented to a student, the institution must be able to explain how the adaptation affects opportunity to learn. These are not legalistic additions to pedagogy. They are conditions of trustworthy assessment.

Subject Specificity and the STEM Imperative

The intervention literature remains uneven across subjects, and generic programs often treat content as interchangeable. This is inadequate because AI performance and assessment inference are content-dependent. In physics, a model may produce a correct numerical result using conceptually flawed reasoning. In mathematics, it may conceal an invalid transformation within fluent prose. In computing, generated code may run on common cases but fail under boundary conditions or introduce security problems. In laboratory science, a model can invent observations that resemble plausible data. Teachers require content knowledge to detect these failures and pedagogical content knowledge to understand how learners might be misled by them.

STEM education also provides strong opportunities for authentic redesign. Students can be assessed on model validation, uncertainty analysis, comparison between simulated and measured data, critique of AI-generated explanations, or the design of tests that expose failure. Oral defense, practical work, version history, and iterative modeling can make reasoning visible. Kotsis and Stylos (2026) emphasize that European governance of AI should be translated into assessment practices that preserve teacher agency and scientific responsibility. The implication for professional development is clear: generic ethical guidance should be paired with disciplinary cases, authentic student work, and subject-specific criteria.

A PRACTICE ARCHITECTURE FOR A-AIK

Design Competence: Preserving Meaningful Evidence

Design competence begins with construct clarity. Before deciding whether or how AI may participate, teachers must specify what the assessment is intended to reveal. Factual recall, conceptual explanation, experimental reasoning, collaborative problem solving, and responsible AI orchestration require different evidential conditions. The key question is therefore whether the resulting performance can still support the intended inference about learning.

Once the construct is explicit, teachers can distinguish AI use that substitutes for the assessed capability from use that legitimately supports performance. Assistance may be bounded by phase, function, or disclosure: for example, AI may be permitted for brainstorming, feedback, or checking, while students remain responsible for the reasoning and accuracy of the final work. Assessment instructions should therefore specify permitted and prohibited uses and require only documentation that is relevant to the assessment purpose.

Evidence should also be distributed across products, processes, and interactions. When authorship is uncertain, a polished final product alone provides limited evidence. Plans, revisions, laboratory records, selected AI interactions, oral questioning, demonstrations, or short transfer tasks can provide triangulation without turning assessment into surveillance.

Finally, resilient design requires testing tasks before deployment. Teachers can compare representative AI outputs with intended standards, identify possible construct substitution, and revise accordingly. Accessibility, unequal access, and the potential effects of disclosure requirements should also be considered. Design competence is therefore both technical and ethical.

Interpretive Competence: Calibrated Trust

Interpretive competence requires treating AI output as evidence to be evaluated rather than authority to be obeyed. Teachers should consider how a system was developed, what data and performance evidence support it, and whether those conditions resemble their own context. Where such information is incomplete, the evidential weight assigned to the output should be reduced.

Calibration also requires distinguishing reliability from validity. A system may produce consistent scores while measuring inappropriate features, agree with human graders overall while failing for particular groups, or provide useful formative feedback while remaining unsuitable for summative decisions. Evaluation must therefore be purpose-specific: trust depends on the learners, tasks, constructs, and consequences involved.

Uncertainty should also remain visible. Scores and risk labels can imply unwarranted precision, so thresholds should be treated as decision aids rather than definitive conclusions. The acceptable cost of error varies with stakes; a false positive in formative support differs fundamentally from one used in a misconduct allegation. Interpretive competence therefore combines statistical reasoning, disciplinary judgment, and consequence awareness.

Justificatory Competence: Accountable Judgment

Justificatory competence makes assessment reasoning transparent and reviewable. A defensible decision should allow another qualified person to reconstruct its purpose, criteria, evidence, use of AI, uncertainties, and human moderation. Documentation should be proportionate to stakes: formative feedback may require little formal record, whereas grades, certification, placement, or integrity decisions require greater transparency.

Procedural fairness requires that students know assessment conditions in advance, are informed when automated systems materially influence judgment, and have access to human review. Teachers should consider the student's account and alternative explanations, avoid treating proprietary probabilities as facts, communicate reasons clearly, and provide a meaningful route for reconsideration or appeal. A-AIK should not, however, transfer institutional obligations to individual teachers. Teachers are responsible for evaluating available evidence, recognizing uncertainty, seeking corroboration, documenting reasoning, and making fair professional judgments. They cannot reasonably be expected to validate proprietary systems, determine institutional data-governance arrangements, conduct technical audits, or establish formal appeal procedures.

Institutions must therefore provide shared protocols, moderation structures, approved-tool procedures, data-protection review, relevant expertise, and clear routes of human review. Responsible AI-mediated assessment depends on both professional competence and organizational capacity. Within this division of responsibility, teachers must also be prepared to conclude that available evidence is insufficient for a punitive or consequential decision. Such restraint is an expression of assessment literacy and professional accountability.

The Relationship Among the Capacities

The capacities can be illustrated through a single case. A teacher assigns a take-home explanation of a physics phenomenon and permits AI for language support but not conceptual generation. Design competence requires the task to make conceptual reasoning visible and the permitted assistance clear. A detector later flags one response. Interpretive competence requires the teacher to understand that the flag is uncertain, compare it with drafts, oral explanation, prior work, and conceptual coherence, and avoid equating polished language with generated reasoning. Justificatory competence requires a fair conversation, documented evidence, proportionate action, and a route of review. Weakness in any capacity undermines the whole assessment argument. **Table 3** translates the framework into observable practices and evidence that can be used in teacher education.

DESIGNING TEACHER EDUCATION AND PROFESSIONAL DEVELOPMENT

Principles for Program Design

An effective A-AIK program should be practice-centered, content-inflected, developmental, collaborative, and sustained. Practice-centered means that participants work on real assessment materials and decisions rather than only discussing hypothetical benefits and risks. Content-inflected means that generic principles are tested against disciplinary cases. Developmental means that complexity increases from low-stakes exploration to ambiguous consequential judgment. Collaborative means that teachers compare interpretations, moderate work, and establish shared norms. Sustained means that participants redesign, enact, collect evidence, and revise over time.

Table 3. A-AIK Capacities, Learning Tasks, and Performance Evidence

A-AIK capacity	Representative learning task	Observable performance evidence
Design	Redesign an existing assessment after testing it with AI	Clear construct; proportionate AI conditions; multiple evidence sources; aligned rubric; accessible alternative
Interpretive	Compare AI, expert, and rubric-based judgments on contrasting student work	Identifies failure modes; explains disagreement; adjusts trust to context and stakes; notes uncertainty and subgroup risk
Justificatory	Resolve an automated-score dispute or integrity case	Separates output from conclusion; seeks corroboration; documents reasoning; communicates a fair decision and review route
Integrated	Pilot, moderate, and revise an AI-mediated assessment	Coherent relationship among design, interpretation, enactment, student experience, and final decision

Such professional development should remain explicitly human-centred, ensuring that teachers develop the capacity to use AI critically while retaining professional agency, disciplinary judgment, and responsibility for pedagogical and assessment decisions (European Commission, 2022; UNESCO, 2024).

These principles respond directly to weaknesses in the emerging literature. Short workshops can provide orientation but should be embedded in a sequence. Tool demonstrations should lead to critique and calibration. Ethical discussion should be attached to decisions rather than treated as a final checklist. Self-efficacy should be supported through mastery experiences, but claims of competence should be based on performance. Communities of practice should do more than exchange prompts; they should moderate judgments, analyze failures, and build local evidence about tools and tasks.

A Four-Phase Professional-Learning Model

The proposed professional-learning model has four phases: calibrate, redesign, enact, and justify. These phases should not be interpreted as four additional competences or as an empirically established developmental sequence. The three A-AIK capacities describe the principal domains of professional judgment, whereas the four phases organize learning activities through which those capacities can be developed, applied, and integrated. The sequence is a theory- and synthesis-informed design hypothesis derived from the evidential logic of the A-AIK framework and from recurrent features of the interventions reviewed, including error analysis, assessment redesign, authentic application, collaborative deliberation, and justification of decisions. It therefore requires empirical evaluation rather than being treated as a validated model of teacher development.

Each phase places different emphasis on the three capacities. Calibration primarily develops interpretive competence by requiring teachers to examine model variability, error, bias, and uncertainty, while also informing subsequent design decisions. Redesign primarily develops design competence, supported by interpretive judgments about how AI may alter the meaning of student performance. Enact is not a fourth competence; it is the integrated application of design, interpretive, and justificatory competence in a real or simulated assessment context. Justify primarily develops justificatory competence, supported by interpretive competence because the quality of a decision depends on the appropriate appraisal and weighting of evidence. The phases can be implemented in initial teacher education, semester-long professional development, or school-based inquiry and may be revisited whenever tools, assessment purposes, or institutional conditions change.

Table 4. Relationship between the Four Professional-Learning Phases and the A-AIK Capacities

Professional-learning phase	Primary capacity	A-AIK Supporting capacities	Principal purpose
Calibrate	Interpretive competence	Design competence	Examine AI variability, limitations, bias, uncertainty, and appropriate evidential weight
Redesign	Design competence	Interpretive competence	Align constructs, tasks, AI-use conditions, criteria, and sources of evidence
Enact	Integrated application	All three capacities	Test the redesigned assessment and examine actual use, judgments, and consequences
Justify	Justificatory competence	Interpretive competence	Explain, document, communicate, and review consequential assessment decisions

The relationship between the professional-learning phases and the three A-AIK capacities is summarized in **Table 4**. The mapping identifies the capacity receiving primary emphasis in each phase while also showing the supporting capacities required for coherent professional judgment. It reinforces the distinction between the architecture of A-AIK and the proposed process for developing it: the capacities define what teachers need to demonstrate, whereas the phases organize learning experiences through which those capacities may be developed and integrated.

As **Table 4** indicates, the phases do not correspond to separate and mutually exclusive competences. Calibration informs subsequent design, redesign depends on an interpretation of how AI affects evidence, and enactment requires all three capacities to operate together. The justify phase makes justificatory competence most visible, but it also depends on appropriate interpretation of the available evidence. Enactment should therefore be understood as the integrative application of A-AIK rather than as a fourth capacity.

In the calibrate phase, participants learn how AI systems generate outputs and examine their domain-specific limitations. They compare machine and expert responses, investigate variability, identify hallucinations or bias, and discuss appropriate levels of trust. The core activity is not prompt optimization but error analysis. Participants should leave able to describe what an output supports, what it does not support, and what additional evidence is needed.

In the redesign phase, participants select an existing assessment and reconstruct its evidential argument. They clarify the construct, test the task with AI, specify permitted assistance, diversify evidence, revise criteria, and check equity and accessibility. Peer critique is essential because task vulnerabilities and hidden assumptions are difficult to see individually. The output is a complete assessment package rather than a list of ideas.

In the enact phase, participants pilot the redesigned assessment in a real or simulated setting. They collect student work, observe how AI is used, compare intended and actual behavior, moderate judgments, and document unexpected consequences. Initial teacher education can use microteaching, cases, or anonymized student artifacts when classroom access is limited. In-service programs should include classroom trials whenever possible.

In the justify phase, participants work with consequential cases. They explain an AI-assisted grade, respond to a disagreement between human and automated feedback, or conduct an integrity inquiry without overreliance on detection. They produce a decision record and communicate it to a student or family in plain language. Peers review whether the decision is valid, fair, transparent, and proportionate. This phase makes justificatory competence visible and connects assessment practice to professional responsibility.

Table 5. Four-Phase Professional-Learning Model for A-AIK

Phase	Central question	Main activity	Demonstrated outcome
Calibrate	What can this system support, and where does it fail?	Compare model and expert outputs across normal and edge cases	Purpose-specific account of reliability, validity, and uncertainty
Redesign	What evidence of learning will remain interpretable?	Reconstruct a current task, rubric, AI conditions, and evidence plan	Complete AI-resilient assessment package
Enact	How does the design work with learners in context?	Pilot, observe use, moderate work, and analyze consequences	Evidence-based revision grounded in student work and experience
Justify	Can the consequential decision be defended fairly?	Conduct a dispute or integrity simulation and communicate the decision	Transparent record, proportionate conclusion, and human review route

The specific purpose, principal activity, and intended performance outcome of each professional-learning phase are summarized in **Table 5**. Whereas **Table 4** maps the phases onto the three A-AIK capacities, **Table 5** describes how the phases may be operationalized within a teacher-education or professional-development program.

As shown in **Table 5**, the sequence begins by calibrating teachers' interpretations of AI-generated information, proceeds to the redesign and enactment of assessment, and concludes with the justification and review of consequential decisions. The sequence is intended to organize professional-learning experiences rather than prescribe a universal developmental pathway. Its effectiveness and ordering therefore require empirical evaluation across different educational and disciplinary contexts.

Learning Activities That Develop Judgment

The reviewed literature provides direct support for several activity formats relevant to A-AIK, although evidence of their effectiveness remains preliminary. Scenario-based learning has been used to develop pre-service teachers' AI literacy and self-efficacy by engaging them with professionally relevant situations (Meegan & Young, 2025). The structured critique of AI-generated materials has been used to expose and develop pre-service science teachers' pedagogical content knowledge (Blonder et al., 2025). Authentic assessment-redesign workshops have engaged educators in reconsidering task validity, permissible AI assistance, and assessment resilience (Awadallah Alkhouk & Khlaif, 2024). Sustained microlearning has also been used to support continuing faculty engagement with generative AI beyond a single workshop (Park et al., 2026). These studies support the educational plausibility of scenario analysis, structured critique, assessment redesign, and sustained professional interaction, although they do not yet establish their effectiveness for developing the complete A-AIK framework.

Other activities proposed here are derived primarily from the conceptual architecture of A-AIK and should be treated as design propositions requiring empirical evaluation. Contrast cases could require participants to distinguish between superficially similar forms of AI assistance that have different implications for the assessed construct. Calibration and error-analysis laboratories could expose participants to model outputs of known and varying quality, allowing them to examine error, instability, bias, and appropriate evidential weight. Moderation conferences could require independent judgments followed by evidence-based reconciliation. Integrity simulations could ask participants to evaluate incomplete or conflicting evidence without treating detector outputs as proof. Policy-translation activities could require participants to convert broad institutional principles into subject-specific assessment guidance. These proposed activities follow logically from the

design, interpretive, and justificatory capacities, but their effectiveness should not be assumed before they are tested.

Whether evidence-supported or framework-derived, the activities should make the participant's professional judgment observable. Participants should annotate decisions, compare AI outputs with disciplinary and contextual evidence, identify uncertainty, explain the evidential consequences of their choices, and justify revisions orally or in writing. Allowing participants simply to ask an AI system to critique its own output would permit the central learning task to be outsourced. AI may remain part of the activity, but the teacher's design, interpretive, and justificatory reasoning must constitute the primary evidence of learning.

Duration, Coherence, and Institutional Support

A-AIK cannot be developed through duration alone, but complex judgment requires repeated opportunities. A practical minimum design might combine an initial intensive workshop, four to six applied sessions, a classroom or simulated trial, and a final moderation and justification conference over one academic term. Microlearning can maintain engagement between sessions, while coaching or peer triads can support implementation. Programs should be aligned with local curriculum, assessment policy, available tools, and data-protection rules.

ASSESSING TEACHERS' A-AIK

The assessment approach proposed in this section is a framework-derived blueprint for future instrument development rather than an empirically validated A-AIK measure. Its components were developed by translating the conceptual architecture of A-AIK into observable forms of professional performance. The proposed tasks represent situations in which teachers must design evidence-generating conditions, interpret AI-mediated information, or justify consequential decisions. The performance indicators specify observable features of judgment within those situations, while the scoring descriptors express provisional hypotheses about qualitative differences in the coordination, consistency, transparency, and adaptability of that judgment.

This translation was guided by three sources. First, argument-based assessment theory requires alignment among the intended construct, the evidence elicited, the interpretation made, and the decision supported (Kane, 2013). Second, contemporary accounts of teacher assessment literacy emphasize situated professional judgment rather than the decontextualized recall of assessment terminology (Xu & Brown, 2016; Pastore & Andrade, 2019). Third, the reviewed AI literature identifies recurrent professional demands involving assessment redesign, appraisal of AI-generated outputs, uncertainty, fairness, and accountability. The resulting tasks and criteria should therefore be understood as theoretically grounded design propositions, not as evidence that A-AIK has already been established as a psychometric construct.

If A-AIK is to be assessed as a professional capacity, tasks should approximate authentic professional work. Self-report may remain useful for examining beliefs, anxiety, perceived preparedness, and confidence, but it should not serve as the primary evidence of competence. Performance assessment can provide stronger evidence of whether teachers apply relevant knowledge under realistic constraints and whether their judgments remain coherent across design, interpretation, and justification. Before the proposed approach is used for certification or other consequential purposes, its tasks, indicators, scoring criteria, and performance descriptors require expert review, cognitive interviewing, pilot testing across subjects and educational levels, rater training and moderation, reliability analysis, and evidence concerning construct representation, fairness, and the consequences of score use.

A portfolio approach can integrate several forms of evidence: an analysis of an existing assessment; a redesigned task and rubric; an AI-use statement; an audit of machine-generated feedback; samples of

moderation; a reflective account of enactment; and a written or recorded defense of a consequential decision. Each artifact should be scored with criteria that distinguish technical activity from educational judgment. For example, the presence of multiple prompt examples does not demonstrate interpretive competence unless the teacher analyzes variability, error, and implications.

Scenario-based assessments can complement portfolios. Participants can be given a case containing a student product, prompt history, detector result, teacher observations, policy excerpt, and student explanation. The task is to identify what is known, what remains uncertain, what evidence should be sought, and what decision is justified. Scoring should reward restraint as well as action. In some cases, the most competent decision is to withhold a punitive judgment because the evidence is insufficient.

Performance indicators should be observable, performance-based, and sensitive to context. For design competence, indicators include explicit construct definition, alignment between task and inference, proportionate AI conditions, evidence triangulation, accessible alternatives, and rubric criteria that prioritize reasoning and validation. For interpretive competence, indicators include identification of model limitations, appropriate comparison evidence, recognition of uncertainty, analysis of differential error, and adjustment of evidential weight to stakes. For justificatory competence, indicators include transparent documentation, procedural fairness, clear communication, human review, proportionate consequences, and a viable route of appeal.

As a provisional basis for instrument development, performance may be described using four descriptors: emerging, developing, proficient, and adaptive. These descriptors are not empirically established developmental stages, proficiency standards, or psychometrically validated score levels. Rather, they specify qualitatively different forms of observable performance that can guide the initial construction and evaluation of scenarios, portfolio tasks, scoring criteria, and rater exemplars.

Emerging performance is characterized by mechanical application of rules, uncritical acceptance of AI-generated outputs, or decisions that are weakly connected to the available evidence. Developing performance demonstrates awareness of AI limitations and relevant assessment principles but applies them inconsistently or without adequate coordination across design, interpretation, and justification. Proficient performance involves aligned, evidence-based, transparent, and context-sensitive judgment in familiar professional situations. Adaptive performance involves the principled application of A-AIK to unfamiliar or ambiguous cases, explicit management of uncertainty and competing considerations, support for collaborative professional judgment, and contributions to the improvement of institutional assessment practices.

The proposed descriptors should be treated as testable design hypotheses. Their number, ordering, boundaries, and scoring criteria require examination through expert review, cognitive interviewing, pilot assessment, rater-moderation studies, and empirical analyses of reliability, validity, and fairness. Until such evidence is available, they should not be used to make high-stakes claims about teachers' competence or to imply a universal sequence of professional development.

Reliability and validity of A-AIK assessments require careful study. Expert raters need shared exemplars and moderation. Scenarios must sample different subjects, stakes, AI roles, and equity issues. Consequences of certification or microcredentials should be proportionate to the maturity of the instrument. Most importantly, A-AIK assessment should not become another decontextualized test of terminology. Its validity depends on representing actual professional judgment.

The ultimate purpose of teacher preparation is improved student learning and fairer assessment. However, direct causal claims should be made cautiously. A-AIK may first influence task quality, clarity of AI-use expectations, feedback accuracy, student understanding of criteria, and procedural fairness. These proximal outcomes can be measured before attempting to link programs to achievement. Longer-term research can

examine whether students show stronger conceptual explanation, transfer, self-regulation, or capacity to critique AI.

INSTITUTIONAL, ETHICAL, AND POLICY IMPLICATIONS

Policy frameworks have established important principles, but teachers need operational guidance. UNESCO's (2024) human-centered framework and the European Commission's (2022) ethical guidelines emphasize agency, fairness, transparency, privacy, and critical engagement. The European AI Act introduces a risk-based regulatory environment with particular significance for educational systems used in access, evaluation, and placement (European Parliament & Council of the European Union, 2024). These principles should inform procurement, impact assessment, data governance, and human oversight.

Translation into assessment practice requires concrete questions. What educational claim will the system influence? What evidence supports its use for this population and subject? What data leave the institution? Can a learner opt for a human alternative? How are errors detected and corrected? Who has authority to override the output? What record is retained, and for how long? How is a decision appealed? These questions are particularly important because the educational use of AI raises substantial concerns regarding the collection, processing, storage, and protection of personal data. International and European guidance consequently emphasizes transparent governance, proportionate data use, meaningful human oversight, and continued human responsibility for decisions affecting learners (European Commission, 2022; European Parliament & Council of the European Union, 2024; UNESCO, 2024). A policy that merely tells teachers to "use AI responsibly" transfers institutional risk to individuals without providing the means to do so.

Purchasing an AI assessment system is not only a technical or financial decision. It embeds assumptions about learning, acceptable evidence, and teacher work. Procurement teams should therefore include educators with assessment expertise, subject specialists, data-protection personnel, accessibility expertise, and student representation. Vendors should provide documentation of intended use, training data where possible, validation evidence, subgroup performance, update procedures, audit access, and routes for correction.

AI-mediated assessment can create both accessibility benefits and new barriers. Translation, multimodal generation, text simplification, and assistive interaction may expand participation. At the same time, unequal access to premium tools, differences in connectivity, language bias, disability-related data, and cultural assumptions can distort evidence. Uniform rules may be inequitable if they ignore legitimate assistive use, while unrestricted use may advantage students with greater technical support.

Equity review should therefore be built into design and interpretation. Teachers should examine who benefits, who bears additional disclosure burdens, whose language is treated as anomalous, and whether alternative pathways are equivalent. They should avoid requiring disclosure of disability or personal circumstances beyond what is necessary. When uncertainty is high, consequences should be cautious and reviewable. Defensive competence is thus an equity practice as well as an accountability practice. Justificatory competence is thus an equity practice as well as an accountability practice.

AI is often justified through efficiency, but poorly governed adoption can increase workload through verification, policy conflict, and student appeals. Teachers may become responsible for correcting machine feedback, investigating false flags, and explaining systems they did not select. Institutions should count calibration, moderation, and redesign as professional work. Time saved in initial generation should not be claimed as institutional efficiency if hidden verification labor is transferred to teachers.

RESEARCH AGENDA

The next phase of research should move from adoption to competence. First, intervention studies should define learning outcomes in terms of observable A-AIK performance. Technical fluency, attitudes, and self-efficacy can remain secondary outcomes, but studies should include artifact quality, calibrated judgment, or decision defense. Measures should be reported with scoring procedures, exemplars, and reliability evidence.

Second, subject-specific studies are essential. Physics, mathematics, computing, languages, arts, and vocational education pose different constructs and error patterns. Research should examine how disciplinary expertise affects the recognition of AI errors and the design of authentic tasks. In STEM, particular attention should be given to conceptual reasoning, modeling, experimental evidence, uncertainty, and multimodal representations.

Third, studies should connect teacher learning to classroom practice and student experience. Do teachers actually change task design? Are AI-use expectations clearer? Does feedback become more accurate or dialogic? Are integrity disputes handled more fairly? Do students become better at evaluating AI output? Mixed-methods and design-based research can capture both outcomes and mechanisms.

Fourth, justificatory competence requires dedicated research. Valid scenarios and rubrics are needed for integrity inquiries, automated scoring disputes, predictive analytics, and communication of uncertainty. Research should examine whether training reduces overreliance on detectors, improves procedural fairness, and changes institutional decision patterns.

Finally, A-AIK itself requires cross-national empirical validation. Studies should examine whether design, interpretation, and justification are distinguishable, how they develop, and which capacities remain stable as models change.

LIMITATIONS OF THE REVIEW

This article is a critical integrative review and conceptual development, not a systematic review of intervention effectiveness. The literature was identified through iterative searching and citation chaining rather than a preregistered, exhaustive database protocol; consequently, relevant studies may have been missed. This methodological choice also limits claims about the prevalence of particular characteristics within the wider literature. For example, the observation that many interventions emphasize confidence, acceptance, self-efficacy, or intended use rather than demonstrated assessment performance should be interpreted as a recurrent pattern within the body of literature examined here, rather than as a quantified estimate of the proportion of all available studies exhibiting that pattern. Similarly, the relative prominence of particular intervention designs or professional populations may partly reflect the search and selection process.

The field is also expanding rapidly, and publications appearing after July 2026 may alter the characterization offered here. Intervention studies include heterogeneous populations, professional settings, disciplinary contexts, durations, intervention designs, and outcome measures. These differences limit direct comparison across studies and do not support pooled estimates of intervention effectiveness. The conclusions of the review should therefore be understood as an interpretive synthesis intended to identify theoretically significant patterns and research gaps, rather than as evidence of their statistical prevalence across the complete literature.

The A-AIK framework is theoretically grounded but not yet validated as a measurement model. The three capacities organize professional work coherently, but their boundaries may overlap and their enactment will vary by subject, educational level, institutional policy, and national regulation. The four-phase professional-learning model should therefore be treated as a synthesis-informed design hypothesis rather than an

established developmental sequence. It requires testing through implementation, expert review, teacher and student participation, and longitudinal evidence.

The article also concentrates on teachers' assessment work. Students' AI literacy, institutional governance, vendor responsibility, and broader political economy are discussed only insofar as they shape that work. A complete account of AI-mediated assessment requires all these perspectives. The narrower focus is intended to correct the relative neglect of teachers' evidential judgment, not to imply that responsibility rests with teachers alone.

CONCLUSION

Artificial intelligence does not remove the need for assessment judgment; it changes the conditions under which that judgment is exercised. When AI contributes to student work, feedback, scoring, classification, or prediction, educators must determine what the resulting information means as evidence, how much evidential weight it deserves, and how consequential decisions can remain fair and reviewable. The literature examined in this review supports the conclusion that generic familiarity with AI tools is insufficient for these assessment-specific responsibilities.

Within the reviewed corpus, recurrent patterns include an emphasis on technical engagement, confidence, acceptance, and intended use, alongside a smaller group of interventions involving structured critique, authentic assessment redesign, scenario-based learning, and sustained professional interaction. These findings should not be interpreted as estimates of prevalence across the entire field. They nevertheless indicate an important direction for research and professional development: greater attention to observable assessment judgment rather than reliance on self-reported preparedness alone.

On the basis of this synthesis and the theoretical foundations of assessment validity, teacher assessment literacy, AI literacy, and professional accountability, this article proposes Assessment-AI Knowledge as an assessment-specific configuration of professional knowledge. Its design, interpretive, and justificatory capacities describe how teachers may coordinate judgment across the learner–task–AI–evidence–decision relationship. The associated commitments to disciplinary authenticity, human agency, equity, transparency and data governance, and epistemic humility specify normative criteria that should constrain all three capacities. This architecture is theoretically grounded but has not yet been empirically validated as a measurement model.

The four professional-learning phases—calibrate, redesign, enact, and justify—should similarly be understood as a synthesis-informed design hypothesis rather than an established developmental sequence. The proposed learning activities, assessment tasks, performance indicators, and descriptors offer a structured agenda for program and instrument development, but their effectiveness, dimensionality, reliability, validity, and fairness require empirical examination across subjects, educational levels, professional populations, and national contexts. They should not yet be used to support high-stakes judgments about teachers.

The practical implication is therefore not the adoption of a finalized competency model but the need to prepare teachers to construct and maintain trustworthy evidential relationships under changing technological conditions. Future research should test whether the proposed framework helps teachers produce stronger assessment designs, interpret AI-mediated information more appropriately, and justify decisions more transparently and fairly. Preparing teachers for these responsibilities is becoming an important component of contemporary assessment literacy and professional formation.

Funding: No funding was received to assist with the preparation of this manuscript.

Declaration of interest: The authors declare no competing or conflicting interests.

Ethical statement: Research did not involve human participants.

AI statement: ChatGPT (OpenAI) was used to improve the English language and clarity of the manuscript. The author takes responsibility for the final text.

Data sharing statement: No data was collected for this study.

REFERENCES

- Awadallah Alkhouk, W., & Khlaif, Z. N. (2024). AI-resistant assessments in higher education: Practical insights from faculty training workshops. *Frontiers in Education*, 9, Article 1499495. <https://doi.org/10.3389/feduc.2024.1499495>
- 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(5), 1160–1173. <https://doi.org/10.1111/bjet.13337>
- Blonder, R., Feldman-Maggor, Y., & Rap, S. (2025). Are they ready to teach? Generative AI as a means to uncover pre-service science teachers' PCK and enhance their preparation program. *Journal of Science Education and Technology*, 34, 1301–1310. <https://doi.org/10.1007/s10956-024-10180-2>
- Bower, M., Torrington, J., Lai, J. W. M., Petocz, P., & Alfano, M. (2024). How should we change teaching and assessment in response to increasingly powerful generative artificial intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technologies*, 29(12), 15403–15439. <https://doi.org/10.1007/s10639-023-12405-0>
- Brandão, A., Pedro, L., & Zagalo, N. (2024). Teacher professional development for a future with generative artificial intelligence: An integrative literature review. *Digital Education Review*, 45, 151–157. <https://doi.org/10.1344/der.2024.45.151-157>
- Corbin, T., Tai, J., & Flenady, G. (2025). Understanding the place and value of GenAI feedback: A recognition-based framework. *Assessment & Evaluation in Higher Education*, 50(5), 718–731. <https://doi.org/10.1080/02602938.2025.2459641>
- 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>
- Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute. <https://doi.org/10.54300/122.311>
- Dawson, P. (2021). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge. <https://doi.org/10.4324/9780429324178>
- Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181–199. <https://doi.org/10.3102/0013189X08331140>
- European Commission. (2022). *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*. Publications Office of the European Union. <https://doi.org/10.2766/153756>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Karakaya Özyer, K., & Aydın, B. (2026). Generative artificial intelligence (GenAI) meets assessment: Experimental insights into teacher candidates' attitudes and acceptance. *International Journal of Education in Mathematics, Science and Technology*, 14(1), 21–45. <https://doi.org/10.46328/ijemst.5190>
- Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, 19(1), Article 010132. <https://doi.org/10.1103/PhysRevPhysEducRes.19.010132>
- Kotsis, K. T. (2025). Misconceptions about artificial intelligence from preservice teachers: A literature review. *EIKI Journal of Effective Teaching Methods*, 3(2), 199–208. <https://doi.org/10.59652/jetm.v3i2.565>
- Kotsis, K. T. (2026). Artificial intelligence and the professional work of physics teachers in STEAM education. *Journal of Artificial Intelligence and Technological Development*, 2(2), 245–253. [https://doi.org/10.59324/jaitd.2026.2\(2\).18](https://doi.org/10.59324/jaitd.2026.2(2).18)

- Kotsis, K. T., & Stylos, G. (2026). The AI Act and the future of STEM education in Europe: Rethinking pedagogy, assessment, and teacher agency. *Frontiers in Education*, 11, Article 1845045. <https://doi.org/10.3389/feduc.2026.1845045>
- Meegan, J., & Young, K. (2025). It's a tool not a crutch: A pilot generative AI intervention to enhance pre-service teachers' self-efficacy and AI literacy. *Technology, Knowledge and Learning*, 30, 1895–1916. <https://doi.org/10.1007/s10758-025-09875-1>
- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, Article 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Panagou, D., Stylos, G., & Kotsis, K. T. (2025). Exploring pre-service teachers' perspectives on integrating artificial intelligence in education. *Journal of Digital Educational Technology*, 5(2), Article ep2514. <https://doi.org/10.30935/jdet/17297>
- Park, E. E., Bodek, M. J., & Valcin, J. (2026). Prompted to explore: A microlearning model for faculty development in generative AI. *Journal of Teaching and Learning with Technology*, 14(1), 3–19. <https://doi.org/10.14434/jotlt.v14i1.41614>
- Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education*, 84, 128–138. <https://doi.org/10.1016/j.tate.2019.05.003>
- Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, 20(2), Article 07. <https://doi.org/10.53761/1.20.02.07>
- Shenoy, P., & Saarela, M. (2026). A critical review and actionable framework for integrating generative AI into teacher professional development. *Discover Education*, 5, Article 570. <https://doi.org/10.1007/s44217-026-01579-7>
- Sims, S., & Fletcher-Wood, H. (2021). Identifying the characteristics of effective teacher professional development: A critical review. *School Effectiveness and School Improvement*, 32(1), 47–63. <https://doi.org/10.1080/09243453.2020.1772841>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, Article 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3, Article 100075. <https://doi.org/10.1016/j.caeai.2022.100075>
- UNESCO. (2024). *AI competency framework for teachers*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000391104>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, Article 26. <https://doi.org/10.1007/s40979-023-00146-z>
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149–162. <https://doi.org/10.1016/j.tate.2016.05.010>