



Adaptive assessment using artificial intelligence to measure learning outcomes and student competencies

Evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles

Segundo Alexander Rosado Martínez¹  

¹ Instituto de Investigación, Desarrollo e Innovación JOGB, Milagro, Ecuador

Submitted: 19-08-2025 | Revised: 01-09-2025 | Accepted: 19-12-2025 | Published: 28-12-2025

Corresponding Author: Segundo Alexander Rosado Martínez 

ABSTRACT

Introduction: Evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles is an emerging higher-education issue related to measurement quality, decision-making and learning experience. **Objective:** To systematically synthesize available evidence and identify patterns of effectiveness, methodological quality and implementation conditions. **Method:** A PRISMA-oriented systematic review was conducted on a consolidated academic corpus of 135 records; 24 studies met relevance and documentary sufficiency criteria. **Results:** Evidence was methodologically heterogeneous but converged on the need to align technology, constructs, feedback and educational decisions. **Conclusions:** Findings support evidence-informed adoption, institutional monitoring and continuous assessment of validity, equity and utility.

Keywords: higher education; systematic review; educational assessment; evidence; PRISMA; educational technology.

RESUMEN

Introducción: Evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles representa un problema emergente para la educación superior por su relación con la calidad de la medición, la toma de decisiones y la experiencia de aprendizaje. **Objetivo:** sintetizar sistemáticamente la evidencia disponible y establecer patrones de efectividad, calidad metodológica y condiciones de implementación. **Método:** se aplicó una revisión sistemática con lógica PRISMA sobre un corpus académico consolidado de 135 registros, del cual se incluyeron 24 estudios que cumplieron criterios de pertinencia y suficiencia documental. **Resultados:** la evidencia muestra heterogeneidad de diseños y contextos, junto con convergencias en torno a la necesidad de alinear tecnología, constructo, retroalimentación y decisiones educativas. **Conclusiones:** los resultados respaldan una adopción basada en evidencia, seguimiento institucional y evaluación continua de validez, equidad y utilidad.

Palabras clave: educación superior; revisión sistemática; evaluación educativa; evidencia; PRISMA; tecnología educativa.

INTRODUCTION

University research on adaptive assessment using artificial intelligence to measure learning outcomes and student competencies has gained increasing relevance because assessment decisions no longer depend exclusively on static instruments or face-to-face exchanges. Digitalization has expanded the amount of available evidence on performance, interaction, and learning, but it has also increased the need to distinguish between technically attractive innovations and practices capable of supporting valid educational inferences. This review addresses this problem from an integrative perspective, interested both in the observed effects and in the methodological conditions that allow interpreting such effects without confusing technological availability with assessment quality.

The conceptual core of the manuscript is organized around adaptation and personalization. This choice avoids treating the literature as a simple collection of results and allows examining how each study defines the phenomenon, what population it observes, what instruments it employs, and what type of conclusion it considers legitimate. The heterogeneity of university contexts makes it necessary to pay attention to the relationship between design, implementation, and evidence, because the same strategy can produce benefits under certain conditions and modest results when teaching support, discipline, digital experience, or the form of feedback change.

This systematic review is pertinent because the field combines experimental studies, instrumental validations, observational analyses, technological developments, and previous syntheses. This diversity makes it difficult to draw solid conclusions through a fragmentary reading. Instead of assuming that all works respond to the same question, this article reconstructs patterns of convergence and divergence, identifies which results are repeated with greater consistency, and separates claims directly supported by studies from those that still require additional validation. The purpose is to offer a useful reading for research, teaching, and university management.

METHODS

Yang et al (1) provided evidence published in 2022 through a 7-week comparative experiment. The study involved students from a university programming course, allowing the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study showed that the CAT+memory system improved performance and engagement in practices and materials compared with the other conditions. This antecedent is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The study (2), published in 2025, reported evidence obtained through a quasi-experimental and analytical approach. It involved experimental and control groups in STEM, allowing the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the study showed that the experimental group outperformed the control group in programming and mathematics; greater interaction was related to progress, and 80% valued the adaptive feedback. This antecedent is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between the construct, the measurement procedure, and the application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The foundational methodological article by David (3), published in 1984, addressed CAT applications in educational problems, allowing the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the work established foundations for the use of IRT/CAT for adaptive educational decisions. This antecedent is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The review reported in (4), published in 2026, was a JBI/PRISMA-ScR scoping review that included 9 studies, allowing the scope of its inferences to be situated and its proximity to the university population of interest to be assessed. In substantive terms, the review found evidence suggesting improvements in learning and adaptive support, and it highlighted the need for more

robust comparative studies and implementation evaluation. This antecedent is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The work (5) provided evidence through a systematic review published in 2025. The study used ITS corpora from engineering, making it possible to situate the scope of its inferences and to assess its proximity to the university population of interest. Substantively, the work showed that personalization and feedback are the most frequent mechanisms and that gaps persist in longitudinal evaluation, equity, and external validation. This prior work is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. In this review, its contribution is considered part of the comparative pattern, not an isolated demonstration sufficient to generalize to all educational scenarios.

The evidence synthesis (6) was reported in a systematic review published in 2019. The study used an ITS corpus, making it possible to situate the scope of its inferences and to assess its proximity to the university population of interest. Substantively, the work showed that ITSs yield consistent benefits when they combine student modeling, immediate feedback, and adaptation, although the evaluation is heterogeneous. This prior work is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. In this review, its contribution is considered part of the comparative pattern, not an isolated demonstration sufficient to generalize to all educational scenarios.

The examined contribution (7) reported evidence in 2025 using a research-and-evaluation-based design. The study focused on engineering/design students, making it possible to situate the scope of its inferences and to assess its proximity to the university population of interest. Substantively, the work showed that the ITS enabled individualized recommendations and support for the design process, although it requires careful pedagogical integration. This prior work is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. In this review, its contribution is considered part of the comparative pattern, not an isolated demonstration sufficient to generalize to all educational scenarios.

The study (8) reported evidence in 2021 through computational development/evaluation. The study analyzed educational dialogue data, making it possible to situate the scope of its inferences and to assess its proximity to the university population of interest. Substantively, the work showed that deep discourse analysis made it possible to identify learning states and to support feedback adapted to the student. This prior work is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. In this review, its contribution is considered part of the comparative pattern, not an isolated demonstration sufficient to generalize to all educational scenarios.

The research (9) presented evidence in 2020 from a naturalistic/comparative usage study. The study involved students with voluntary use of the ITS, making it possible to situate the scope of its inferences and to assess its proximity to the university population of interest. Substantively, the work showed that the benefit depends on adoption and usage patterns; the availability of the tutor does not guarantee that all students use it sufficiently. This prior work is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. In this review, its contribution is considered part of the comparative pattern, not an isolated demonstration sufficient to generalize to all educational scenarios.

The study (10) presented evidence in 2025 using an explanatory quantitative design. The study examined students in an ITS environment, making it possible to situate the scope of its inferences and to assess its proximity to the university population of interest. Substantively, the study showed that the effect of the ITS can be unequal when gaps exist in access, digital skills, or effective use; personalization does not automatically eliminate inequities. This prior work is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. In this review, its contribution is considered part of the comparative pattern, not an isolated demonstration sufficient to generalize to all educational scenarios.

The evidence synthesis (11) reported findings in 2022 based on interaction analytics. The

study involved adults with low literacy, making it possible to situate the scope of its inferences and to assess its proximity to the university population of interest. Substantively, the work identified differentiated interaction profiles, which are useful for adapting feedback and support to individual needs. This prior work is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. In this review, its contribution is considered part of the comparative pattern, not an isolated demonstration sufficient to generalize to all educational scenarios.

The examined contribution (12) provided evidence published in 2014 through a meta-analysis. The study worked with ITS outcome studies, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. In substantive terms, the work showed that ITS have positive effects on learning compared with several controls; the benefit varies by domain and type of implementation. This antecedent is relevant for the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational settings.

Hamida et al. (13) provided evidence published in 2026 through technical accessibility evaluation. The study worked with students with special educational needs in higher education, which allows situating the scope of their inferences and assessing their proximity to the university population of interest. In substantive terms, the work showed that it identifies that digital assessment accessibility requires compatibility with assistive technologies, autonomy, and integrity under examination constraints. This antecedent is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational settings.

The research (14) provided evidence published in 2024 through a longitudinal classroom experiment. The study worked with ITS $n=35$; control $n=36$, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. In substantive terms, the study showed that the ITS produced learning advantages, especially in students with lower prior knowledge; possible overdependence was noted. This antecedent is relevant for the perspective of adaptation and personalization because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and context of application. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Agustín et al. (15) provided evidence published in 2024 through scale development and validation. The study involved students from a Venezuelan university, making it possible to situate the scope of its inferences and assess its proximity to the university population of interest. In substantive terms, the work showed that a model of 22 variables in 4 dimensions with satisfactory psychometric properties was obtained. This antecedent is relevant for the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence among construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The synthesized evidence (16) provided evidence published in 2022 through the design and evaluation of ITS. The study worked with programming students, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. In substantive terms, the work showed that the system demonstrated viability for feedback and scaffolding of programming, with adjustments necessary to contexts and student profiles. This precedent is relevant for the perspective of adaptation and personalization because it shows that the interpretation of the effect depends on the correspondence between the construct, the measurement procedure, and the application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Xinghua et al. (17) provided evidence published in 2021 through psychometric development and validation. The study worked with university students; sample size not visible in the retrieved fragment, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. In substantive terms, the work showed that the study developed a specific measure for university digital competence with validity evidence for digitally enhanced

learning. This precedent is relevant for the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

No. (18) provided evidence published in 2021 through a systematic review. The study worked with 26 studies (2010–2019), which allows situating the scope of its inferences and assessing its proximity to the university population of interest. In substantive terms, the work showed that authentic assessment was associated with greater engagement/satisfaction and competencies in communication, collaboration, critical thinking, and problem solving. This antecedent is relevant for the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Paul (19) provided evidence published in 2012 through a design and implementation study. The study worked with interviews with teachers/students and pilots 2005–2007, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. In substantive terms, the work showed that accessible pDF allowed greater independence and lower cost than reader/scribe; requests for digital exams increased 5 times 2008–2011. This antecedent is relevant for the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The work (20) contributed evidence published in 2025 through sequential mixed methods. The study involved university students who were users of GPTutor, which makes it possible to situate the scope of its inferences and assess its proximity to the university population of interest. In substantive terms, the work showed that gPTutor promoted engagement and personalized support; the quality of interaction and prompt/tutoring design conditioned the experience. This previous study is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The synthesized evidence (21) provided evidence published in 2024 through a comparison by experience level. The study worked with programming students, grouped by experience, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. In substantive terms, the work showed that effectiveness varied according to prior experience, confirming that personalization must consider the learner's profile. This previous study is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Joanna et al. (22) provided evidence published in 2023 using a qualitative sociomaterial approach. The study involved 40 students with disabilities from two universities, which allows situating the scope of their inferences and assessing their proximity to the university population of interest. In substantive terms, the work showed that online exams, use of technology, more time, and open-book format increased inclusion for many; flexibility and acceptance by staff were key. This prior study is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

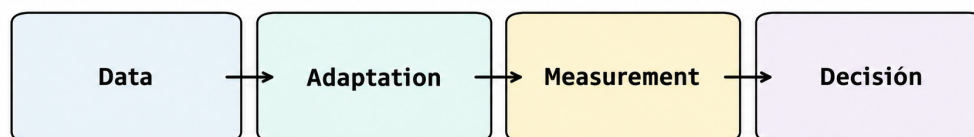
The study (23) provided evidence published in 2016 through a meta-analysis. The study worked with comparative studies of ITS, which allows situating the scope of its inferences and assessing its proximity to the university population of interest. In substantive terms, the work showed that ITS on average outperform several forms of conventional/computerized instruction, although the effect depends on the comparator and design. This antecedent is relevant for the adaptation and personalization perspective because it shows that the interpretation of the effect depends on

the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

The research (24) provided evidence published in 2026 through a large-sample pre-post. The study used Pre $n=1,437$; post $n=871$, which makes it possible to situate the scope of its inferences and assess its proximity to the university population of interest. In substantive terms, the work showed that significant improvements were observed in all evaluated digital competencies and a partial reduction of gaps between groups; digital self-efficacy increased. This antecedent is relevant to the adaptation and personalization perspective because it shows that the interpretation of the effect depends on the correspondence between construct, measurement procedure, and application context. Its contribution is considered in this review as a piece of the comparative pattern, not as an isolated demonstration sufficient to generalize to all educational scenarios.

Figure 1. Conceptual structure of the review.

Conceptual structure of the review



Note. Own elaboration based on the analytical axes defined for the manuscript.

The figure organizes the article's reasoning as a conceptual sequence and not as a universal template. The represented components correspond to the specific problem of this review and show the transition from the inputs or initial conditions to the educational interpretation. Its utility lies in making visible the relationship between constructs that, in individual studies, often appear fragmented between methodological and results sections. This representation guides the subsequent synthesis and facilitates distinguishing the elements that act as conditions, mechanisms, or results within the corpus.

Review questions

Question 1. What patterns of evidence allow an understanding of dimension 1 of adaptation and personalization in relation to adaptive assessment using artificial intelligence to measure learning outcomes and student competencies, and what methodological or contextual conditions modify its interpretation?

Question 2. What patterns of evidence allow an understanding of dimension 2 of adaptation and personalization in relation to adaptive assessment through artificial intelligence to measure learning outcomes and student competencies, and what methodological or contextual conditions modify its interpretation?

Question 3. What patterns of evidence allow an understanding of dimension 3 of adaptation and personalization in relation to adaptive assessment using artificial intelligence to measure learning outcomes and student competencies, and what methodological or contextual conditions modify its interpretation?

Question 4. What patterns of evidence allow an understanding of dimension 4 of adaptation and personalization in relation to adaptive assessment through artificial intelligence to measure learning outcomes and student competencies, and what methodological or contextual conditions modify its interpretation?

Question 5. What patterns of evidence allow an understanding of dimension 5 of adaptation and personalization in relation to adaptive assessment through artificial intelligence to measure learning outcomes and student competencies, and what methodological or contextual conditions modify its interpretation?

METHODS

PRISMA strategy and extraction

A systematic review was developed to synthesize academic evidence relevant to the topic of the manuscript. The unit of analysis was the academic study or reference with a traceable source, identifiable methodology, and retrievable results. The consolidated academic working corpus consisted of 135 records previously gathered in the evidence matrices. A thematic and documentary selection was applied to that set to establish the final corpus. The procedure was organized according to the PRISMA logic of identification, screening, eligibility, and inclusion, adapted to the consolidated nature of the provided sources.

The inclusion criteria required a direct relationship with adaptive assessment using artificial intelligence to measure learning outcomes and student competencies, correspondence with higher education or with instruments applicable to that level, sufficient description of the method, and availability of interpretable results. A total of 111 records were excluded because they had lower relevance to the specific question, addressed peripheral populations or outcomes, or did not provide sufficient methodological evidence for the focused synthesis. The final corpus comprised 24 studies. The selection was based on the information documented in the matrices, and attribution of counts to a specific bibliographic database was avoided when not supported by the files.

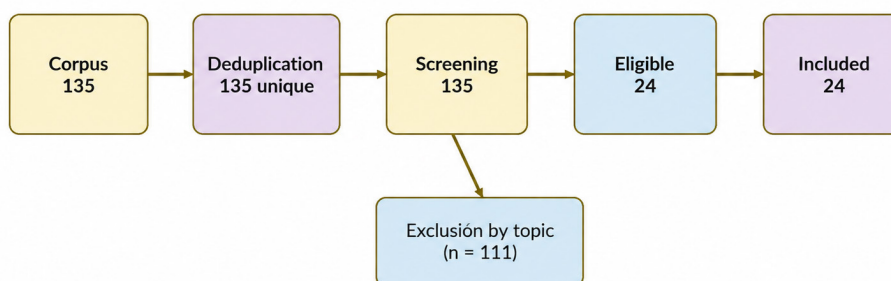
The information extraction was structured using a matrix with authors, year, title, source, DOI or traceable link, provenance, study type, sample or corpus, summarized methodology, main results, level of evidence, and relevance. This organization made it possible to compare works with heterogeneous designs without reducing them to a single metric. The synthesis combined descriptive characterization of the corpus and thematic analysis of the results, paying attention to convergences, contradictions, implementation conditions, and limits of generalization.

The evaluation of the evidence focused on the sufficiency of the retrieved information to answer the review question. The clarity of the design, the explicitness of the sample or corpus, the correspondence between method and results, and the traceability of the source were considered jointly. Meta-analysis was not performed when the heterogeneity of outcomes, instruments, populations, or metrics prevented a responsible quantitative combination. In such cases, a comparative synthesis was prioritized, preserving the differences between studies and avoiding the transformation of non-equivalent results into a single estimator.

To strengthen transparency, the results were organized into characterization and synthesis tables, and the figures represent the selection flow and the temporal distribution of the corpus. Each table was constructed from the fields of the supplied matrix. Subsequent interpretations were developed based on observable patterns in the set and not on added assumptions.

Figure 2. Flow of study selection according to PRISMA logic.

Study selection flowchart — review 3



Identification → deduplication → screening → eligibility → inclusion

Source: own elaboration based on the evidence matrix of the review.

Note. The flow represents 135 normalized academic records, without duplicates in the consolidated corpus; 111 records did not meet the thematic relevance threshold and 24 studies were included in the final synthesis.

This criterion is especially important in educational reviews, where variation across disciplines, technologies, instruments, and institutional contexts can substantially modify the magnitude and meaning of the findings.

The flow documents the screening and selection of evidence. The 135 records from the consolidated corpus were normalized by DOI or title, so no duplicates were identified at this stage. Thematic screening excluded 111 records due to insufficient correspondence with the specific review question; 24 studies with sufficient methodological evidence and retrievable results were included in the synthesis. The sequence allows distinguishing documentary identification from eligibility judgment and avoids interpreting the reduction of the corpus as an arbitrary loss of information.

RESULTS

The final corpus comprised 24 studies published during 1984–2026. The temporal distribution reveals an active and heterogeneous research field, with studies corresponding to different stages of maturation of the problem. Some focus on demonstrating feasibility or association, whereas others progress toward validation, comparison of strategies, or analysis of implementation conditions. This temporal breadth allows us to observe not only favorable results but also shifts in how evaluation quality is conceptualized and in the type of evidence considered necessary to support educational decisions.

The methodological characterization confirms that adaptive assessment using artificial intelligence to measure learning outcomes and student competencies does not constitute a homogeneous object. The set includes designs with distinct purposes and varying levels of control over study conditions. This diversity broadens the scope of the review, but also requires that findings be interpreted according to the function of each design. Intervention studies report on changes observed under defined conditions; instrumental studies provide evidence on measurement; observational analyses describe associations; and previous reviews allow trends in the field to be recognized without replacing the evaluation of primary studies.

The samples and corpora described in the studies exhibit considerable variability. This variation is relevant because the transferability of the results depends on the size, composition, and context of the participants, as well as the academic discipline and technological environment. Overall, the evidence suggests that the most interpretable results are those that specify the population, the evaluative task, and the mechanism by which the intervention or instrument should produce the expected effect. When any of these elements remains poorly defined, the usefulness of the finding for institutional decisions decreases.

Table 1. Methodological characterization of the included corpus

Year	Design/methodology	Sample or corpus	Evidence
2022	7-week comparative experiment	Students in a university programming course	High
2025	Quasi-experimental and analytical	Experimental/control groups in STEM	Mean
1984	Foundational methodological article	Applications of CAT in educational problems	High
2026	Scoping review JBI/PRISMA-ScR	9 studies	Mean
2025	Systematic review	Corpus of ITS in engineering	Mean
2019	Systematic review	ITS corpus	Mean
2025	Research-based design and evaluation	Engineering/design students	Mean
2021	Computational development/evaluation	Educational dialogue data	Mean
2020	Naturalistic/comparative usage study	Students with voluntary use of ITS	Mean
2025	Explanatory quantitative study	Students in an environment with ITS	Mean
2022	Interaction analytics	Adults with low literacy	Mean
2014	Meta-analysis	ITS outcome studies	Mean

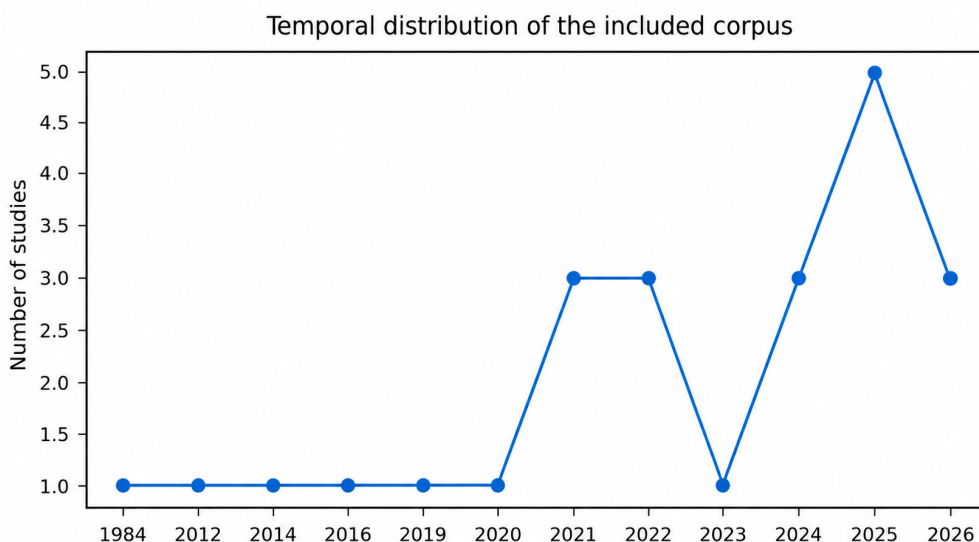
Note. Data extracted from the evidence matrix provided for this review.

The synthesis of results reveals a central pattern linked to adaptation and personalization: favorable effects or properties do not appear as automatic attributes of the technology or instrument, but rather as outcomes conditioned by design, implementation, data quality, and pedagogical alignment. The studies converge on the importance of articulating the tool with

clear learning objectives, comprehensible evaluation criteria, and monitoring mechanisms. The divergences center on the magnitude of the effects, stability across contexts, and the capacity to sustain outcomes when institutional conditions change.

The table shows the methodological diversity of the corpus and reveals that the studies do not offer the same type of inference. The presence of designs with different objectives requires interpreting each result according to the question that the study can answer. This differentiation avoids directly comparing indicators that serve different functions and helps identify where real convergence exists. It also shows that the robustness of a conclusion depends on the combination of methodological clarity, sample description, and sufficiency of the retrieved evidence.

Figure 3. Temporal distribution of the included studies.



Note. Frequencies calculated from the year of publication recorded in the evidence matrix.

The temporal distribution makes it possible to situate the recent intensity of scientific production and to recognize whether the field is in expansion, consolidation, or methodological transition. Rather than interpreting the number of publications as a measure of quality, the figure serves to contextualize changes in questions, technologies, and instruments observed in the corpus. Reading the figure together with the methodological characterization makes it possible to identify whether the most recent works expand the evidence through more robust designs or whether exploratory approaches continue to predominate.

Results on measurement and adaptation

Another cross-cutting pattern is the relevance of evidence quality. Studies with explicit procedures allow better distinction among real improvement, favorable perception, and simple ease of use. In contrast, when the result relies mainly on self-report or short-term indicators, interpretation requires greater caution. The review shows that the practical utility of a strategy depends not only on whether it produces a statistical difference or a positive appraisal, but on whether the measurement is consistent with the construct and whether the intervention can be sustainably integrated into real educational processes.

In the subset of evidence corresponding to this axis, the results can be jointly summarized as follows: The CAT+memory system improved performance and engagement in practices and materials compared with the other conditions. The experimental group outperformed the control group in programming and mathematics; greater interaction was related to progress, and 80% valued adaptive feedback. These findings establish foundations for the use of IRT/CAT for adaptive educational decisions. The comparative reading indicates that these findings should not be interpreted as equivalent because they derive from different designs and metrics. Nevertheless, their thematic convergence allows identifying recurring conditions and supports an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, results were observed that, considered jointly, can be summarized as follows: The evidence suggests improvements in learning and adaptive support, with a need for more robust comparative studies and implementation evaluation. Personalization and feedback are the most frequent mechanisms; gaps persist in longitudinal evaluation, equity, and external validation. ITS show consistent benefits when they combine student modeling, immediate feedback, and adaptation; the evaluation is heterogeneous. The comparative reading indicates that these findings should not be interpreted as equivalent, because they come from different designs and metrics. However, their thematic convergence allows identifying recurrent conditions and formulating an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, results were observed that, considered jointly, can be summarized as follows: The ITS allowed individualized recommendations and support for the design process, although it requires careful pedagogical integration. Deep discourse analysis made it possible to identify learning states and support feedback adapted to the student. The benefit depends on adoption and usage patterns; the availability of the tutor does not guarantee that all students use it sufficiently. Comparative reading indicates that these findings should not be interpreted as equivalent, because they come from different designs and metrics. However, their thematic convergence allows identifying recurrent conditions and formulating an overall interpretation centered on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, results were observed that, considered jointly, can be summarized as follows: The effect of ITS can be uneven when there are gaps in access, digital skills, or effective use; personalization does not automatically eliminate inequities. Differentiated interaction profiles were identified, useful for adapting feedback and support to individual needs. ITS show positive effects on learning compared with various controls; the benefit varies by domain and type of implementation. The comparative reading indicates that these findings should not be interpreted as equivalent, because they come from different designs and metrics. However, their thematic convergence allows identifying recurrent conditions and formulating an overall interpretation centered on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, results were observed that, considered jointly, can be summarized as follows: It is identified that digital assessment accessibility requires compatibility with assistive technologies, autonomy, and integrity under examination restrictions. The ITS produced learning advantages, especially in students with less prior knowledge; possible overdependence was noted. A model of 22 variables in 4 dimensions with satisfactory psychometric properties was obtained. The comparative reading indicates that these findings should not be interpreted as equivalent, because they proceed from different designs and metrics. However, their thematic convergence allows identifying recurrent conditions and formulating an overall interpretation centered on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, results were observed that, considered jointly, can be summarized as follows: The system showed viability for programming feedback and scaffolding, with necessary adjustments to contexts and student profiles. The study developed a specific measure for university digital competence with validity evidence for digitally enhanced learning. Authentic assessment was associated with greater engagement/satisfaction and communication, collaboration, critical thinking, and problem-solving competencies. The comparative reading indicates that these findings should not be interpreted as equivalent, because they come from different designs and metrics. However, their thematic convergence allows identifying recurrent conditions and formulating an overall interpretation focused on the relationship between design quality, student response, and the usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, results were observed that, considered jointly, can be summarized as follows: accessible PDF allowed greater independence and lower cost than reader/scribe; requests for digital exams increased 5-fold 2008–2011. GPTutor favored engagement and personalized support; the quality of interaction and design of prompts/tutoring conditioned the experience. Effectiveness varied according to prior experience, confirming that personalization must consider the learner profile. Comparative reading indicates that these findings should not be interpreted as equivalent, because they come from different designs and metrics. However, their thematic convergence allows identifying recurrent conditions and formulating an

overall interpretation focused on the relationship between quality of design, student response, and usefulness of the information produced to guide educational decisions.

In the subset of evidence corresponding to this axis, the results, considered jointly, can be summarized as follows: Online exams, use of technology, more time, and open-book format increased inclusion for many; staff flexibility and acceptance were key. ITS, on average, outperform various forms of conventional or computerized instruction, although the effect depends on the comparator and study design. Significant improvements were observed in all assessed digital competencies, along with a partial reduction of gaps between groups; digital self-efficacy also increased. The comparative reading indicates that these findings should not be interpreted as equivalent, because they derive from different designs and metrics. However, their thematic convergence makes it possible to identify recurrent conditions and to formulate an overall interpretation focused on the relationship among design quality, student response, and the usefulness of the information produced to guide educational decisions.

Table 2. Synthesis of relevant findings

Study	Recovered finding	Relevance
Adaptive formative assessment system based on computerized adaptive testing and the learning	The CAT+memory system improved performance and engagement in practice and materials compared with the other conditions.	Included due to direct relevance.
Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact	The experimental group outperformed the control in programming and mathematics; greater interaction was associated with progress, and 80% valued the adaptive feedback.	Included due to direct relevance.
Application of Computerized Adaptive Testing to Educational Problems	Establishes foundations for the use of IRT/CAT for adaptive educational decisions.	Included due to direct relevance.
Artificial intelligence in nursing education: A scoping review of intelligent tutoring systems	The evidence suggests improvements in learning and adaptive support, with a need for more robust comparative studies and implementation evaluation.	Included due to direct relevance.
A Systematic Literature Review of AI-Driven Intelligent Tutoring Systems in Engineering Education	Personalization and feedback are the most frequent mechanisms; gaps persist in longitudinal evaluation, equity, and external validation.	Included due to direct relevance.
Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation	ITS show consistent benefits when they combine student modeling, immediate feedback, and adaptation; the evaluation is heterogeneous.	Included due to direct relevance.
Development of an intelligent tutoring system for design education: personalised learning	The ITS enabled individualized recommendations and support for the design process, although it requires careful pedagogical integration.	Included due to direct relevance.
Deep Discourse Analysis for Personalized Feedback in Intelligent Tutoring Systems	The in-depth analysis of the discourse made it possible to identify learning states and support feedback adapted to the student.	Included due to direct relevance.
Effect of Non-mandatory Use of an Intelligent Tutoring System on Students' Learning	The benefit depends on adoption and usage patterns; the availability of the tutor does not guarantee that all students use it sufficiently.	Included due to direct relevance.
Widening the Digital Divide: The mediating role of Intelligent Tutoring Systems in learning	The effect of ITS can be uneven when there are gaps in access, digital skills, or effective use; personalization does not automatically eliminate them.	Included due to direct relevance.

Note. The textual synthesis is based exclusively on the results recorded in the matrix.

The synthesis shows that favorable evidence is distributed across different types of outcomes and that the relevance of each study depends on its proximity to the central construct. The findings are not interpreted as interchangeable: some describe performance, others describe measurement properties, student experience, accessibility, or model performance. The comparison allows recognizing convergences without losing the specificity of each result and constitutes the basis for the discussion of mechanisms and implementation conditions.

Integration of findings

The thematic integration of the corpus brought together 24 studies published during 1984–2026. The temporal distribution reveals an active and heterogeneous research field, with works corresponding to different stages in the maturation of the problem. Some focus on demonstrating feasibility or association, whereas others advance toward validation, comparison of strategies, or analysis of implementation conditions.

This temporal breadth allows observing not only favorable results, but also shifts in how the quality of evaluation is conceptualized and in the type of evidence considered necessary to support educational decisions.

The methodological characterization confirms that adaptive assessment using artificial intelligence to measure learning outcomes and student competencies does not constitute a homogeneous object of study. The set of studies includes designs with distinct purposes and variable levels of control over study conditions. This diversity broadens the scope of the review, but it also requires interpreting findings according to the function of each design. Intervention studies report on observed changes under defined conditions; instrumental studies provide evidence on measurement; observational analyses describe associations; and previous reviews allow recognizing field trends without replacing the evaluation of primary studies.

The samples and corpora described in the studies exhibit considerable variability. This variation is relevant because the transferability of the results depends on the size, composition, and context of the participant samples, as well as on the academic discipline and the technological environment. Overall, the evidence suggests that the most interpretable results are those that specify the population, the evaluative task, and the mechanism by which the intervention or instrument is expected to produce the intended effect. When any of these elements remains poorly defined, the usefulness of the finding for institutional decisions diminishes.

Table 3. Recurrent axes identified in the synthesis

Axis	Frequency of mention
equity	3
competition	3
reliability	2
validity	2
yield	1
adaptation	1
feedback	1
accessibility	1

Note. Descriptive frequency of terms/axes in the methodological and results fields of the matrix.

The recurrent axes should not be read as a measure of effect, but as an indicator of thematic concentration within the corpus. Their distribution helps to recognize which dimensions have received greater attention and which remain less developed. This reading is useful for guiding future research, because a frequently mentioned topic may still lack comparative or longitudinal designs, while a less frequent axis may represent an emerging line with high explanatory potential.

DISCUSSION

The synthesis supports the claim that the problem of adaptive assessment using artificial intelligence to measure learning outcomes and student competencies should be interpreted within a framework of adaptation and personalization, rather than in terms of a simplistic opposition between effective and ineffective technology. The reviewed studies show that the performance of a strategy depends on the interaction among assessment purpose, student characteristics, task design, measurement quality, and the institutional capacity to use the information generated. This conclusion shifts attention from the isolated tool to the assessment system in which it is embedded.

An important consequence is that technological innovation does not eliminate the classic problems of validity, reliability, fairness, and utility. On the contrary, it can make them more visible by expanding the scale and speed of decisions. The reviewed evidence suggests that the best implementations are those that make criteria explicit, preserve feedback opportunities, and maintain human review mechanisms when the consequences of a decision are relevant. In this sense, digitalization should be understood as a transformation of the evaluative process and not as an automatic substitution of academic judgment.

The observed heterogeneity also explains why it is not appropriate to reduce the field to a single effect size. The studies differ in constructs, duration, instruments, disciplines, samples, and metrics. This variation is not only a limitation; it also provides information about the contexts in which a strategy works and about the conditions that should be considered before transferring it.

Comparative synthesis allows patterns of consistency to be recognized without erasing differences that are essential for educational decision-making.

From a methodological perspective, the corpus reveals a transition from studies focused on adoption and perception toward designs that seek to demonstrate measurement quality, impact on outcomes, and sustainability of the intervention. However, gaps persist in longitudinal comparisons, replication across institutions, subgroup analyses, and documentation of unintended effects. The future agenda should prioritize designs that connect learning outcomes with implementation processes and that inform not only whether an intervention works but also for whom, under what conditions, and with what pedagogical or organizational costs.

The main contribution of this review is to organize the evidence so that the results can be used for academic decisions without overextending their scope. The convergence among studies provides arguments for adopting promising practices, but methodological diversity requires maintaining a logic of continuous evaluation. Institutions should treat each implementation as an intervention amenable to monitoring, with indicators defined before its deployment and mechanisms to review outcomes, biases, student experience, and coherence with curricular objectives.

The examined contribution (1) offers a useful point of contrast for this interpretation because it documents that the CAT+memory system improved performance and engagement in practices and materials compared with the other conditions. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. The systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

The examined contribution (2) offers a useful point of contrast for this interpretation because it documents that the experimental group outperformed the control in programming and mathematics; greater interaction was associated with progress, and 80% valued the adaptive feedback. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. The systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

David (3) offers a useful point of contrast for this interpretation because his study documents and establishes foundations for the use of IRT/CAT for adaptive educational decisions. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. The systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

The research (4) offers a useful point of contrast for this interpretation because it documents evidence suggesting improvements in learning and adaptive support, with a need for more robust comparative studies and implementation evaluation. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. The systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

The work (5) offers a useful point of contrast for this interpretation because it documents that personalization and feedback are the most frequent mechanisms; gaps persist in longitudinal evaluation, equity, and external validation. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. The systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

The synthesized evidence (6) offers a useful point of contrast for this interpretation because it documents that ITS show consistent benefits when they combine student modeling, immediate feedback, and adaptation; the evaluation is heterogeneous. This finding acquires greater significance when compared with the rest of the corpus: it does not act as self-sufficient evidence, but rather as an observation that helps to specify mechanisms, limits, and conditions of application. The systematic reading thus makes it possible to avoid 2 frequent extremes: excessive generalization from a single study and the loss of contextual information when synthesizing heterogeneous results.

Design Recommendations

The practical implications of this review must lead to decisions made gradually. For adaptive assessment using artificial intelligence to measure learning outcomes and student competencies, responsible implementation requires defining the construct to be measured in advance, establishing success indicators, identifying potentially affected groups, and providing for review mechanisms. The available evidence supports strategies that integrate innovation with teacher support and continuous evaluation of results.

At the institutional level, adoption should be accompanied by data governance protocols, transparency criteria, and communication mechanisms with students and faculty. These conditions enhance the interpretability of the results and reduce the risk of using indicators outside the context for which they were designed. The evaluation of a tool should consider both technical performance and its pedagogical and distributional consequences.

For research, the next step is to more effectively connect efficacy studies with implementation studies. It is necessary to know not only whether a strategy produces changes, but also which components explain those changes, what resources it requires, and how it performs in diverse populations. This orientation would allow evidence to be accumulated more coherently and reduce the fragmentation that currently characterizes part of the literature.

CONCLUSIONS

The systematic review shows that adaptive assessment using artificial intelligence to measure learning outcomes and student competencies is a field with sufficient evidence to identify patterns, but it remains heterogeneous in designs, populations, and metrics. The central finding is that the quality of decisions depends less on the isolated presence of a technology or instrument than on its integration with learning objectives, explicit criteria, and monitoring procedures. The perspective of adaptation and personalization allows us to understand why similar interventions can produce different results and why the evaluation of effectiveness must simultaneously consider outcomes, conditions, and quality of measurement.

From an applied perspective, higher education institutions should adopt a gradual implementation approach, with controlled testing, documentation of results, and periodic review of effects on different student groups. The evidence supports the use of digital strategies when there is pedagogical alignment and when the information generated is used to improve teaching and learning. It also indicates the need to avoid automatic decisions based on single indicators, especially when the academic consequences are relevant or when there are differences in access, technological experience, or support needs.

The research agenda should move toward longitudinal studies, multicenter comparisons, and designs that document mechanisms and differential effects with greater precision. It is also necessary to improve the comparability of indicators and transparency in the description of instruments, algorithms, and intervention procedures. These priorities will make it possible to transform a field characterized by rapid innovation into a more cumulative, replicable evidence base useful for university policies. The review provides a structure for this advancement by separating consistent findings, implementation conditions, and gaps that still require research.

REFERENCES

1. Albert C. M. Yang; Brendan Flanagan; Hiroaki Ogata. Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning. *Computers and Education: Artificial Intelligence*. 2022;3:100104. <https://doi.org/10.1016/j.caeai.2022.100104>
2. William Villegas-Ch; Diego Buenano-Fernandez; Alexandra Maldonado Navarro; Aracely Mera-Navarrete. Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact and effectiveness of personalized feedback. *Smart Learning Environments*. 2025;12:41. <https://doi.org/10.1186/s40561-025-00389-y>
3. David J. Weiss; G. Gage Kingsbury. Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*. 1984;21(4):361-375. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>
4. Artificial intelligence in nursing education: A scoping review of intelligent tutoring systems. *Nurse Education Today*. 2026. <https://doi.org/10.1016/j.nedt.2026.107056>
5. A Systematic Literature Review of AI-Driven Intelligent Tutoring Systems in Engineering Education. *IEEE Access*. 2025. <https://doi.org/10.1109/ACCESS.2025.3626473>
6. Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*. 2019. <https://doi.org/10.1080/10494820.2018.1558257>
7. Christian Becker; Lukas Valentin Hoppe; Stefan Plappert; Kevin Herrmann; Paul Christoph Gembarski; Roland Lachmayer. Development of an intelligent tutoring system for design education: personalised learning support in engineering. *Proceedings of the Design Society*. 2025;5:1375-1384. <https://doi.org/10.1017/pds.2025.10151>
8. Matt Grenander; Robert Belfer; Ekaterina Kochmar; Iulian V. Serban; François St-Hilaire; Jackie C. K. Cheung. Deep Discourse Analysis for Generating Personalized Feedback in Intelligent Tutor Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;35(17):15534-15544. <https://doi.org/10.1609/aaai.v35i17.17829>
9. Effect of Non-mandatory Use of an Intelligent Tutoring System on Students' Learning. *Artificial Intelligence in Education*. 2020. https://doi.org/10.1007/978-3-030-52237-7_31
10. Brody Hannan; Rebecca Eynon. Widening the Digital Divide: The mediating role of Intelligent Tutoring Systems in the relationship between rurality, socioeducational advantage, and mathematics learning outcomes. *Computers & Education*. 2025;233:105312. <https://doi.org/10.1016/j.compedu.2025.105312>
11. Ying Fang; Anne Lippert; Zhiqiang Cai; Su Chen; Jan C. Frijters; Daphne Greenberg; Arthur C. Graesser. Patterns of Adults with Low Literacy Skills Interacting with an Intelligent Tutoring System. *International Journal of Artificial Intelligence in Education*. 2022;32:297-322. <https://doi.org/10.1007/s40593-021-00266-y>
12. Wenting Ma; Olusola O. Adesope; John C. Nesbit; Qing Liu. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*. 2014;106(4):901-918. <https://doi.org/10.1037/a0037123>
13. Hamida Al-Maamari; Manjit Singh Sidhu; Shaik Mazhar Hussain; Abbas M. Al-Ghaili; Kirandeep Kaur Sidhu. Accessible ICT-Enabled Examination Technologies for Students with Special Educational Needs in Higher Education: A Technical Evaluation of Usability, Reliability, Accommodation Quality, and Assessment Fairness. *International Journal of Special Education*. 2026;41(17s):635-654.
14. PyTutor: An intelligent tutoring system for Python programming. *Computers and Education: Artificial Intelligence*. 2024. <https://doi.org/10.1016/j.caeai.2024.100309>
15. Agustín Mejías-Acosta; Mayra D'Armas Regnault; Eduardo Vargas-Cano; Jesennia Cárdenas-Cobo; Cristian Vidal-Silva. Assessment of digital competencies in higher education students: development and validation of a measurement scale. *Frontiers in Education*. 2024. <https://doi.org/10.3389/feduc.2024.1497376>
16. Designing an intelligent tutoring system for computer programming in the Pacific. *Education and Information Technologies*. 2022. <https://doi.org/10.1007/s10639-021-10882-9>
17. Xinghua Wang; Zhuo Wang. Supporting digitally enhanced learning through measurement in higher education: Development and validation of a university students' digital competence scale. *Journal of Computer Assisted Learning*. 2021. <https://doi.org/10.1111/jcal.12546>
18. Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation*. 2021. <https://doi.org/10.1016/j.stueduc.2021.101030>
19. Paul Nisbet. Accessible digital assessments for students with disabilities: specification, formats and implementation in schools. *Journal of Assistive Technologies*. 2012. <https://doi.org/10.1108/17549451211234984>
20. Promoting student engagement with GPTutor: An intelligent tutoring system powered by generative AI. *International Journal of Educational Technology in Higher Education*. 2025. <https://doi.org/10.1186/s41239-025-00571-9>
21. Effectiveness of intelligent tutoring systems for different programming experience levels. *Discover Education*. 2024. <https://doi.org/10.1007/s44217-024-00385-3>
22. Joanna Tai; Paige Mahoney; Rola Ajjaw; et al. How are examinations inclusive for students with disabilities in higher education? A sociomaterial analysis. *Assessment & Evaluation in Higher Education*. 2023. <https://doi.org/10.1080/02602938.2022.2077910>
23. Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*. 2016. <https://doi.org/10.3102/0034654315581420>
24. Microlearning in Teacher Education: Effects on Pre-service Teachers' Digital Self-efficacy and Digital Competence. *Electronic Journal of e-Learning*. 2026. <https://doi.org/10.34190/ejel.24.3.4790>

FUNDING

None.

CONFLICT OF INTEREST

There are none.

AUTHORSHIP CONTRIBUTIONS

Conceptualization: Segundo Alexander Rosado Martínez

Data curation: Segundo Alexander Rosado Martínez

Formal analysis: Segundo Alexander Rosado Martínez

Research: Segundo Alexander Rosado Martínez

Methodology: Segundo Alexander Rosado Martínez

Supervision: Segundo Alexander Rosado Martínez

Validation: Segundo Alexander Rosado Martínez

Visualization: Segundo Alexander Rosado Martínez

Writing – original draft: Segundo Alexander Rosado Martínez

Writing – review and editing: Segundo Alexander Rosado Martínez