frontiers | Frontiers in Education

TYPE Review
PUBLISHED 23 October 2025
DOI 10.3389/feduc.2025.1710992

OPEN ACCESS

EDITED BY
Gloria Concepcion Tenorio-Sepulveda, Tecnológico Nacional de México/TES de Chalco, Mexico

REVIEWED BY
Katherine Muñoz, Corporación Educacional Naguilan, Chile
Maria del Carmen Gomez Pezuela Reyes, Autonomous Metropolitan University, Mexico

*CORRESPONDENCE
Andrés Chiappe
✉ andres.chiappe@unisabana.edu.co

RECEIVED 23 September 2025
ACCEPTED 10 October 2025
PUBLISHED 23 October 2025

CITATION
Fajardo-Ramos DC, Chiappe A and Mella-Norambuena J (2025) Human-in-the-loop assessment with AI: implications for teacher education in Ibero-American universities. *Front. Educ.* 10:1710992. doi: 10.3389/feduc.2025.1710992

Human-in-the-loop assessment with AI: implications for teacher education in Ibero-American universities

Diana Carolina Fajardo-Ramos[1], Andrés Chiappe[1]* and Javier Mella-Norambuena[2]

[1]Universidad de La Sabana, Chía, Colombia, [2]Universidad de Las Américas, Santiago, Chile

This scoping review examines how artificial intelligence (AI) reshapes assessment in Ibero-American higher education and specifies the teacher-training capacities and ethical safeguards needed for responsible adoption. Guided by PRISMA procedures and an eligibility scheme based on PPCDO (Population–Phenomenon–Context–Design–Outcomes), we searched Scopus and screened records (2015–2025; English/Spanish/Portuguese), yielding 76 peer-reviewed studies. Synthesis combined qualitative thematic analysis with quantitative descriptors and an exploratory correlational look at tool–outcome pairings. Rather than listing generic ICT, we propose a function-by-purpose taxonomy of assessment technologies that distinguishes pre-AI baselines from AI-specific mechanisms–generativity, adaptivity, and algorithmic feedback/analytics. Read through this lens, AI's value emerges when benefits are paired with conditions of use: explainability practices, data stewardship, audit trails, and clearly communicated assistance limits. The review translates these insights into a decision-oriented agenda for teacher education, specifying five competency clusters: (1) feedback literacy with AI (criterion-anchored prompting, sampling and audits, revision-based workflows); (2) rubric/item validation and traceability; (3) data interpretation and fairness; (4) integrity and transparency in AI-involved assessment; and (5) orchestration of platforms and moderation/double-marking when AI assists scoring. Exploratory correlations reinforce these priorities, signaling where training should concentrate. We conclude that Ibero-American systems are technically ready yet pedagogically under-specified: progress depends less on adding tools and more on professionalizing human-in-the-loop assessment within robust governance. The article offers a replicable taxonomy, actionable training targets, and a research agenda on enabling conditions for trustworthy, AI-enhanced assessment.

KEYWORDS

artificial intelligence in education, teacher training, ICT integration, educational technology, digital competence, pedagogical innovation

1 Introduction

The Fourth Industrial Revolution, characterized by the convergence of digital, physical, and biological technologies, is rapidly transforming all sectors of society, including education (Oke and Fernandes, 2020). This accelerated change, driven by the introduction of advanced technologies such as Artificial Intelligence (AI), Robotics, the Internet of Things, and Augmented Reality, is redefining the skills needed for the future (Miranda et al., 2024). As a result, the education sector faces a dual challenge: on the one hand, it must adapt its practices to prepare students for an increasingly digital and automated world, and on the other, it must leverage these new technologies to improve the teaching-learning process (Chituc, 2021). In this context of transformation, Information and Communication Technologies (ICT) are expected to effectively contribute to revolutionizing the way education is delivered and received, facilitating more personalized, accessible, and collaborative learning (Lawrence and Tar, 2018). For almost three decades now, online learning platforms, electronic devices, videoconferencing tools, and interactive educational applications have allowed educators to create more dynamic and effective learning environments, which has been documented in a growing process of research and practice, as shown in Figure 1.

Now, we are currently witnessing explosive and equally growing processes of implementation of disruptive technologies such as artificial intelligence in education, which is expected to achieve not only improvement as has already been achieved with other digital technologies, but true educational transformation processes.

2 Literature review

2.1 Key competencies for the AI era

In this constantly evolving digital landscape, it is crucial that both teachers and students develop a series of key competencies to face the challenges of the future. In this sense, González-Pérez and Ramírez-Montoya (2022) point out that digital competencies have acquired fundamental relevance in current education, transcending mere technical management. Thus, educators must possess the ability to critically evaluate digital resources, design technology-enriched learning experiences, and foster responsible digital citizenship among students. Complementarily, González-Salamanca et al. (2020) emphasize the importance of cultivating skills such as creativity, critical thinking, problem-solving, collaboration and teamwork, effective communication, and digital literacy. These competencies are essential to prepare individuals for Industry 4.0 and 21st-century society, where the integration of disruptive technologies and adaptability are crucial for personal and professional development (Miranda et al., 2024). In this sense, it is imperative that teacher training programs, both for future teachers and those already in practice, include specific components that prepare them to effectively integrate a diverse set of digital technologies, both new and already consolidated, into their pedagogical practices (Cabero Almenara and Martínez Gimeno, 2019; Marimon-Martí et al., 2022). This continuous and updated training is fundamental to equip all educators, regardless of their experience, with the necessary tools to navigate this new educational landscape and respond to the changing demands of the digital era.

2.2 Teacher training and digital technologies: an imperative for 21st century education

In this section we explicitly acknowledge the distinction between teacher training and teacher education, and we consider both for the purposes of this review. We use teacher education to denote the broader, programmatic formation of the profession–typically pre-service or postgraduate routes that develop pedagogical, ethical, disciplinary and research capacities. We use teacher training to refer to targeted, practice-oriented development–often short-cycle or in-service activities focused on skill acquisition and classroom enactment. Because digital transformation–and especially AI–cuts across these layers, we adopt an inclusive usage: studies are coded by setting (pre-service teacher education, in-service teacher training, or mixed), and findings are interpreted through this dual lens to capture impacts on both immediate instructional practice and the wider educational formation of teachers.

In the context of the digital revolution we are experiencing, teacher training in the use of digital technologies has become an unavoidable imperative for the education system. The ability of educators to effectively integrate ICT into their pedagogical practices not only improves the quality of teaching but also prepares students for an increasingly digitalized future. As Cisneros-Barahona et al. (2024) point out, ICT teacher training must go beyond mere development of technical skills, encompassing also the pedagogical understanding of how these technologies can enrich the teaching-learning process.

From this perspective, the importance of this training becomes even more evident when we consider the rapid advancement of emerging technologies. Teachers well-trained in digital technologies are better equipped to adapt to new tools and methodologies, allowing them to keep up with changing educational demands. Furthermore, as argued by both Mishra and Koehler (2006) and Gómez Sánchez et al. (2024) in the TPACK (Technological Pedagogical Content Knowledge) framework, it is considered key to develop in teachers the knowledge that effectively integrates technology into the teaching of specific content, which is evidently achieved through adequate teacher training processes.

Likewise, adequate teacher training involving the educational use of digital technologies empowers teachers to address educational inequalities, as in a world where access to information and learning opportunities are increasingly mediated by technology, ICT-competent educators can help close current digital divides, providing all students with the skills necessary to thrive in the knowledge society (Martín-Párraga et al., 2024). As highlighted by Arkorful et al. (2024), this training also fosters critical reflection on the use of technology, allowing teachers to cultivate responsible digital citizenship among their students.

Ultimately, investment in ICT teacher training is an investment in the future of education, as equipping educators with the tools and knowledge necessary to harness the potential of ICT

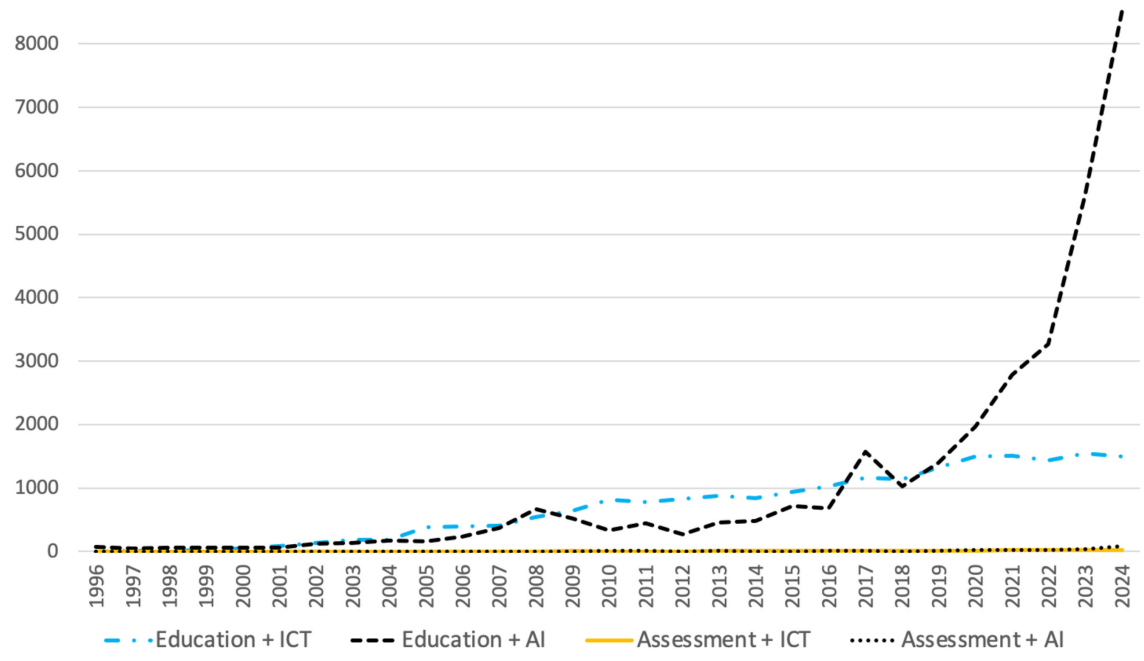

FIGURE 1
Research on education and assessment with ICT and artificial intelligence.

lays the foundation for a more flexible, inclusive, and relevant education system. This training not only benefits current teachers and students but also contributes to creating a culture of lifelong learning and adaptability, which are essential qualities for navigating the changing educational landscape of the 21st century.

2.3 Current challenges of assessment in the AI era

The rapid diffusion of AI across higher education has unsettled established assessment logics by introducing automated scoring, predictive analytics, and conversational agents into evaluative practice. While these affordances promise timelier feedback and personalization, they also raise core ethical concerns–transparency and explainability of model decisions, equity and non-discrimination, privacy and data protection, responsibility, and trust (Flores-Viva and García-Peñalvo, 2023). In parallel, the much-vaunted benefits of AI–automation of low-level grading and personalization–are contingent on the presence of robust digital ecosystems that render algorithms intelligible and data flows governable (Rojas and Chiappe, 2024).

At the task level, generative tools complicate the evidentiary status of traditional assignments: students can now produce competent surface texts with minimal cognitive engagement, which pressures assessment to pivot toward authentic performances that better signal individual understanding and originality. Experimental evidence shows that targeted instructional interventions can recalibrate perceived usefulness and ease-of-use of generative AI and reduce over-reliance in assessed tasks (Qian, 2025). Relatedly, students' attitudes, interest, usage, and literacy interact to shape AI self-efficacy–implications that matter for academic integrity and assessment design (Bewersdorff et al., 2025).

Finally, the field faces a normative tension as AI opens new pathways for personalized, real-time assessment and learner profiling, yet simultaneously elevates concerns over bias, opacity, and displacement of pedagogical judgment–conditions that demand strengthened educator competencies and ethics-by-design (Barrera Castro et al., 2024). In response, proposals gathered under Evaluation 4.0 advocate evidence-rich, non-linear logics (e.g., fuzzy systems) that value metacognition and self-regulation, thereby increasing the need for transparent criteria and informed teacher oversight of human–AI assessment processes.

2.4 Urgency of the review: teacher training in the AI-enhanced assessment era

Within this landscape, teacher training is the decisive lever to align AI capabilities with educational values in Ibero-American higher education. AI can scale feedback and continuous assessment only when educators are prepared to audit outputs, document assistance limits, and communicate rationales–requirements tightly coupled to data governance and algorithmic transparency (Wang, 2024). In practice, this entails cultivating algorithmic literacy, data stewardship, and pedagogical judgment for task redesign toward authentic evidence, competencies repeatedly flagged by high-impact reviews of AI-enabled personalization and platforms (Amoako et al., 2024).

Besides the above, urgency also stems from the changing nature of academic work. As generative systems trivialize routine text production, assessment must pivot from verifying recall

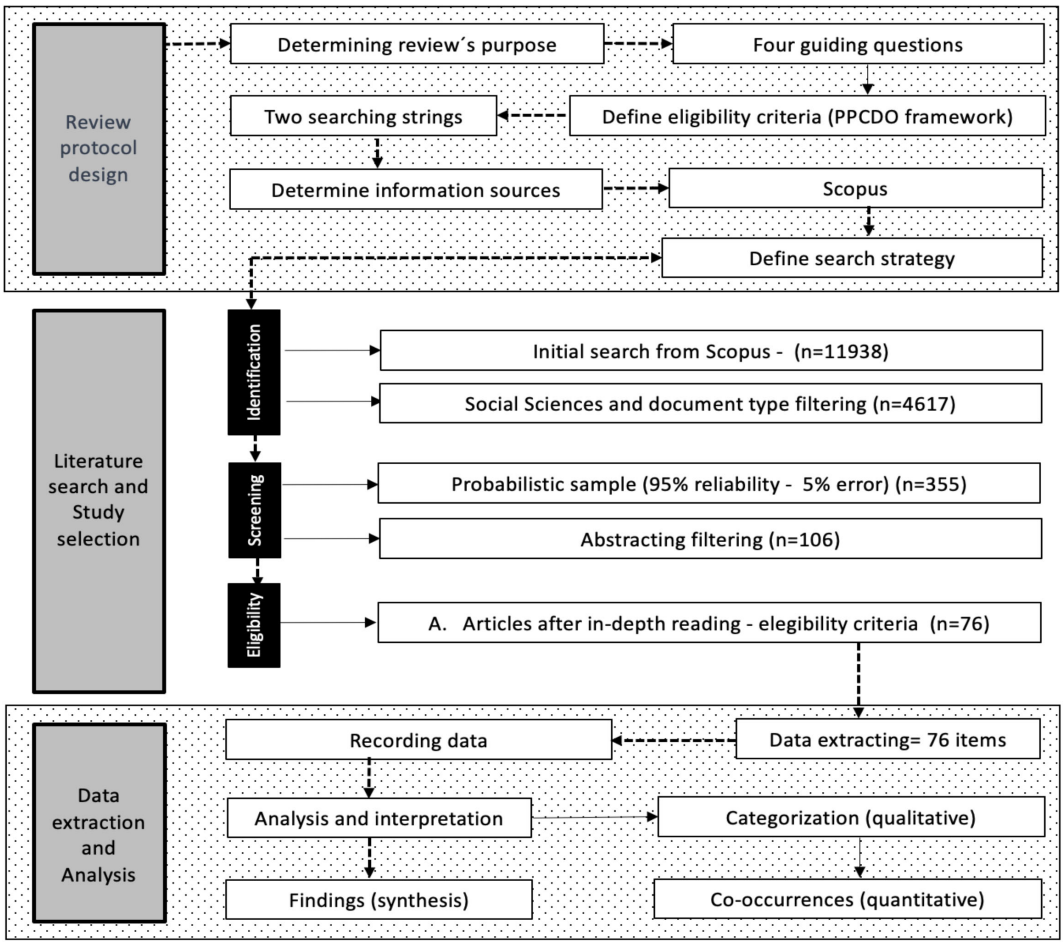

FIGURE 2
Review process.

to eliciting analysis, transfer, originality, and oral/interactive performances at scale. Evidence shows that active-learning interventions can dissuade inappropriate reliance on generative AI and reshape students' behavioral intentions–underscoring the need to prepare faculty to enact such designs (Bewersdorff et al., 2025; Van Niekerk et al., 2025).

Accordingly, this review is both timely and necessary: it consolidates dispersed evidence on digital and AI-enabled assessment in Ibero-American higher education and translates it into actionable requirements for teacher training. By offering (1) a taxonomy of assessment tools before and with AI, (2) an analysis of their pedagogical affordances and risks (e.g., bias, opacity, privacy, integrity), and (3) a set of competency targets and governance safeguards for educators (algorithmic literacy, audit trails, transparency-by-design, ethical use policies), the study aims to guide the redesign of initial and continuing teacher-education programs.

3 Methods

To address the objectives of this scope review, an approach proposed by Page et al. (2021) was integrated with some of the main processes of the PRISMA statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). Figure 2 shows, as a visual representation, the main phases, workflow, and key elements of the methodological process used in this scope review.

3.1 Review protocol design

3.1.1 Determining review's purpose

To align the review with the evidence and to make explicit the distinction between pre-AI ICT and AI-enabled technologies in assessment and teacher training, we structured the study around four guiding questions that jointly frame the search, coding, and synthesis. Specifically, we ask:

> RQ1 - What digital tools have been used for student assessment in Ibero-American higher education?

> RQ2 - What new elements–both positive and negative–does AI introduce into assessment that should be incorporated into teacher-training programs?

> RQ3 - What ethical principles, institutional policies, and governance safeguards are required for the responsible use of AI in assessment, and how should these be embedded in teacher education?

> RQ4 - Under what infrastructural and organizational conditions do the benefits of AI-enhanced assessment materialize, and what are the implications for the design, implementation, and evaluation of teacher training?

Together, these questions provide a coherent lens to interpret the corpus, attribute effects to AI where explicitly reported, and connect findings to practical conditions for responsible integration.

3.1.2 Eligibility criteria

To ensure transparent, reproducible screening aligned with our revised research questions, we operationalized eligibility using the PPCDO framework–Population, Phenomena, Context, Design, and Outcomes. PPCDO makes explicit who the evidence concerns (higher-education teacher–learners and assessment stakeholders), what technologies and uses are in scope (digital assessment tools, distinguishing pre-AI and AI-enhanced variants), where activity occurs (Ibero-American higher education and teacher-education programmes), how evidence is generated (empirical designs), and which results must be reported (tool taxonomy, affordances/risks, training implications, ethics/governance, and enabling conditions). This structure reduces selection ambiguity across diverse programmes and delivery modes and supports consistent downstream coding. In keeping with our AI lens, we pre-specified a conservative operational rule: tools are coded as AI-enabled only when primary studies explicitly report generative, predictive/recommender, or automated scoring/feedback features; otherwise, they are treated as pre-AI.

Population: We included studies involving pre-service teachers enrolled in university programmes and in-service university educators (lecturers, instructors, TAs) as learners or designers of assessment. Studies centered exclusively on K-12 populations were excluded unless the participants were pre-service teachers within higher-education settings. Research with higher-education students was eligible when the focal phenomenon was the use of digital tools for assessment and findings were interpretable for teacher training. Studies with teacher educators were eligible only when the instructional focus was the training of teachers.

Phenomena of interest: Eligible papers examined student assessment (diagnostic, formative, summative, authentic/performance) using digital tools, explicitly distinguishing:

– Pre-AI tools (e.g., LMS quizzes/e-exams, e-portfolios, rubric systems, plagiarism detection, clickers); and

– AI-enhanced tools (e.g., automated/granular feedback, adaptive testing, learning-analytics-supported grading, generative item/rubric creation, LLM-assisted evaluation, AI-supported proctoring), with AI functionality explicitly reported. We excluded works addressing general classroom technology not used for assessment, organizational IT unrelated to pedagogy, or tool descriptions without educational application.

Context: We accepted Ibero-American higher-education settings (undergraduate/postgraduate), including formal teacher-education programmes (initial or continuing), practicum contexts, and structured professional-development initiatives delivered face-to-face, blended, or online (including MOOCs and short courses) provided they were purposefully designed for teacher training or yielded directly transferable implications for teacher education in assessment. Informal or unstructured uses of technology with no training design were excluded.

Design: We included empirical studies–qualitative, quantitative, or mixed-methods–reporting original data, and systematic/scoping reviews centered on assessment technologies. We excluded conceptual essays, editorials, second-order reviews/meta-analyses without primary synthesis of assessment tools, and purely technical system papers without educational deployment. To maintain comparability, we limited the corpus to peer-reviewed journal articles (Education/Educational Technology/Learning Analytics and related Social Sciences) published 2015–2025 in English, Spanish, or Portuguese, with full text available. Duplicates were removed prior to screening.

Outcomes: To be eligible, studies had to report extractable outcomes for at least one of the following:

Type/taxonomy of digital assessment tools used.

Pedagogical affordances/effects (e.g., timeliness and quality of feedback, personalization/adaptivity, reliability/validity, authenticity, workload/time).

Risks/challenges (e.g., bias, opacity/explainability, privacy and data protection, academic integrity, displacement of pedagogical judgment).

Implications for teacher training (competencies, curricular elements, professional-learning designs).

Ethics and governance requirements (transparency, consent, fairness monitoring, audit trails, assistance limits).

Enabling conditions (digital ecosystems, data quality, technical–pedagogical support).

Papers without outcome evidence (e.g., proposals or descriptions without results) were excluded.

3.1.3 Information sources

We used Scopus as the sole primary source. This choice reflects a trade-off between coverage, metadata quality, and reproducibility aligned with our review questions. First, Scopus provides broad, multidisciplinary indexing across education, social sciences, psychology, and computing, which is essential for capturing teacher-training studies where ICT–and particularly AI-enabled tools–often intersect with adjacent fields (Chaparro-Martínez et al., 2016). Its consistent field tags (title/abstract/keywords), subject-area filters, and document-type controls allow us to target peer-reviewed journal articles within Social Sciences, reducing noise from non-comparable outputs (e.g., non-empirical pieces or technical demos with no training application).

Second, Scopus offers stable identifiers and exportable, structured metadata (DOIs, author keywords, affiliations, reference lists), which improves traceability and facilitates transparent PRISMA reporting (search dates, filters applied, and full record logs). Using a single database with uniform indexing policies also enhances reproducibility: another researcher can rerun the same Boolean strings and filters and obtain closely comparable corpora without resolving cross-database taxonomy conflicts.

Also, relying on one database intentionally avoids heterogeneity introduced by mixing sources with different coverage rules (e.g., varying inclusion of gray literature or conference proceedings) and incompatible thesauri. While multi-database searches may increase recall, they also raise deduplication and harmonization burdens that can obscure the very distinctions central to this study (pre-AI ICT vs. AI-enabled variants). Because our goal is to map tool families and attribute mechanisms rather than to estimate population parameters, we prioritized metadata consistency over maximal recall. To address single-database bias, we complemented the Scopus query with forward–backward citation chasing on sentinel studies surfaced during screening. This step helps recover influential papers that might be missed by indexing idiosyncrasies, without compromising the comparability afforded by a single primary database.

3.1.4 Search strategy

We designed the search to maximize recall of digital assessment uses in higher education–distinguishing pre-AI and AI-enhanced variants–while preserving precision for empirical, peer-reviewed work and teacher-training implications. Following PRISMA guidance, we proceeded in four steps: term development, query construction, filtering, and verification/calibration.

Term development: We compiled controlled expressions and synonyms for the population, phenomenon, and context from sentinel papers and indexing terms.

Population: "teacher education," "teacher training," "pre-service teacher*," "in-service teacher*," "faculty development," "teacher professional development."

Phenomenon (assessment): "assess*," "educational assessment," "student assessment," "evaluation," "grading," "rubric*," "e-portfolio*," "quiz*," "e-exam*," "proctor*," "formative feedback," "automated feedback," "learning analytics," "adaptive test*," "computerized adaptive testing," "automated scoring," "item generation."

AI variants (for auxiliary blocks): "artificial intelligence," "machine learning," "generative AI," "large language model*," "chatbot*," "LLM," "recommender."

Context/region: "higher education," "universit*," plus Ibero-American boosters (e.g., "Ibero-America*," "Latin America*," and country names such as "Argentina," "Brazil," "Chile," "Colombia," "Mexico," "Peru," "Spain," "Portugal," etc.).

To improve recall, we generated multilingual equivalents for core terms in Spanish (p. ej., formación docente, evaluación, retroalimentación, rúbrica, portafolio, examen en línea) and Portuguese (formação de professores, avaliação, feedback, rubrica, portfólio, exame online).

Query construction: Searches targeted titles, abstracts, and keywords using field codes (e.g., Scopus TITLE-ABS-KEY) with Boolean operators, quotation marks for phrases, truncation, and proximity where helpful. The core Scopus string was:

[TITLE-ABS-KEY ("teacher education" OR "teacher training" OR "pre-service teacher*" OR "in-service teacher*" OR "faculty development" OR "teacher professional development") AND TITLE-ABS-KEY (assess* OR "student assessment" OR "educational assessment" OR evaluation OR grading OR rubric* OR "e-portfolio*" OR quiz* OR "e-exam*" OR proctor* OR "formative feedback" OR "learning analytics" OR "adaptive test*" OR "computerized adaptive testing" OR "automated scoring" OR "item generation") AND TITLE-ABS-KEY ("higher education" OR universit*)].

Because many tool families include AI and non-AI variants, we did not require AI terms in the core query to avoid biasing retrieval toward explicitly AI-labeled studies; AI was identified at full-text coding when authors explicitly reported generative, predictive/recommender, or automated scoring/feedback features. For recall checks, we ran an auxiliary AI block ed with AND ("artificial intelligence" OR "machine learning" OR "generative AI" OR "large language model*" OR chatbot* OR LLM OR "automated feedback" OR "automated scoring" OR recommender) and a region booster ed with country/region terms. Equivalent Spanish/Portuguese query variants were executed with translated keywords.

Filtering: We searched Scopus, in which database-level limits aligned with scope: Subject areas = Education/Educational Research (plus Learning Analytics/Education-related Social Sciences where available); Document type = Article and Review (peer-reviewed journals); Years = 2015–2025; Languages = English, Spanish, Portuguese. No hard country filter was applied in the core runs (to avoid false negatives); the region booster ensured Ibero-American coverage.

Verification and calibration: To check sensitivity/precision trade-offs, we piloted the query on a sentinel set of known relevant articles and inspected a random slice of results to refine synonyms and suppress off-topic retrieval (e.g., organizational IT unrelated to pedagogy). We compared yields from the AI and region auxiliaries against the core query, confirming that unique inclusions were material for coverage but did not alter downstream coding logic.

3.2 Literature search and study selection

3.2.1 Identification

The initial search in Scopus yielded a total of 11938 results. To refine these results and ensure the relevance of the selected studies, a series of specific filters were applied. These included limiting the subject area to Social Sciences and selecting exclusively articles as the document type, reducing the number of studies to 4617.

3.2.2 Screening

To generate a manageable set of documents, a representative probabilistic sample was calculated with a reliability of 95% and an error of 5%, thereby reducing the number of items to 355 documents. At this stage, the review focused on the title and abstract of the studies, evaluating whether they met the established criteria. As a result of this selection, the total number of documents was reduced to 106, which were considered the most pertinent to the objectives of the review.

3.2.3 Eligibility

To increase the relevance of the selected studies, previously mentioned inclusion/exclusion criteria were applied through in-depth reading. The studies had to meet those requirements and also, presenting research results, topic relevance, and framing within an educational perspective. After this process, 76 studies were selected for data extraction and analysis.

3.3 Data extraction

In this phase, a detailed analysis of the 76 selected documents was carried out. The extraction process focused on identifying the specific ICT tools used in teacher training, the reported advantages of using these tools, and the disadvantages or limitations identified in their implementation. To facilitate this process and prepare the ground for a comprehensive analysis, a data extraction matrix was designed. This table was created in a spreadsheet and served as a central tool for organizing and structuring the information extracted from each study. Each row of the table represented an individual study, while the columns corresponded to the aforementioned information categories, in addition to bibliographic data such as authors, year of publication, and article title. This systematic approach to data extraction and tabulation not only provided a clear view of the reviewed literature but also facilitated the identification of patterns, trends, and relationships between different aspects of ICT use in teacher training.

3.4 Data analysis

Data analysis was performed combining qualitative and quantitative methods to obtain a comprehensive understanding of the results. First, a qualitative analysis was carried out, consisting of a detailed description of the ICT tools used in teacher training, as well as the advantages and disadvantages identified in their implementation, capturing the richness and complexity of the reported experiences. Second, a quantitative analysis was performed, identifying the frequencies of use of different ICT tools, as well as the recurrence of advantages and disadvantages mentioned in the selected articles, providing an overview of predominant trends. Finally, a correlation analysis was established between the use of tools and the observed advantages, looking for information patterns that could offer valuable insights into the effectiveness of different technological approaches in teacher training. Initially, ICT tools and their corresponding advantages were systematically identified through an exhaustive literature review. This phase included categorizing various technological tools, such as virtual platforms, simulators, and social networks, as well as enumerating associated benefits like improved time management, practical skill development, and collaboration promotion. Subsequently, a variable coding process was implemented, quantifying the presence and frequency of each identified advantage using a five-point Likert scale. This allowed for a numerical evaluation of the impact reported in the literature.

The Pearson correlation coefficient was then applied to measure the strength of the relationship between tools and their associated advantages. The interpretation of these correlations was based on standardized thresholds, ranging from very low (0.00–0.19) to very high (0.80–1.00). Finally, a statistical significance analysis ($p < 0.05$) was conducted to validate the robustness of the correlations found. This methodology not only established relationships between ICT tools and their benefits in teacher training but also identified the most promising technologies in this field.

4 Results

In reporting the findings, we distinguish between pre-AI ICT (e.g., LMS, repositories, forums, videoconferencing, non-intelligent simulators) and AI-enabled technologies (generative systems for content transformation, predictive/analytic engines for feedback and personalization, and automation for assessment and task design). This distinction clarifies what changes in assessment when AI intervenes: not merely the tool type, but the pedagogical mechanism (dynamic adaptation, generative scaffolding, agent-mediated tutoring) and the redistribution of instructional work (design, facilitation, assessment, and metacognitive regulation). We do not infer AI where primary studies do not state it; AI is flagged only when authors explicitly report machine-learning, natural-language generation, recommendation/prediction, or automated scoring/feedback.

For each category below, we first summarize patterns already observed with pre-AI ICT and then make explicit the AI-specific differentials where present–e.g., a shift from channel-based communication to agent-mediated interaction, from static resources to generative supports, and from periodic grading to continuous algorithmic feedback subject to human oversight. Where tool families include both AI and non-AI variants (e.g., simulators, feedback systems), we indicate when reported outcomes are attributable to AI features as described by the original studies and note the implementation conditions (explainability, data governance, and limits on assistance) under which those gains materialize in pre-service and in-service teacher education.

4.1 Results regarding the question: What digital tools have been used for student assessment in Ibero-American higher education?

To respond convincingly to RQ1, we organize the evidence by assessment function rather than by generic ICT families. Accordingly–and to make the comparison analytically useful–we mark, within each function, the pre-AI baseline and the AI-enhanced mechanism (generativity, adaptivity, and algorithmic feedback). This moves not only inventories tools but also reveals what educators must learn to design, implement, and moderate AI-aware assessment. Moreover, because verification-oriented instructional designs demonstrably reduce uncritical reliance on generative systems while preserving efficiency, human oversight is treated as constitutive of the tool landscape rather than as a *post hoc* safeguard. Table 1 synthesizes the landscape and, where available, includes shares from our corpus to indicate relative prevalence.

4.1.1 Assessment delivery and integrity (LMS, e-exams, proctoring)

Before AI, platforms largely scheduled, delivered, and graded tests. In our corpus, LMS-based assessment appears in 17.4% ($n = 45$) of references, with Moodle accounting for 35.6% ($n = 16$) of LMS mentions, and with email (30.7%; $n = 23$) and videoconferencing (25.3%; $n = 19$) frequently supporting submission, oral defenses, or live Q&A. With AI features, however,

TABLE 1 Digital tools for assessment in Ibero-American higher education (pre-AI vs. AI-enhanced).

Assessment function	Tools	Purpose	Prevalence in our corpus	Pre-AI → AI-enhanced mechanism	Implications for teacher training
Delivery and integrity (LMS/e-exam/proctoring)	LMS quizzes/e-exams (e.g., Moodle), proctoring suites	Summative/diagnostic at scale	LMS mentioned in 17.4% (n = 45); within LMS mentions Moodle appears in 35.6% (n = 16); email 30.7% (n = 23) and videoconference 25.3% (n = 19) support assessment logistics	Channel management → identity checks, reminders, anomaly flags	Policy-aware assessment design; proportionate integrity measures; transparent communication to students; documenting automated checks
Criteria and rubrics (authoring/co-creation)	Rubric tools; LLM-assisted rubric/item drafting	Formative/summative	Rubrics/portfolios 20.7% (n = 6) in assessment-specific subsector	Manual rubrics → generative drafts and criterion exemplars; risk of construct drift	"Rubric engineering"; alignment/coverage checks; decision logs and traceability
Automated feedback and tutoring	Writing-feedback engines; conversational agents	Formative	Interactive quizzes/tests 44.8% (n = 13); ∼70% of these incorporate interactivity	Batch comments → near real-time, criterion-referenced feedback; reliability varies	Feedback literacy with AI; prompt-and-audit cycles; revision-based workflows (human-in-the-loop)
Analytics and early warning	Learning-analytics dashboards; risk flags	Diagnostic/formative	AI tools 8.9% (n = 23) (cross-cutting)	Descriptive dashboards → prediction/recommendation	Data interpretation; consent/minimization; fairness checks; clear escalation paths
Portfolios, peer and self-assessment	e-portfolios; peer-review platforms	Authentic/longitudinal	20.7% (n = 6) (with rubrics)	Manual review → assisted clustering, rubric suggestions	Calibrated moderation/double-marking with AI; criteria alignment; workload planning
Authentic and multimodal evidence	Video/oral tasks, code notebooks, audio artifacts	Performance/competency	Video resources appear in 26.9% (n = 18) of educational-resource mentions; VR/AR remains 0.8% (n = 2); programming tools 2.7% (n = 7)	Human scoring → AI-assisted transcription/scoring; construct shift risks	Task design for authenticity; override/escalation rules; multimodal assessment literacy; careful interpretation of auto-scores

Coding rule. A tool is labeled AI-enhanced only when primary studies explicitly report generation, prediction/recommendation, or automated scoring/feedback.

10.3389/feduc.2025.1710992

systems add chatbot Q&A, recommendations, and anomaly detection, which shift routine assistance and parts of integrity checking to the platform. Consequently–and this becomes central for teacher education–the task evolves from merely "posting a test" to setting transparent criteria, explaining automated checks, and documenting decisions. Verification-oriented routines are key because structured human oversight has been shown to curb naïve trust in automation during assessed tasks.

Qualitatively, the evidence portrays a field still governed by institutional risk management logics rather than pedagogical transformation. Proctoring and secure e-exams are framed as safeguards to preserve legacy assessment formats at scale, while LMS-based delivery standardizes processes but rarely reimagines the epistemic aims of assessment. Educators' narratives reveal ambivalence: confidence in the logistical reliability of these systems coexists with doubts about their fairness, cultural sensitivity, and potential to exacerbate student anxiety. In practice, integrity tools are deployed as compliance infrastructures, which stabilize assessment operations yet may inadvertently narrow opportunities for dialogic evaluation, authentic demonstration of competence, and equitable accommodation of diverse learners.

4.1.2 Criteria, rubrics, and generative co-design

Rubric and exemplar creation has long been manual; nevertheless, with generative tools, instructors can prototype prompts, obtain draft rubrics/items, and iterate to criteria-aligned versions. While this yields speed and variety, it also raises the risk of construct drift; therefore, competencies shift toward rubric engineering (coverage, difficulty, alignment) and traceability (what was generated, how it was edited, and why). Notably, artifacts tied to rubrics and portfolios constitute 20.7% (n = 6) of assessment-specific mentions, which underscores both their salience and the need to prepare educators to validate AI-assisted drafts rather than accept them at face value. Moreover, interventions that require learners to check sources, detect hallucinations, and justify edits foster the critical stance we expect from assessors.

Across studies, rubrics emerge as boundary objects where human judgment and AI affordances intersect. However, the qualitative accounts suggest that rubric work remains predominantly top-down: generative tools are enlisted to draft criteria and descriptors, but iterative co-design with students–crucial for transparency and shared understanding–appears underdeveloped. Educators welcome efficiency gains (e.g., faster criterion phrasing, exemplars at multiple levels) yet express concern about construct drift and misalignment with disciplinary norms when AI proposes decontextualized language. These tensions indicate that generative co-design adds value when anchored in local assessment cultures–through calibration sessions, exemplification with real student work, and explicit negotiation of meaning–rather than treated as a one-shot content generator.

4.1.3 Automated formative feedback and tutoring

Whereas pre-AI feedback typically arrived in batches, AI introduces near real-time, criterion-referenced feedback and stepwise suggestions, thereby enabling denser formative cycles. In our data slice, interactive quizzes/tests represent 44.8% (n = 13) of the assessment tools cataloged in this subsection, and ≈70% of those implementations include interactive elements that align well with iterative practice. Yet timeliness without reliability is fragile; consequently, programs should build feedback literacy with AI–designing criterion-anchored prompts, auditing samples for consistency/bias, and requiring revision-based workflows so that students demonstrate improvement rather than merely accepting outputs. Experimental interventions that embed verification and reflection consistently reduce over-reliance while preserving efficiency.

The qualitative pattern points to a "use–trust gap." Educators acknowledge the motivational and pacing benefits of rapid, tailored feedback, yet they hesitate to delegate judgment in areas requiring nuanced interpretation (argumentation, interdisciplinarity, ethical reasoning). Students appreciate immediacy but question accuracy and relevance when feedback is generic or insufficiently grounded in task criteria. Where human moderation and feedback triage are embedded–e.g., teachers auditing samples, editing AI comments, and closing loops with brief in-class clarifications–participants describe higher perceived usefulness and fairness. These accounts suggest that automated tutoring is most pedagogically credible when it functions as a scaffold within a human-facilitated feedback ecology, not as a substitute for expert sense-making.

4.1.4 Learning analytics and early-warning models

Analytics dashboards have moved from descriptive reporting to prediction/recommendation. Although AI-labeled tools account for 8.9% (n = 23) overall in our corpus, their impact depends less on raw prevalence and more on ecosystem quality–namely, data pipelines, documentation, and clear role definitions. Hence, teacher training should include data interpretation, consent/minimization, and fairness monitoring; likewise, because interest and literacy shape willingness to engage, programs should blend hands-on experience with explicit critical reflection to build self-efficacy for responsible use.

Narratives around analytics emphasize potential for timely support but also surface concerns about reductionism, labeling, and student agency. In this sense, qualitative reports from instructors show value in triangulating dashboards with situated knowledge (attendance patterns, clinical placement feedback, students' self-reports), which tempers the risk of over-reacting to noisy signals. Students, in turn, respond better when indicators are explained, actionable, and framed as growth-oriented rather than predictive verdicts. A recurrent theme is that analytics catalyze meaningful intervention only where relational practices and clear referral pathways exist; without these, alerts remain informational artifacts that can entrench deficit framings rather than mobilize supportive action.

4.1.5 Portfolios, peer/self-assessment, and authentic/multimodal evidence

Portfolios and authentic tasks provide rich, longitudinal evidence; nevertheless, they are labor-intensive to scale. In our corpus, the rubrics/portfolios subsector (20.7%; n = 6) confirms their relevance; yet VR/AR remains marginal (0.8%; n = 2) and programming tools 2.7% (n = 7) –signaling that AI-assisted multimodal assessment is still emergent. AI can facilitate transcription, clustering, or pre-marking, but these supports bring reliability questions to the fore; hence, moderation and double

marking become central competencies. Importantly, the shift toward multimodal artifacts aligns with evidence that alternative modalities can boost engagement and achievement–particularly when well scaffolded–which in turn justifies training educators to design, judge, and communicate expectations for non-textual evidence. Furthermore, because attitudes, usage, and interest interact to build AI self-efficacy, programs should incorporate low-stakes, supervised practice that normalizes verification and reflection across these modalities.

The qualitative corpus highlights portfolios and peer/self-assessment as loci of student voice and epistemic agency, especially when AI is positioned to help with organization, reflection prompts, and accessibility (e.g., transcriptions, translation, multimodal curation). Still, educators caution that AI-supported curation can drift toward performative showcase if reflective depth and criterion-referenced judgment are not foregrounded. Moreover, effective implementations pair AI assistance with structured dialogic moments–studio critiques, calibration with exemplars, and guided metacognitive prompts–so that multimodal artifacts become evidence of learning processes rather than mere products. In these conditions, participants report stronger authenticity, identity expression, and transfer across contexts.

4.2 Results regarding question 2: What new elements–both positive and negative–does AI introduce into assessment that should be incorporated into teacher-training programs?

Framed through a comparative lens, our synthesis distinguishes enduring patterns from pre-AI ICT (e.g., access, communication, practice, coordination) and the AI-specific differentials that matter for assessment–namely generativity, adaptivity, and algorithmic feedback/analytics. Where primary studies explicitly report AI features, we interpret benefits such as faster iteration, targeted guidance, and semi-automated formative feedback as AI-driven, and we analyze them together with the institutional and pedagogical conditions under which they materialize (explainability requirements, data-governance protocols, principled assistance limits). This clarifies not only what appears to work, but also when and why it works in AI-mediated assessment. Consistent with evidence that instructional designs requiring students to interrogate AI outputs temper naïve adoption without sacrificing efficiency, these insights translate directly into teacher-education assessment tasks (Van Niekerk et al., 2025). Regarding this, Table 2 summarizes the AI-introduced elements for assessment, the core affordances/concerns of each, and the corresponding implications for teacher training.

In parallel with these elements, large cross-national evidence shows that attitudes and actual use of AI strongly predict interest, which–together with baseline AI literacy–builds self-efficacy; hence, training designs should couple authentic AI use with explicit verification routines (Bewersdorff et al., 2025).

4.2.1 Innovation in assessment design, feedback, and communication

Before AI, "innovation" in assessment largely meant richer resource integration and activity redesign, but now, with AI, it shifts toward co-design with generative systems: rapid prototyping of tasks, exemplars, and rubrics; scaffolded drafting/redrafting; and agent-mediated dialogue that elicits reasoning for formative purposes. Reported gains–time saved in preparing assessment materials, finer differentiation of feedback, and richer dialogic interaction–are maximized when programs embed vetting of AI outputs into the assessment workflow itself (e.g., mandatory source verification, justification prompts, and reflection on model limitations). Thus, empirical interventions show that when learners must test citations, check for hallucinations, and judge relevance/quality, they develop critical assessment literacy without discarding useful efficiencies; in effect, oversight becomes the learning target (Van Niekerk et al., 2025).

For teacher training, the implication is straightforward: formative modules should combine AI-supported production (draft rubrics or feedback suggestions) with structured auditing, including explicit criteria and decision logs. In this configuration, AI catalyzes iterative assessment design while the human assessor retains epistemic authority. Studies implementing such "inspect-and-explain" routines report reduced uncritical reliance on AI and measurable shifts in behavioral intentions–precisely the type of

TABLE 2 Artificial intelligence (AI)-introduced elements in assessment and implications for teacher training.

AI element in assessment	Affordance/concern	Training implication
Automated, criterion-referenced feedback (LLMs, analytic engines)	Timely, scalable, personalized feedback; risk of opacity/hallucination	Feedback literacy with AI; prompt-and-audit cycles; documentation of assistance limits
Adaptive testing and personalization	Finer measurement across ability ranges; item exposure/fairness issues	Practitioner basics of IRT/CAT; bias monitoring; fairness checks
Generative item/rubric drafting	Rapid prototyping and variant generation; content-validity and alignment risks	Rubric engineering and item vetting; coverage and difficulty balance
Learning analytics and early-warning models	Diagnostic insight; profiling and over-reach risks	Data interpretation; consent/minimization; purpose limitation and communication
Multimodal evidence and AI-assisted scoring	Richer evidence and scale; construct shift and reliability concerns	Authentic task design; moderation/double-marking; override/escalation rules
AI-supported proctoring/integrity tools	Deterrence/verification; privacy and due-process concerns	Integrity policy literacy; proportionate use; appeals and human review

outcome sought in professional preparation for assessment (Abuzar et al., 2025).

Qualitatively, innovation appears less as the introduction of novel tools and more as the orchestration of practices that reconfigure roles, timelines, and evidence. Cases perceived as most generative share three traits: (a) task redesign toward iterative production with micro-feedback cycles, (b) transparent communication of what AI may and may not assist, and (c) explicit alignment between criteria and the forms of evidence students can credibly produce with AI support. Where these traits are missing, the same tools yield incrementalism–faster grading, nicer rubrics–without altering how learners engage with standards, reflection, or audience. Thus, innovation hinges on communicative clarity and design intentionality, not on technical novelty alone.

4.2.2 Adoption dynamics: acceptance, interest, and self-efficacy in AI-supported assessment

Where pre-AI ICT adoption often hinged on usability and perceived usefulness, AI introduces additional drivers and frictions: perceived explainability of automated judgments, trust in model behavior, and the redistribution of effort from content production to oversight. A large multi-country study finds that AI use and positive attitudes significantly predict interest, which–together with AI literacy–enhances AI self-efficacy; hence, teacher-education should integrate hands-on assessment tasks with AI to cultivate attitudes and interest while normalizing verification (Bewersdorff et al., 2025). Moreover, the same work identifies meaningful learner profiles ("AI Advocates," "Cautious Critics," "Pragmatic Observers"), implying differentiated supports: some teacher-learners need stronger scaffolds for risk appraisal and ethics, whereas others require structured opportunities to translate enthusiasm into disciplined assessment practice (Trajkovski and Hayes, 2025).

Educators' and students' narratives converge on a pathway in which perceived relevance and low entry-barriers nurture early use, which in turn fosters confidence and more ambitious applications. Yet acceptance is fragile: it stalls when institutional signals are mixed (e.g., permissive rhetoric coupled with punitive enforcement), when supports are generic rather than discipline-specific, or when exemplars remain abstract. Reports from successful sites describe local champions, brief targeted workshops using authentic tasks, and quick-win templates that lower cognitive load. In short, self-efficacy grows where adoption is social and situated–anchored in credible peers, practical exemplars, and feedback on first attempts–rather than mandated or left to individual improvisation.

4.2.3 Emerging risks and workload reconfiguration

Although familiar ICT constraints persist, AI raises distinct assessment challenges. First, data governance is central: consent, retention, and student privacy must be addressed when prompts or outputs include sensitive assessment data. Second, model behavior is variable; updates can shift output quality and hallucinations remain a known failure mode–hence the need for systematic verification and transparency about model/version use. Third, equity concerns follow from the attitudinal and efficacy gradients noted above: without low-stakes guided interaction to build self-efficacy, AI-enabled assessment may amplify disparities; conversely, early authentic use embedded in a balanced curriculum (benefits and risks) can raise interest and, through it, self-efficacy (Stephenson and Harvey, 2022).

The qualitative evidence reframes "efficiency" as a redistribution rather than a net reduction of work: time saved in marking can be reallocated to rubric refinement, audit sampling, student conferencing, and governance activities (consent, disclosures, documentation). Participants also surface ethical and relational risks–opacity, bias, learned helplessness–that require new competencies (explainability, boundary-setting) and institutional safeguards (clear assistance limits, appeal routes, data stewardship). Where these are absent, educators report precautionary underuse or covert practices that undermine coherence. The message is not that risks outweigh benefits, but that benefits materialize when workload models, training, and policy evolve to recognize the distinct labor of human-in-the-loop assessment.

4.3 Correlational analysis

To remain coherent with the preceding results, we reinterpreted the correlational evidence through the assessment functions mapped before. Accordingly, we report associations between tool families and assessment-relevant advantages that inform teacher training. As before, coefficients (r) indicate association strength and p-values its statistical reliability. Although correlations do not imply causation, they help prioritize competency targets for AI-aware assessment. In line with intervention studies that show the value of human oversight and verification during assessed tasks, we read the strongest associations as signals for where training should concentrate orchestration and audit skills. Regarding the above, Table 3 presents the matrix with the original coefficients, now grouped by assessment function.

First, the very high association for evaluation platforms and continuous feedback ($r = 0.83$; $p = 0.002$) coheres with Subsection "4.1.3 Automated formative feedback and tutoring" and reinforces the need to prepare educators for feedback literacy with AI–that is, designing criterion-anchored prompts, auditing samples, and running revision-based workflows rather than accepting outputs at face value. Second, strong links for virtual platforms ($r = 0.78$; $p = 0.002$) indicate that delivery and scheduling functions materially affect time management in assessment; therefore, training should also cover transparent communication of automated checks and documentation of decisions when chatbot Q&A or anomaly flags intervene in graded activities.

Third, the top coefficients for simulators ($r = 0.85$) and augmented reality ($r = 0.81$) point to the promise of authentic and multimodal assessment; nonetheless, scaling such tasks requires explicit competencies in moderation/double-marking, reliability checks, and override rules when AI assistance (e.g., transcription or pre-marking) is used–competencies that should be embedded early in teacher education. Complementarily, positive associations for authoring tools ($r = 0.74$) suggest opportunities for generative co-design of rubrics and items; yet, because construct drift is a known risk, training must emphasize rubric engineering (coverage, alignment) and traceability.

TABLE 3 Correlations between assessment-aligned tool families and observed advantages in teacher training.

Assessment-aligned tool family (example)	Observed advantage	r	*p*
Delivery and integrity – virtual platforms (LMS/e-exam)	Improvement in time management	0.78	0.002
Authentic and multimodal evidence – simulators	Development of practical skills	0.85	0.001
Collaboration channels – social networks	Promotion of collaboration	0.69	0.004
Authentic and multimodal evidence – multimedia tools	Increased motivation	0.72	0.003
Peer/self-assessment discourse – online forums	Development of critical thinking	0.65	0.006
Automated feedback systems – evaluation platforms	Continuous feedback	0.83	0.002
Criteria and rubrics/authoring – authoring tools	Personalization of learning	0.74	0.003
Peer knowledge production – blogs and wikis	Knowledge sharing	0.68	0.004
Synchronous assessment events – videoconferences	Interaction and participation	0.76	0.002
Ubiquitous capture – mobile devices	Access to assessment resources	0.71	0.003
Authentic and multimodal evidence – augmented reality	Immersion in learning environments	0.81	0.001
Process orchestration – school-management software	Optimization of administrative processes	0.77	0.002

Also, although effects for blogs/wikis (r = 0.68) and social networks (r = 0.69) are lower relative to simulators or evaluation systems, they remain meaningful for peer and self-assessment. Here, outcomes are more context-dependent, which implies that teacher training should differentiate supports by learner profile–building interest and self-efficacy with supervised practice–so that collaboration produces assessable evidence rather than noise. Large multi-country evidence shows that attitudes, use, and interest interact to build AI self-efficacy, which in turn conditions responsible uptake of assessment technologies. See also recent work on AI empowerment, which highlights the value of guided, low-stakes engagement to develop confidence with AI-mediated problem solving.

Now, since these are bivariate correlations, they do not establish causality; moreover, several outcomes (e.g., motivation) are broader than assessment *per se*. Nevertheless, taken together–and read through the assessment lens adopted in Subsection "4.1 Results regarding the question: What digital tools have been used for student assessment in Ibero-American higher education?" –the pattern suggests where teacher-training curricula should concentrate: (1) automated-feedback design and auditing; (2) generative rubric/item validation; (3) authentic, multimodal task moderation; and (4) policy-aware orchestration of delivery and integrity tools. Finally, because verification-centered designs demonstrably temper naïve reliance on AI while preserving efficiency, embedding human-in-the-loop routines across these tool families is not optional but foundational.

Interpreted qualitatively, the observed associations appear to reflect underlying sociotechnical configurations rather than tool effects *per se*. Environments with clearer assistance boundaries, audit routines, and shared exemplars tend to report more formative uses of AI and stronger perceptions of fairness; conversely, contexts emphasizing surveillance or unmoderated automation report weaker student buy-in and limited pedagogical gains. These patterns suggest that correlations are contingent on enabling conditions–competency development, governance, and task design–that shape how technologies are enacted in practice. Thus, the quantitative patterns are best read as signals pointing to institutional arrangements that make trustworthy, learning-oriented assessment with AI possible.

5 Discussion

Across the corpus, what matters for the integration of AI into assessment in teacher education is not the accumulation of tools *per se*, but the reallocation of pedagogical work: from manual production to design, oversight, and documentation of AI-involved processes. Accordingly, the central task for Ibero-American higher education is to professionalize human-in-the-loop assessment–that is, to make verification, justification, and transparent communication routine elements of assessment design rather than exceptional safeguards. Experimental evidence shows that when verification and reflection are embedded into assessed tasks, reliance on generative systems becomes more discerning without sacrificing efficiency, which is precisely the stance needed in teacher training (Van Niekerk et al., 2025).

5.1 Implications for teacher-education curricula and institutional policy

First, programs should pivot from tool operation to capability building around five assessable competency clusters: (1) Feedback literacy with AI (criterion-anchored prompting, sampling and audit cycles, revision-based workflows); (2) Rubric/item validation (coverage, alignment, difficulty, and traceability to demonstrate how AI-drafted artifacts were accepted or corrected); (3) Data interpretation and stewardship (consent/minimization, fairness checks, and proportionate use of analytics and proctoring); (4) Integrity and transparency (clear assistance limits, explanation of automated checks, and student rights/appeals); and (5) Orchestration (documenting decisions across platforms and coordinating moderation/double-marking when AI assists scoring). Because attitudes, usage, and interest jointly build AI self-efficacy, these competencies should be taught

through guided, low-stakes practice rather than prohibition (Bewersdorff et al., 2025).

Second, at the institutional level, governance must travel with pedagogy. Concretely, providers should (a) require decision logs for significant AI involvement in grading or feedback; (b) implement sampling protocols for periodic audits of AI outputs; and (c) adopt procurement and deployment checklists (model/version notes, data-flow mapping, and fairness monitoring). These measures align responsibility with capability and make assessment explainable and contestable to learners.

5.2 Conceptual promises vs. situated realities in Ibero-American higher education

The international literature frames AI as a lever for formative assessment, personalization, and feedback literacy, if data pipelines, governance, and teacher preparation will co-evolve to support trustworthy, transparent use. In contrast, the Ibero-American corpus reveals a more incremental, operations-first trajectory: institutions prioritize proctoring, grading efficiency, and integrity controls that stabilize existing summative formats at scale, while formative redesign advances slowly and unevenly. Where the literature anticipates co-designed rubrics, dialogic feedback, and shared understandings of assistance limits, practice often manifests as instrumental uptake–automation of discrete tasks with limited visibility into models, datasets, or auditability. The gap is less about tool availability than about enabling conditions: policy clarity, assessment redesign capacity, and programmatic professional learning. In short, the conceptual promise is pedagogical transformation; the situated reality is controlled modernization, with learning gains contingent on pockets of strong design and governance rather than system-wide alignment.

5.3 Technical implementation vs. pedagogical acceptance: emerging tensions

Our synthesis surfaces recurrent tensions at the interface of technical deployment and pedagogical credibility. First, opacity vs. trust: educators and students question the provenance and validity of AI feedback when criteria mapping and error modes are not made explicit, dampening acceptance even where tools function reliably. Second, scale vs. relationship: automation accelerates turnaround but can displace dialogic moments unless workflows deliberately reintroduce human moderation (sampling, conferencing, calibration). Third, compliance vs. learning: integrity tooling and surveillance logics reduce misconduct risk yet risk narrowing demonstrations of competence and heightening anxiety, particularly in high-stakes e-exams. Fourth, efficiency vs. rigor: time saved in marking is offset by new labor in rubric refinement, prompt engineering, documentation, and audits; acceptance improves where institutions recognize and resource this reconfigured workload. Finally, innovation vs. coherence: local pilots flourish, but uneven policy signals and fragmented support produce "islands of practice," making it difficult to build shared norms about what AI may assist and how that assistance is disclosed and evaluated. Pedagogical acceptance grows when implementations foreground explainability, preserve space for human judgment, and embed feedback loops that make AI outputs accountable to stated criteria.

5.4 Which teaching roles are most challenged by AI-assisted assessment?

AI-assisted assessment reconfigures, but does not replace, core academic roles; the challenges cluster around roles that arbitrate quality, meaning, and fairness.

- Assessors and graders. These roles face the steepest shift in practice. They must interpret and edit AI-generated comments, run audit samples, and justify decisions when human and machine judgments diverge. Without calibration time and traceability tools, perceived fairness–and thus acceptance–suffers.

- Task and rubric designers. Designers bear responsibility for aligning prompts, criteria, and acceptable assistance. The challenge is preventing construct drift (AI suggesting decontextualized criteria) while maintaining disciplinary integrity. Effective practice requires iterative co-design with exemplars and student-facing plain-language criteria.

- Program-level coordinators and quality assurers. At program scale, coordinators must harmonize assistance policies, disclosure norms, and appeal routes across courses. They balance innovation with comparability of standards, an administrative and cultural task as much as a technical one.

- Academic advisors and learning support staff. As analytics and early-warning systems expand, advisors must translate indicators into humane, actionable guidance. The challenge is resisting reductive labeling and ensuring that data-driven flags trigger relational support rather than punitive responses.

- Clinical/practicum supervisors and instructors of authentic, multimodal work. These roles must adjudicate evidence that blends human and AI contributions. They report the greatest need for protocols that preserve authorship, identity, and reflective depth when curation or drafting is AI-assisted.

5.5 Conceptual contribution

The review advances a function-by-purpose view of assessment technology that distinguishes pre-AI baselines from AI-specific mechanisms (generativity, adaptivity, algorithmic feedback/analytics). This lens treats assessment as a socio-technical practice whose reliability depends on the coupling of tool affordances with trained human oversight. In doing so, it reframes "innovation" from adoption of platforms to redesign of assessment workflows around design quality, evidence credibility, and fairness.

5.6 Limitations

It is important to note that this study reviewed only Scopus as an academic database, and therefore, the identified and analyzed studies are limited to this environment and search engine. However,

other databases may contain complementary studies. In the field of technology, where scientific output advances rapidly, it is crucial to stay updated with the latest research findings. Consequently, researchers using the results of this study should consider the timeframe of the reviewed studies and compare them with more recent ones.

5.7 Future research

Future work should: (a) run workflow-level trials comparing assessment designs with/without AI while holding learning goals constant, measuring not only performance but also oversight workload and fairness, (b) test the teachability of the competency clusters above through controlled training interventions, using validated measures of AI self-efficacy/empowerment, (c) evaluate integrity and transparency regimes (assistance limits, audit trails, appeal processes) under real grading conditions; and, (d) examine enabling conditions (ecosystem quality, support services) that determine when AI-enhanced assessment delivers trustworthy personalization at scale.

As a final insight, it is noteworthy to mention that the sector is positioned to benefit from AI not by replacing judgment, but by elevating it: educators who can co-design with generative systems, audit algorithmic feedback, and communicate rationales clearly will turn AI from a productivity add-on into a credible assessment partner–one that improves frequency, personalization, and trust without eroding professional agency.

Author contributions

DF-R: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. AC: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. JM-N: Conceptualization, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We thank Universidad de La Sabana [Technologies for Academia - Proventus Research Group (EDUPHD-20-2022 Project)], for the support received for the preparation of this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Abuzar, M., Mahmudulhassan, and Muthoifin. (2025). University students' trust in AI: Examining reliance and strategies for critical engagement. *Int. J. Interact. Mobile Technol.* 19, 70–82. doi: 10.3991/ijim.v19i07.52875

Amoako, K., Asante, A., and Owusu, K. (2024). AI-powered tools for personalized learning in educational technology. *Int. J. Technol. Model.* 3, 46–56. doi: 10.63876/ijtm.v3i1.115

Arkorful, V., Salifu, I., Arthur, F., and Abam Nortey, S. (2024). Exploring the nexus between digital competencies and digital citizenship of higher education students: A PLS-SEM approach. *Cogent Educ.* 11, 1–20. doi: 10.1080/2331186X.2024.2326722

Barrera Castro, G. P., Chiappe, A., Becerra Rodriguez, D. F., and Sepulveda, F. G. (2024). Harnessing AI for education 4.0: Drivers of personalized learning. *Electron. J. eLearn.* 22, 01–14. doi: 10.34190/ejel.22.5.3467

Bewersdorff, A., Hornberger, M., Nerdel, C., and Schiff, D. S. (2025). AI advocates and cautious critics: How AI attitudes, AI interest, use of AI, and AI literacy build university students' AI self-efficacy. *Comput. Educ. Artif. Intell.* 8:100340. doi: 10.1016/j.caeai.2024.100340

Cabero Almenara, J., and Martínez Gimeno, A. (2019). Las TIC y la formación inicial de los docentes. Modelos y competencias digitales. *Profesor. Rev. Currículum Form. Profesor.* 23, 247–268. doi: 10.30827/profesorado.v23i3.9421

Chaparro-Martínez, E., Álvarez-Muñoz, P., and Armas-Regnault, M. (2016). Gestión de la información: Uso de las bases de datos scopus y web of science con fines académicos. *Univers. Ciencia Tecnol.* 20, 166–175.

Chituc, C.-M. (2021). "A framework for education 4.0 in digital education ecosystems," in *Smart and sustainable collaborative networks 4.0*, Vol. 629, eds L. M. Camarinha-Matos, X. Boucher, and H. Afsarmanesh (Cham: Springer International Publishing), 702–709. doi: 10.1007/978-3-030-85969-5_66

Cisneros-Barahona, A. S., Marqués-Molías, L., Samaniego-Erazo, G., and Mejía-Granizo, C. (2024). Assessing teacher digital competence. An analysis integrating

descriptive, inferential, and multivariate perspectives. *RIED Rev. Iberoam. Educ. Dist.* 27:22. doi: 10.5944/ried.27.2.39122

Flores-Viva, J.-M., and García-Peñalvo, F.-J. (2023). Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4). *Comun. Med. Educ. Res. J.* 31, 35–44. doi: 10.3916/C74-2023-03

Gómez Sánchez, T. F., Bobadilla-Pérez, M., Arcas, B. R., Fraga-VIñas, L., and Galán-Ridríguez, N. M. (2024). ICT integration in FLT: An analysis of TPACK implementation in Spanish primary teacher education. *Digit. Educ. Rev.* 45, 214–221. doi: 10.1344/der.2024.45.214-221

González-Pérez, L. I., and Ramírez-Montoya, M. S. (2022). Components of education 4.0 in 21st century skills frameworks: Systematic review. *Sustainability* 14:1493. doi: 10.3390/su14031493

González-Salamanca, J. C., Agudelo, O. L., and Salinas, J. (2020). Key competences, education for sustainable development and strategies for the development of 21st century skills. A systematic literature review. *Sustainability* 12:10366. doi: 10.3390/su122410366

Lawrence, J. E., and Tar, U. A. (2018). Factors that influence teachers' adoption and integration of ICT in teaching/learning process. *Educ. Media Int.* 55, 79–105. doi: 10.1080/09523987.2018.1439712

Marimon-Martí, M., Cabero, J., Castañeda, L., Coll, C., De Oliveira, J. M., and Rodríguez-Triana, M. J. (2022). Construir el conocimiento en la era digital: Retos y reflexiones. *Rev. Educ. Dist.* 22:61. doi: 10.6018/red.505661

Martín-Párraga, L., Llorente-Cejudo, C., and Almenara, J. C. (2024). ICTs as a space for progress towards sustainable development goal 4 (SDG4). *Rev. Lusofona Educ.* 61, 75–88. doi: 10.24140/issn.1645-7250.rle61.05

Miranda, J. P. P., Dianelo, R. F. B., Gamboa, A. B., Bansil, J. A., Regala, A. R., and Simpao, L. S. (2024). Identification, validity, and reliability of the 21st-century workplace skills for on-the-job training practicum. *Int. J. Eval. Res. Educ.* 13, 3485–3492. doi: 10.11591/ijere.v13i5.28919

Mishra, P., and Koehler, M. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teach. Coll. Rec.* 108, 1017–1054. doi: 10.1111/j.1467-9620.2006.006

Oke, A., and Fernandes, F. A. P. (2020). Innovations in teaching and learning: Exploring the perceptions of the education sector on the 4th industrial revolution (4IR). *J. Open Innov. Technol. Mark. Complex.* 6, 31. doi: 10.3390/joitmc6020031

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 372:n71. doi: 10.1136/bmj.n71

Qian, Y. (2025). Pedagogical applications of generative ai in higher education: A systematic review of the field. *TechTrends* doi: 10.1007/s11528-025-01100-1

Rojas, M. P., and Chiappe, A. (2024). Artificial intelligence and digital ecosystems in education: A review. *Technol. Knowledge Learn.* 29, 2153–2170. doi: 10.1007/s10758-024-09732-7

Stephenson, B., and Harvey, A. (2022). "Student equity in the age of ai-enabled assessment," in *Assessment for inclusion in higher education*, 1st Edn, eds R. Ajjawi, J. Tai, D. Boud, and T. Jorre De St Jorre (London: Routledge), 120–130. doi: 10.4324/9781003293101-14

Trajkovski, G., and Hayes, H. (2025). "AI in assessment analysis and improvement," in *AI-assisted assessment in education*, Digital education and learning (Cham: Palgrave Macmillan), 159–192. doi: 10.1007/978-3-031-88252-4_4

Van Niekerk, J., Delport, P. M. J., and Sutherland, I. (2025). Addressing the use of generative AI in academic writing. *Comput. Educ. Artif. Intell.* 8:100342. doi: 10.1016/j.caeai.2024.100342

Wang, Y. (2024). Algorithmic decisions in education governance: Implications and challenges. *Discov. Educ.* 3:229. doi: 10.1007/s44217-024-00337-x