



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

Recibido: 19-08-2025 | Revisado: 01-09-2025 | Aceptado: 19-12-2025 | Publicado: 28-12-2025

Autor de correspondencia: 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.

INTRODUCCIÓN

La investigación universitaria sobre evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles ha adquirido una relevancia creciente porque las decisiones de evaluación ya no dependen exclusivamente de instrumentos estáticos ni de intercambios presenciales. La digitalización ha ampliado la cantidad de evidencias disponibles sobre desempeño, interacción y aprendizaje, pero también ha incrementado la necesidad de distinguir entre innovaciones técnicamente atractivas y prácticas capaces de sostener inferencias educativas válidas. Esta revisión aborda ese problema desde una perspectiva integradora, interesada tanto en los efectos observados como en las condiciones metodológicas que permiten interpretar dichos efectos sin confundir disponibilidad tecnológica con calidad evaluativa.

El núcleo conceptual del manuscrito se organiza alrededor de adaptación y personalización. Esta elección evita tratar la literatura como una simple colección de resultados y permite examinar cómo cada estudio define el fenómeno, qué población observa, qué instrumentos emplea y qué tipo de conclusión considera legítima. La heterogeneidad de contextos universitarios obliga a prestar atención a la relación entre diseño, implementación y evidencia, porque una misma estrategia puede producir beneficios bajo determinadas condiciones y resultados modestos cuando cambian el soporte docente, la disciplina, la experiencia digital o la forma de retroalimentación.

La revisión sistemática resulta pertinente porque el campo combina estudios experimentales, validaciones instrumentales, análisis observacionales, desarrollos tecnológicos y síntesis previas. Esa diversidad dificulta obtener conclusiones sólidas mediante una lectura fragmentaria. En lugar de asumir que todos los trabajos responden a una misma pregunta, el presente artículo reconstruye patrones de convergencia y divergencia, identifica qué resultados se repiten con mayor consistencia y separa las afirmaciones directamente sustentadas por los estudios de aquellas que todavía requieren validación adicional. El propósito es ofrecer una lectura útil para investigación, docencia y gestión universitaria.

METODOLOGÍA

Yang et al¹ aportaron evidencia publicada en 2022 mediante experimento comparativo de 7 semanas. El estudio trabajó con Estudiantes de un curso universitario de programación, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el sistema CAT+memoria mejoró rendimiento y engagement en prácticas y materiales frente a las otras condiciones. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La contribución examinada² aportó evidencia publicada en 2025 mediante cuasi-experimental + analítica. El estudio trabajó con Grupos experimental/control en STEM, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el grupo experimental superó al control en programación y matemáticas; mayor interacción se relacionó con progreso y 80% valoró el feedback adaptativo. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

David³ aportaron evidencia publicada en 1984 mediante artículo metodológico fundacional. El estudio trabajó con Aplicaciones de CAT en problemas educativos, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que establece fundamentos de uso de IRT/CAT para decisiones educativas adaptativas. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La investigación⁴ aportó evidencia publicada en 2026 mediante scoping review jbi/prisma-scr. El estudio trabajó con 9 estudios, lo que permite situar el alcance de sus inferencias y valorar

su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la evidencia sugiere mejoras en aprendizaje y soporte adaptativo, con necesidad de estudios comparativos más robustos y evaluación de implementación. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El trabajo⁵ aportó evidencia publicada en 2025 mediante revisión sistemática. El estudio trabajó con Corpus de ITS en ingeniería, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la personalización y el feedback son los mecanismos más frecuentes; persisten vacíos en evaluación longitudinal, equidad y validación externa. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La evidencia sintetizada (6) aportó evidencia publicada en 2019 mediante revisión sistemática. El estudio trabajó con Corpus de ITS, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que los ITS muestran beneficios consistentes cuando combinan modelado del estudiante, feedback inmediato y adaptación; la evaluación es heterogénea. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La contribución examinada⁷ aportó evidencia publicada en 2025 mediante diseño basado en investigación + evaluación. El estudio trabajó con Estudiantes de ingeniería/diseño, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el ITS permitió recomendaciones individualizadas y apoyo al proceso de diseño, aunque requiere integración pedagógica cuidadosa. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El estudio⁸ aportó evidencia publicada en 2021 mediante desarrollo/evaluación computacional. El estudio trabajó con Datos de diálogos educativos, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el análisis profundo del discurso permitió identificar estados de aprendizaje y apoyar retroalimentación adaptada al estudiante. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La investigación⁹ aportó evidencia publicada en 2020 mediante estudio de uso naturalista/comparativo. El estudio trabajó con Estudiantes con uso voluntario del ITS, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el beneficio depende de adopción y patrones de uso; la disponibilidad del tutor no garantiza que todos los estudiantes lo utilicen suficientemente. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El trabajo¹⁰ aportó evidencia publicada en 2025 mediante estudio cuantitativo explicativo. El estudio trabajó con Estudiantes en entorno con ITS, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el efecto de ITS puede ser desigual cuando existen brechas de acceso, habilidades digitales o uso efectivo; la personalización no elimina automáticamente inequidades. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra

que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La evidencia sintetizada¹¹aportó evidencia publicada en 2022 mediante analítica de interacción. El estudio trabajó con Adultos con baja alfabetización, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que se identificaron perfiles de interacción diferenciados, útiles para adaptar feedback y apoyo a necesidades individuales. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La contribución examinada¹²aportó evidencia publicada en 2014 mediante meta-análisis. El estudio trabajó con Estudios de resultados de ITS, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que los ITS presentan efectos positivos sobre aprendizaje frente a varios controles; el beneficio varía por dominio y tipo de implementación. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

Hamida et al.¹³aportaron evidencia publicada en 2026 mediante evaluación técnica de accesibilidad. El estudio trabajó con Estudiantes con necesidades educativas especiales en educación superior, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que identifica que accesibilidad de evaluación digital exige compatibilidad con tecnologías de apoyo, autonomía e integridad bajo restricciones de examen. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La investigación¹⁴aportó evidencia publicada en 2024 mediante experimento longitudinal en aula. El estudio trabajó con ITS n=35; control n=36, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el ITS produjo ventajas de aprendizaje, especialmente en estudiantes con menor conocimiento previo; se señaló posible sobredependencia. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

Agustín et al.¹⁵aportaron evidencia publicada en 2024 mediante desarrollo y validación de escala. El estudio trabajó con Estudiantes de una universidad venezolana, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que se obtuvo un modelo de 22 variables en cuatro dimensiones con propiedades psicométricas satisfactorias. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La evidencia sintetizada¹⁶aportó evidencia publicada en 2022 mediante diseño y evaluación de its. El estudio trabajó con Estudiantes de programación, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el sistema mostró viabilidad para feedback y andamiaje de programación, con ajustes necesarios a contextos y perfiles estudiantiles. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

Xinghua et al.¹⁷aportaron evidencia publicada en 2021 mediante desarrollo y validación

psicométrica. El estudio trabajó con Estudiantes universitarios; tamaño no visible en el fragmento recuperado, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el estudio desarrolló una medida específica para competencia digital universitaria con evidencia de validez para aprendizaje digitalmente mejorado. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

No18aportaron evidencia publicada en 2021 mediante revisión sistemática. El estudio trabajó con 26 estudios (2010–2019), lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la evaluación auténtica se asoció con mayor engagement/satisfacción y competencias de comunicación, colaboración, pensamiento crítico y resolución de problemas. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

Paul19aportaron evidencia publicada en 2012 mediante estudio de diseño e implementación. El estudio trabajó con Entrevistas a docentes/estudiantes y pilotos 2005–2007, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que pDF accesible permitió mayor independencia y menor coste que lector/escriba; las solicitudes de exámenes digitales aumentaron cinco veces 2008–2011. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El trabajo20aportó evidencia publicada en 2025 mediante métodos mixtos secuenciales. El estudio trabajó con Estudiantes universitarios usuarios de GPTutor, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que gPTutor favoreció engagement y apoyo personalizado; la calidad de interacción y diseño de prompts/tutoría condicionó la experiencia. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La evidencia sintetizada21aportó evidencia publicada en 2024 mediante comparativo por nivel de experiencia. El estudio trabajó con Estudiantes de programación; grupos por experiencia, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la efectividad varió según experiencia previa, confirmando que la personalización debe considerar el perfil del aprendiz. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

Joanna et al.22aportaron evidencia publicada en 2023 mediante cualitativo sociomaterial. El estudio trabajó con 40 estudiantes con discapacidad de dos universidades, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que exámenes online, uso de tecnología, más tiempo y formato open-book aumentaron inclusión para muchos; la flexibilidad y aceptación del personal fueron claves. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

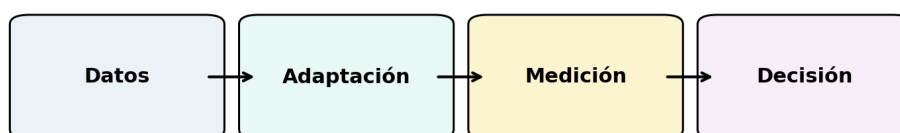
El estudio23aportó evidencia publicada en 2016 mediante meta-análisis. El estudio trabajó con Estudios comparativos de ITS, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que los ITS superan en promedio a varias formas de instrucción convencional/computarizada, aunque

el efecto depende del comparador y diseño. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La investigación²⁴ aportó evidencia publicada en 2026 mediante pre-post de gran muestra. El estudio trabajó con Pre $n=1,437$; post $n=871$, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que se observaron mejoras significativas en todas las competencias digitales evaluadas y reducción parcial de brechas entre grupos; aumentó la autoeficacia digital. Este antecedente es relevante para la perspectiva de adaptación y personalización porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

Figura 1. Estructura conceptual de la revisión

Estructura conceptual de la revisión



Nota. Elaboración propia a partir de los ejes analíticos definidos para el manuscrito.

La figura organiza el razonamiento del artículo como una secuencia conceptual y no como una plantilla universal. Los componentes representados corresponden al problema específico de esta revisión y muestran el tránsito desde los insumos o condiciones iniciales hasta la interpretación educativa. Su utilidad consiste en hacer visible la relación entre constructos que, en los estudios individuales, suelen aparecer fragmentados entre secciones metodológicas y de resultados. Esta representación orienta la síntesis posterior y facilita distinguir los elementos que actúan como condiciones, mecanismos o resultados dentro del corpus.

Preguntas de revisión

Pregunta 1. ¿Qué patrones de evidencia permiten comprender la dimensión 1 de adaptación y personalización en relación con evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 2. ¿Qué patrones de evidencia permiten comprender la dimensión 2 de adaptación y personalización en relación con evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 3. ¿Qué patrones de evidencia permiten comprender la dimensión 3 de adaptación y personalización en relación con evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 4. ¿Qué patrones de evidencia permiten comprender la dimensión 4 de adaptación y personalización en relación con evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 5. ¿Qué patrones de evidencia permiten comprender la dimensión 5 de adaptación y personalización en relación con evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles y qué condiciones metodológicas o contextuales modifican su interpretación?

o contextuales modifican su interpretación?

MÉTODOS

Estrategia PRISMA y extracción

Se desarrolló una revisión sistemática orientada a sintetizar evidencia académica pertinente para el tema del manuscrito. La unidad de análisis fue el estudio o referencia académica con fuente trazable, metodología identificable y resultados recuperables. El corpus académico consolidado de trabajo estuvo compuesto por 135 registros previamente reunidos en las matrices de evidencia. Sobre ese conjunto se aplicó una selección temática y documental para establecer el corpus final. El procedimiento se organizó conforme a la lógica de identificación, cribado, elegibilidad e inclusión de PRISMA, adaptada al carácter consolidado de las fuentes suministradas.

Los criterios de inclusión exigieron relación directa con evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles, correspondencia con educación superior o con instrumentos aplicables a ese nivel, descripción suficiente del método y disponibilidad de resultados interpretables. Se excluyeron 111 registros por menor pertinencia respecto de la pregunta específica, por tratar poblaciones o desenlaces periféricos, o por no aportar evidencia metodológica suficiente para la síntesis focal. El corpus final quedó constituido por 24 estudios. La selección se efectuó sobre la información documentada en las matrices y se evitó atribuir a una base bibliográfica particular conteos que no estuvieran respaldados por los archivos.

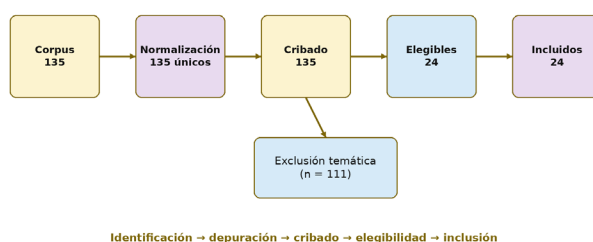
La extracción de información se estructuró mediante una matriz con autores, año, título, fuente, DOI o enlace trazable, procedencia, tipo de estudio, muestra o corpus, metodología resumida, resultados principales, nivel de evidencia y pertinencia. Esta organización permitió comparar trabajos con diseños heterogéneos sin reducirlos a una única métrica. La síntesis combinó caracterización descriptiva del corpus y análisis temático de los resultados, prestando atención a convergencias, contradicciones, condiciones de implementación y límites de generalización.

La evaluación de la evidencia se centró en la suficiencia de la información recuperada para responder la pregunta de revisión. Se consideraron de forma conjunta la claridad del diseño, la explicitación de la muestra o corpus, la correspondencia entre método y resultados y la trazabilidad de la fuente. No se realizó metaanálisis cuando la heterogeneidad de desenlaces, instrumentos, poblaciones o métricas impedía una combinación cuantitativa responsable. En esos casos se priorizó una síntesis comparativa que preserve las diferencias entre estudios y evita transformar resultados no equivalentes en un único estimador.

Para fortalecer la transparencia, los resultados se organizaron en tablas de caracterización y síntesis, mientras que las figuras representan el flujo de selección y la distribución temporal del corpus. Cada tabla fue construida a partir de los campos de la matriz suministrada. Las interpretaciones posteriores se elaboraron sobre patrones observables en el conjunto y no sobre supuestos añadidos.

Figura 2. Flujo de selección de estudios según la lógica PRISMA

Flujo de selección de estudios — revisión 3



Fuente: elaboración propia a partir de la matriz de evidencia de la revisión.

Nota. El flujo representa 135 registros académicos normalizados, sin duplicados en el corpus consolidado; 111 registros no cumplieron el umbral de pertinencia temática y 24 estudios integraron la síntesis final.

Este criterio es especialmente importante en revisiones educativas, donde la variación entre disciplinas, tecnologías, instrumentos y contextos institucionales puede modificar sustancialmente la magnitud y el significado de los hallazgos.

El flujo documenta la depuración y selección de la evidencia. Los 135 registros del corpus consolidado se encontraban normalizados por DOI o título, por lo que no se identificaron duplicados en esta etapa. El cribado temático excluyó 111 registros por insuficiente correspondencia con la pregunta específica de la revisión, mientras que 24 estudios conservaron evidencia metodológica y resultados suficientemente recuperables para integrar la síntesis. La secuencia permite distinguir la identificación documental del juicio de elegibilidad y evita interpretar la reducción del corpus como una pérdida arbitraria de información.

RESULTADOS

El corpus final reunió 24 estudios publicados durante 1984–2026. La distribución temporal muestra un campo de investigación activo y heterogéneo, con trabajos que responden a distintas fases de maduración del problema. Algunos se concentran en demostrar viabilidad o asociación, mientras otros avanzan hacia validación, comparación de estrategias o análisis de condiciones de implementación. Esta amplitud temporal permite observar no solo resultados favorables, sino también desplazamientos en la forma de conceptualizar la calidad de la evaluación y en el tipo de evidencia que se considera necesaria para respaldar decisiones educativas.

La caracterización metodológica confirma que evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles no constituye un objeto homogéneo. En el conjunto aparecen diseños con finalidades distintas y niveles variables de control sobre las condiciones de estudio. Esta diversidad amplía el alcance de la revisión, aunque también obliga a interpretar los hallazgos de acuerdo con la función de cada diseño. Los trabajos de intervención informan sobre cambios observados bajo condiciones definidas; los estudios instrumentales aportan evidencia sobre medición; los análisis observacionales describen asociaciones; y las revisiones previas permiten reconocer tendencias del campo sin sustituir la evaluación de los estudios primarios.

Las muestras y corpus descritos en los estudios presentan una variabilidad considerable. Esa variación es relevante porque la transferibilidad de los resultados depende del tamaño, composición y contexto de los participantes, así como de la disciplina académica y del entorno tecnológico. En conjunto, la evidencia sugiere que los resultados más interpretables son aquellos que explicitan la población, la tarea evaluativa y el mecanismo mediante el cual la intervención o instrumento debería producir el efecto esperado. Cuando alguno de esos elementos permanece poco definido, la utilidad del hallazgo para decisiones institucionales disminuye.

Tabla 1. Caracterización metodológica del corpus incluido

Año	Diseño/metodología	Muestra o corpus	Evidencia
2022	Experimento comparativo de 7 semanas	Estudiantes de un curso universitario de programación	Alta
2025	Cuasi-experimental + analítica	Grupos experimental/control en STEM	Media
1984	Artículo metodológico fundacional	Aplicaciones de CAT en problemas educativos	Alta
2026	Scoping review JBI/PRISMA-ScR	9 estudios	Media
2025	Revisión sistemática	Corpus de ITS en ingeniería	Media
2019	Revisión sistemática	Corpus de ITS	Media
2025	Diseño basado en investigación + evaluación	Estudiantes de ingeniería/diseño	Media
2021	Desarrollo/evaluación computacional	Datos de diálogos educativos	Media
2020	Estudio de uso naturalista/comparativo	Estudiantes con uso voluntario del ITS	Media
2025	Estudio cuantitativo explicativo	Estudiantes en entorno con ITS	Media
2022	Analítica de interacción	Adultos con baja alfabetización	Media
2014	Meta-análisis	Estudios de resultados de ITS	Media

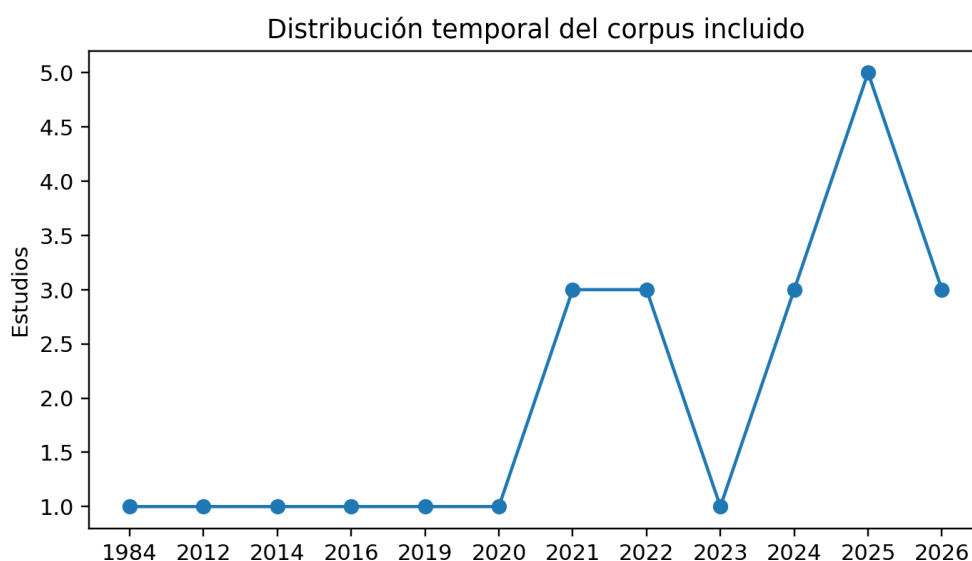
Nota. Datos extraídos de la matriz de evidencia suministrada para esta revisión.

La síntesis de resultados revela un patrón central vinculado con adaptación y personalización: los efectos o propiedades favorables no aparecen como atributos automáticos de la tecnología o

del instrumento, sino como resultados condicionados por diseño, implementación, calidad de datos y alineación pedagógica. Los estudios convergen en la importancia de articular la herramienta con objetivos de aprendizaje claros, criterios de evaluación comprensibles y mecanismos de seguimiento. Las divergencias se concentran en la magnitud de los efectos, la estabilidad entre contextos y la capacidad de sostener resultados cuando cambian las condiciones institucionales.

La tabla evidencia la diversidad metodológica del corpus y permite observar que los estudios no ofrecen el mismo tipo de inferencia. La presencia de diseños con objetivos distintos exige interpretar cada resultado según la pregunta que el estudio puede responder. Esta diferenciación evita comparar directamente indicadores que cumplen funciones diferentes y ayuda a identificar dónde existe convergencia real. También muestra que la solidez de una conclusión depende de la combinación entre claridad metodológica, descripción de la muestra y suficiencia de la evidencia recuperada.

Figura 3. Distribución temporal de los estudios incluidos



Nota. Frecuencias calculadas a partir del año de publicación registrado en la matriz de evidencia.

La distribución temporal permite situar la intensidad reciente de la producción científica y reconocer si el campo se encuentra en expansión, consolidación o transición metodológica. Más que interpretar el número de publicaciones como una medida de calidad, la figura sirve para contextualizar los cambios en preguntas, tecnologías e instrumentos observados en el corpus. La lectura conjunta con la caracterización metodológica permite identificar si los trabajos más recientes amplían la evidencia mediante diseños más robustos o si continúan predominando aproximaciones exploratorias.

Resultados sobre medición y adaptación

Otro patrón transversal es la relevancia de la calidad de la evidencia. Los estudios con procedimientos explícitos permiten distinguir mejor entre mejora real, percepción favorable y simple facilidad de uso. En cambio, cuando el resultado se apoya principalmente en autoinforme o en indicadores de corto plazo, la interpretación requiere mayor cautela. La revisión muestra que la utilidad práctica de una estrategia no depende únicamente de que produzca una diferencia estadística o una valoración positiva, sino de que la medición sea coherente con el constructo y de que la intervención pueda integrarse de manera sostenible en procesos educativos reales.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: El sistema CAT+memoria mejoró rendimiento y engagement en prácticas y materiales frente a las otras condiciones. El grupo experimental superó al control en programación y matemáticas; mayor interacción se relacionó con progreso y 80% valoró el feedback adaptativo. Establece fundamentos de uso de IRT/CAT para decisiones educativas adaptativas.

La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: La evidencia sugiere mejoras en aprendizaje y soporte adaptativo, con necesidad de estudios comparativos más robustos y evaluación de implementación. La personalización y el feedback son los mecanismos más frecuentes; persisten vacíos en evaluación longitudinal, equidad y validación externa. Los ITS muestran beneficios consistentes cuando combinan modelado del estudiante, feedback inmediato y adaptación; la evaluación es heterogénea. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: El ITS permitió recomendaciones individualizadas y apoyo al proceso de diseño, aunque requiere integración pedagógica cuidadosa. El análisis profundo del discurso permitió identificar estados de aprendizaje y apoyar retroalimentación adaptada al estudiante. El beneficio depende de adopción y patrones de uso; la disponibilidad del tutor no garantiza que todos los estudiantes lo utilicen suficientemente. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: El efecto de ITS puede ser desigual cuando existen brechas de acceso, habilidades digitales o uso efectivo; la personalización no elimina automáticamente inequidades. Se identificaron perfiles de interacción diferenciados, útiles para adaptar feedback y apoyo a necesidades individuales. Los ITS presentan efectos positivos sobre aprendizaje frente a varios controles; el beneficio varía por dominio y tipo de implementación. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: Identifica que accesibilidad de evaluación digital exige compatibilidad con tecnologías de apoyo, autonomía e integridad bajo restricciones de examen. El ITS produjo ventajas de aprendizaje, especialmente en estudiantes con menor conocimiento previo; se señaló posible sobredependencia. Se obtuvo un modelo de 22 variables en cuatro dimensiones con propiedades psicométricas satisfactorias. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: El sistema mostró viabilidad para feedback y andamiaje de programación, con ajustes necesarios a contextos y perfiles estudiantiles. El estudio desarrolló una medida específica para competencia digital universitaria con evidencia de validez para aprendizaje digitalmente mejorado. La evaluación auténtica se asoció con mayor engagement/satisfacción y competencias de comunicación, colaboración, pensamiento crítico y resolución de problemas. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: PDF accesible permitió mayor independencia y menor coste que lector/escritura; las solicitudes de exámenes digitales aumentaron cinco veces 2008–2011. GPTutor favoreció engagement y apoyo personalizado; la calidad de interacción y diseño de prompts/tutoría condicionó la experiencia. La efectividad varió según experiencia previa, confirmando que la personalización debe considerar el perfil del aprendiz. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: Exámenes online, uso de tecnología, más tiempo y formato open-book aumentaron inclusión para muchos; la flexibilidad y aceptación del personal fueron claves. Los ITS superan en promedio a varias formas de instrucción convencional/computarizada, aunque el efecto depende del comparador y diseño. Se observaron mejoras significativas en todas las competencias digitales evaluadas y reducción parcial de brechas entre grupos; aumentó la autoeficacia digital. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

Tabla 2. Síntesis de hallazgos relevantes

Estudio	Hallazgo recuperado	Pertinencia
Adaptive formative assessment system based on computerized adaptive testing and the learni	El sistema CAT+memoria mejoró rendimiento y engagement en prácticas y materiales frente a las otras condiciones.	Incluido por pertinencia directa con la
Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact	El grupo experimental superó al control en programación y matemáticas; mayor interacción se relacionó con progreso y 80% valoró el feedback adaptativo	Incluido por pertinencia directa con la
Application of Computerized Adaptive Testing to Educational Problems	Establece fundamentos de uso de IRT/CAT para decisiones educativas adaptativas.	Incluido por pertinencia directa con la
Artificial intelligence in nursing education: A scoping review of intelligent tutoring sys	La evidencia sugiere mejoras en aprendizaje y soporte adaptativo, con necesidad de estudios comparativos más robustos y evaluación de implementación.	Incluido por pertinencia directa con la
A Systematic Literature Review of AI-Driven Intelligent Tutoring Systems in Engineering Ed	La personalización y el feedback son los mecanismos más frecuentes; persisten vacíos en evaluación longitudinal, equidad y validación externa.	Incluido por pertinencia directa con la
Intelligent tutoring systems: A systematic review of characteristics, applications, and ev	Los ITS muestran beneficios consistentes cuando combinan modelado del estudiante, feedback inmediato y adaptación; la evaluación es heterogénea.	Incluido por pertinencia directa con la
Development of an intelligent tutoring system for design education: personalised learning	El ITS permitió recomendaciones individualizadas y apoyo al proceso de diseño, aunque requiere integración pedagógica cuidadosa.	Incluido por pertinencia directa con la
Deep Discourse Analysis for Personalized Feedback in Intelligent Tutoring Systems	El análisis profundo del discurso permitió identificar estados de aprendizaje y apoyar retroalimentación adaptada al estudiante.	Incluido por pertinencia directa con la
Effect of Non-mandatory Use of an Intelligent Tutoring System on Students' Learning	El beneficio depende de adopción y patrones de uso; la disponibilidad del tutor no garantiza que todos los estudiantes lo utilicen suficientemente.	Incluido por pertinencia directa con la
Widening the Digital Divide: The mediating role of Intelligent Tutoring Systems in learnin	El efecto de ITS puede ser desigual cuando existen brechas de acceso, habilidades digitales o uso efectivo; la personalización no elimina automáticamente	Incluido por pertinencia directa con la

Nota. Síntesis textual basada exclusivamente en los resultados registrados en la matriz.

La síntesis muestra que la evidencia favorable se distribuye entre distintos tipos de resultados y que la pertinencia de cada estudio depende de su proximidad con el constructo central. Los hallazgos no se interpretan como intercambiables: algunos describen rendimiento, otros propiedades de medición, experiencia estudiantil, accesibilidad o desempeño de modelos. La comparación permite reconocer convergencias sin perder la especificidad de cada resultado y constituye la base para la discusión de mecanismos y condiciones de implementación.

Integración de hallazgos

La integración temática del corpus reunió 24 estudios publicados durante 1984–2026. La distribución temporal muestra un campo de investigación activo y heterogéneo, con trabajos que responden a distintas fases de maduración del problema. Algunos se concentran en demostrar viabilidad o asociación, mientras otros avanzan hacia validación, comparación de estrategias o análisis de condiciones de implementación. Esta amplitud temporal permite observar no solo resultados favorables, sino también desplazamientos en la forma de conceptualizar la calidad de la evaluación y en el tipo de evidencia que se considera necesaria para respaldar decisiones educativas.

La caracterización metodológica confirma que evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles no constituye un objeto homogéneo. En el conjunto aparecen diseños con finalidades distintas y niveles variables de control sobre las condiciones de estudio. Esta diversidad amplía el alcance de la revisión, aunque también obliga a interpretar los hallazgos de acuerdo con la función de cada diseño. Los trabajos de intervención informan sobre cambios observados bajo condiciones definidas; los estudios instrumentales aportan evidencia sobre medición; los análisis observacionales describen asociaciones; y las revisiones previas permiten reconocer tendencias del campo sin sustituir la evaluación de los estudios primarios.

Las muestras y corpus descritos en los estudios presentan una variabilidad considerable. Esa variación es relevante porque la transferibilidad de los resultados depende del tamaño, composición y contexto de los participantes, así como de la disciplina académica y del entorno tecnológico. En conjunto, la evidencia sugiere que los resultados más interpretables son aquellos que explicitan la población, la tarea evaluativa y el mecanismo mediante el cual la intervención o instrumento debería producir el efecto esperado. Cuando alguno de esos elementos permanece poco definido, la utilidad del hallazgo para decisiones institucionales disminuye.

Tabla 3. Ejes recurrentes identificados en la síntesis

Eje	Frecuencia de mención
equidad	3
competencia	3
fiabilidad	2
validez	2
rendimiento	1
adaptación	1
retroalimentación	1
accesibilidad	1

Nota. Frecuencia descriptiva de términos/ejes en los campos metodológicos y de resultados de la matriz.

Los ejes recurrentes no deben leerse como una medida de efecto, sino como un indicador de concentración temática dentro del corpus. Su distribución ayuda a reconocer qué dimensiones han recibido mayor atención y cuáles permanecen menos desarrolladas. Esta lectura resulta útil para orientar futuras investigaciones, porque un tema frecuentemente mencionado puede seguir careciendo de diseños comparativos o longitudinales, mientras que un eje menos frecuente puede representar una línea emergente con alto potencial explicativo.

DISCUSIÓN

La síntesis permite sostener que el problema de evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles debe interpretarse desde una lógica de adaptación y personalización y no desde una oposición simplificada entre

tecnología eficaz e ineficaz.

Los estudios revisados muestran que el desempeño de una estrategia depende de la interacción entre propósito evaluativo, características de los estudiantes, diseño de la tarea, calidad de la medición y capacidad institucional para utilizar la información generada. Esta conclusión desplaza la atención desde la herramienta aislada hacia el sistema de evaluación en el que se inserta.

Una consecuencia importante es que la innovación tecnológica no elimina los problemas clásicos de validez, fiabilidad, equidad y utilidad. Por el contrario, puede hacerlos más visibles al ampliar la escala y velocidad de las decisiones. La evidencia revisada sugiere que las mejores implementaciones son aquellas que hacen explícitos los criterios, preservan oportunidades de retroalimentación y mantienen mecanismos de revisión humana cuando las consecuencias de una decisión son relevantes. En este sentido, la digitalización debe entenderse como una transformación del proceso evaluativo y no como una sustitución automática del juicio académico.

La heterogeneidad observada también explica por qué no resulta adecuado reducir el campo a un único tamaño de efecto. Las investigaciones difieren en constructos, duración, instrumentos, disciplinas, muestras y métricas. Esa variación no constituye únicamente una limitación; también aporta información sobre los contextos en los que una estrategia funciona y sobre las condiciones que deben ser consideradas antes de transferirla. La síntesis comparativa permite reconocer patrones de consistencia sin borrar diferencias que son esenciales para la toma de decisiones educativas.

Desde el punto de vista metodológico, el corpus evidencia una transición desde estudios centrados en adopción y percepción hacia diseños que buscan demostrar calidad de medición, impacto en resultados y sostenibilidad de la intervención. Sin embargo, persisten vacíos relacionados con comparaciones longitudinales, replicación entre instituciones, análisis de subgrupos y documentación de efectos no previstos. La agenda futura debería priorizar diseños que conecten resultados de aprendizaje con procesos de implementación y que informen no solo si una intervención funciona, sino para quién, bajo qué condiciones y con qué costos pedagógicos u organizacionales.

La principal contribución de esta revisión consiste en organizar la evidencia de forma que los resultados puedan utilizarse para decisiones académicas sin sobreextender sus alcances. La convergencia entre estudios ofrece argumentos para adoptar prácticas prometedoras, pero la diversidad metodológica exige preservar una lógica de evaluación continua. Las instituciones deberían tratar cada implementación como una intervención susceptible de seguimiento, con indicadores definidos antes de su despliegue y mecanismos para revisar resultados, sesgos, experiencia estudiantil y coherencia con los objetivos curriculares.

No ofrece un punto de contraste útil para esta interpretación porque su estudio documenta el sistema CAT+memoria mejoró rendimiento y engagement en prácticas y materiales frente a las otras condiciones.. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

La contribución examinada² ofrece un punto de contraste útil para esta interpretación porque su estudio documenta el grupo experimental superó al control en programación y matemáticas; mayor interacción se relacionó con progreso y 80% valoró el feedback adaptativo.. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

David³ ofrece un punto de contraste útil para esta interpretación porque su estudio documenta establece fundamentos de uso de IRT/CAT para decisiones educativas adaptativas.. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

La investigación⁴ ofrece un punto de contraste útil para esta interpretación porque su estudio documenta la evidencia sugiere mejoras en aprendizaje y soporte adaptativo, con necesidad de estudios comparativos más robustos y evaluación de implementación.

Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

El trabajo 5 ofrece un punto de contraste útil para esta interpretación porque su estudio documenta la personalización y el feedback son los mecanismos más frecuentes; persisten vacíos en evaluación longitudinal, equidad y validación externa. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

La evidencia sintetizada 6 ofrece un punto de contraste útil para esta interpretación porque su estudio documenta los ITS muestran beneficios consistentes cuando combinan modelado del estudiante, feedback inmediato y adaptación; la evaluación es heterogénea. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

Recomendaciones de diseño

Las implicaciones prácticas derivadas de la revisión deben traducirse en decisiones graduales. Para evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles, una implementación responsable requiere definir previamente el constructo que se desea medir, establecer indicadores de éxito, identificar grupos potencialmente afectados y prever mecanismos de revisión. La evidencia disponible favorece estrategias que integran la innovación con acompañamiento docente y evaluación continua de resultados.

En el plano institucional, la adopción debería acompañarse de protocolos de gobernanza de datos, criterios de transparencia y mecanismos de comunicación con estudiantes y docentes. Estas condiciones aumentan la interpretabilidad de los resultados y reducen el riesgo de utilizar indicadores fuera del contexto para el que fueron diseñados. La evaluación de una herramienta debe considerar tanto desempeño técnico como consecuencias pedagógicas y distributivas.

Para la investigación, el siguiente paso consiste en conectar mejor los estudios de eficacia con los de implementación. Resulta necesario saber no solo si una estrategia produce cambios, sino qué componentes explican esos cambios, qué recursos demanda y cómo se comporta en poblaciones diversas. Esa orientación permitiría acumular evidencia de forma más coherente y reducir la fragmentación que actualmente caracteriza parte de la literatura.

CONCLUSIONES

La revisión sistemática muestra que evaluación adaptativa mediante inteligencia artificial para medir resultados de aprendizaje y competencias estudiantiles constituye un campo con evidencia suficiente para identificar patrones, pero todavía heterogéneo en diseños, poblaciones y métricas. El resultado central es que la calidad de las decisiones depende menos de la presencia aislada de una tecnología o instrumento que de su integración con objetivos de aprendizaje, criterios explícitos y procedimientos de seguimiento. La perspectiva de adaptación y personalización permite comprender por qué intervenciones similares pueden producir resultados diferentes y por qué la evaluación de efectividad debe considerar simultáneamente resultados, condiciones y calidad de la medición.

En términos aplicados, las instituciones de educación superior deberían adoptar una lógica gradual de implementación, con pruebas controladas, documentación de resultados y revisión periódica de efectos sobre distintos grupos de estudiantes. La evidencia respalda el uso de estrategias digitales cuando existe alineación pedagógica y cuando la información generada se utiliza para mejorar la enseñanza y el aprendizaje. También muestra la necesidad de evitar decisiones automáticas basadas en indicadores únicos, especialmente cuando las consecuencias académicas son relevantes o cuando existen diferencias de acceso, experiencia tecnológica o necesidades de apoyo.

La agenda de investigación debe avanzar hacia estudios longitudinales, comparaciones

multicéntricas y diseños que documenten con mayor precisión mecanismos y efectos diferenciales. También se requiere mejorar la comparabilidad de indicadores y la transparencia en la descripción de instrumentos, algoritmos y procedimientos de intervención. Estas prioridades permitirán transformar un campo caracterizado por rápida innovación en una base de evidencia más acumulativa, replicable y útil para políticas universitarias. La revisión aporta una estructura para ese avance al separar hallazgos consistentes, condiciones de implementación y vacíos que todavía requieren investigación.

Referencias

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/educ.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>

Financiación

Ninguna.

Conflictos de interés

No existen.

Contribución de autoría

Conceptualización: Segundo Alexander Rosado Martínez

Curación de datos: Segundo Alexander Rosado Martínez

Análisis formal: Segundo Alexander Rosado Martínez

Investigación: Segundo Alexander Rosado Martínez

Metodología: Segundo Alexander Rosado Martínez

Supervisión: Segundo Alexander Rosado Martínez

Validación: Segundo Alexander Rosado Martínez

Visualización: Segundo Alexander Rosado Martínez

Redacción – borrador original: Segundo Alexander Rosado Martínez

Redacción – revisión y edición: Segundo Alexander Rosado Martínez