



Use of item response theory in the validation of digital educational instruments

Uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales

Karla Andrea Paredes Rosado¹  

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

Recibido: 11-06-2025 | Revisado: 01-07-2025 | Aceptado: 19-10-2025 | Publicado: 27-10-2025

Autor de correspondencia: Karla Andrea Paredes Rosado 

ABSTRACT

Introduction: Uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales 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: Uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales 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 uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales 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 TRI y validación digital. 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.

Teoría de respuesta al ítem en entornos digitales

No¹aportaron evidencia publicada en 2026 mediante irt 2pl en examen computarizado. El estudio trabajó con 500 estudiantes universitarios de primer año; 10 ítems de álgebra, 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 evalúa equidad y fiabilidad de selección aleatoria de ítems en pruebas computarizadas. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.

Kai et al.²aportaron evidencia publicada en 2017 mediante psicométrico irt. El estudio trabajó con 735 estudiantes universitarios, 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 confirmó estructura unifactorial y evaluó utilidad de la escala de aprendizaje activo mediante IRT. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.

Koen et al.³aportaron evidencia publicada en 2014 mediante validación con teoría de respuesta al ítem. El estudio trabajó con 560 estudiantes; 56 ítems, 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 medida final de 27 ítems fue especialmente precisa en niveles bajo y medio de habilidad. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.

Amal⁴aportaron evidencia publicada en 2023 mediante psicométrico irt. El estudio trabajó con 450 estudiantes universitarios, 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 evaluó propiedades psicométricas de Grit-S con IRT y estructura de consistencia de interés/perseverancia. Este antecedente es relevante para la perspectiva de TRI y validación

digital 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.

Siu-Cheung et al.⁵ aportaron evidencia publicada en 2022 mediante validación irt. El estudio trabajó con 13,670 estudiantes de primaria, 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 evidenció validez rigurosa; identificó déficits conceptuales, crecimiento por grado y ventaja pequeña de los varones. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.

Ece et al.⁶ aportaron evidencia publicada en 2024 mediante desarrollo y validación rasch. El estudio trabajó con Estudiantes de secundaria, 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 produjo 20 ítems válidos y fiables con valores estadísticos adecuados. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.

Fitria et al. (7) aportaron evidencia publicada en 2024 mediante desarrollo/evaluación rasch. El estudio trabajó con Estudiantes en aprendizaje de ciencias, 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 propone un formato de prueba eficiente y psicométricamente analizado para alfabetización digital. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 trabajo⁸ aportó evidencia publicada en 2026 mediante revisión/diseño educativo. El estudio trabajó con Corpus documental, 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 aporta criterios de segmentación, carga cognitiva y autonomía relevantes para diseño de microcontenidos, aunque no es exclusivamente microlearning. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.

Wibowo et al.⁹ aportaron evidencia publicada en 2025 mediante cuantitativo rasch. El estudio trabajó con 251 estudiantes, 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 halló diferencias por género, educación parental y uso de Internet; el instrumento permitió mapear competencias digitales. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.

Muhammad et al.¹⁰ aportaron evidencia publicada en 2025 mediante estudio piloto rasch. El estudio trabajó con 261 futuros docentes universitarios, 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 todos los ítems cumplieron criterios de ajuste; aportó mapeo de dificultad por área de competencia digital. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.¹¹ aportaron 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 TRI y validación digital 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.

David12aportaron 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 TRI y validación digital 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 trabajo13aportó evidencia publicada en 2026 mediante revisión/modelización educativa. El estudio trabajó con Corpus documental, 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 ofrece sustento contextual para secuenciar microcontenidos según progreso y desempeño; pertinencia indirecta al microlearning móvil. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 sintetizada14aportó evidencia publicada en 2025 mediante revisión sistemática prisma. El estudio trabajó con 24 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 reporta mejoras en comprensión, motivación, habilidades prácticas y retención; barreras: costo, infraestructura, formación docente y resistencia. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 examinada15aportó evidencia publicada en 2025 mediante revisión sistemática prisma. El estudio trabajó con 291 registros; 50 textos completos, 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 literatura muestra expansión rápida y beneficios potenciales, pero exige terminología consistente, mejor diseño y estudios comparables. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 estudio16aportó evidencia publicada en 2025 mediante multicéntrico de aceptación/experiencia. El estudio trabajó con Estudiantes de 5 universidades europeas, 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 alta motivación y emociones positivas, especialmente con interactividad; se reportaron cibermareo y problemas técnicos en algunos contextos. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 investigación17aportó evidencia publicada en 2024 mediante revisión sistemática cualitativa. El estudio trabajó con 50 artículos (2020–2024), 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 barreras económicas, técnicas, pedagógicas y sociales y plantea accesibilidad como condición para adopción sostenible. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 trabajo¹⁸ aportó evidencia publicada en 2023 mediante dos experimentos. El estudio trabajó con Estudiantes universitarios, 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 beneficios fueron mayores en participantes con alta tendencia de absorción; características individuales moderan efectos de la inmersión. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 sintetizada¹⁹ aportó evidencia publicada en 2022 mediante diseño y evaluación educativa. El estudio trabajó con Estudiantes; muestra reportada en artículo, 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 experiencia VR apoyó aprendizaje activo y pensamiento computacional, con relevancia del diseño de interacción y feedback. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 examinada²⁰ aportó evidencia publicada en 2022 mediante estudio de intervención educativa. El estudio trabajó con Estudiantes universitarios, 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 beneficios en motivación e implicación, junto con barreras de infraestructura y diseño instruccional. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 estudio²¹ aportó evidencia publicada en 2022 mediante comparativo/cuasi-experimental. El estudio trabajó con Estudiantes de odontología, 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 vR mejoró experiencia de aprendizaje y algunas medidas de desempeño/conocimiento en educación clínica. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 investigación²² aportó evidencia publicada en 2022 mediante scoping review. El estudio trabajó con Estudios de VR con estudiantes de medicina, 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 VR se usa ampliamente en anatomía/procedimientos y suele mejorar engagement; faltan estudios longitudinales y comparadores estandarizados. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.

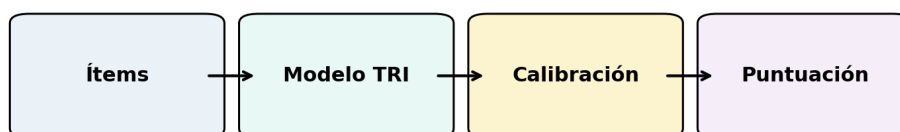
El trabajo²³ aportó evidencia publicada en 2022 mediante encuesta/estudio institucional. El estudio trabajó con Académicos/estudiantes de universidad, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la aceptación es favorable, pero capacitación, soporte técnico, costo y alineación curricular condicionan la integración. Este antecedente es relevante para la perspectiva de TRI y validación digital 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.

Katerina et al.²⁴ aportaron evidencia publicada en 2022 mediante desarrollo y validación psicométrica. El estudio trabajó con 156 estudiantes de grado y posgrado, 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 escala mostró ajuste, cargas, validez interna y fiabilidad

adecuadas; edad, experiencia informática y área de estudio se asociaron con dimensiones, no el género. Este antecedente es relevante para la perspectiva de TRI y validación digital 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 TRI y validación digital en relación con uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 2. ¿Qué patrones de evidencia permiten comprender la dimensión 2 de TRI y validación digital en relación con uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 3. ¿Qué patrones de evidencia permiten comprender la dimensión 3 de TRI y validación digital en relación con uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 4. ¿Qué patrones de evidencia permiten comprender la dimensión 4 de TRI y validación digital en relación con uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales y qué condiciones metodológicas o contextuales modifican su interpretación?

MÉTODOS

Método de revisión sistemática

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 uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales, 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,

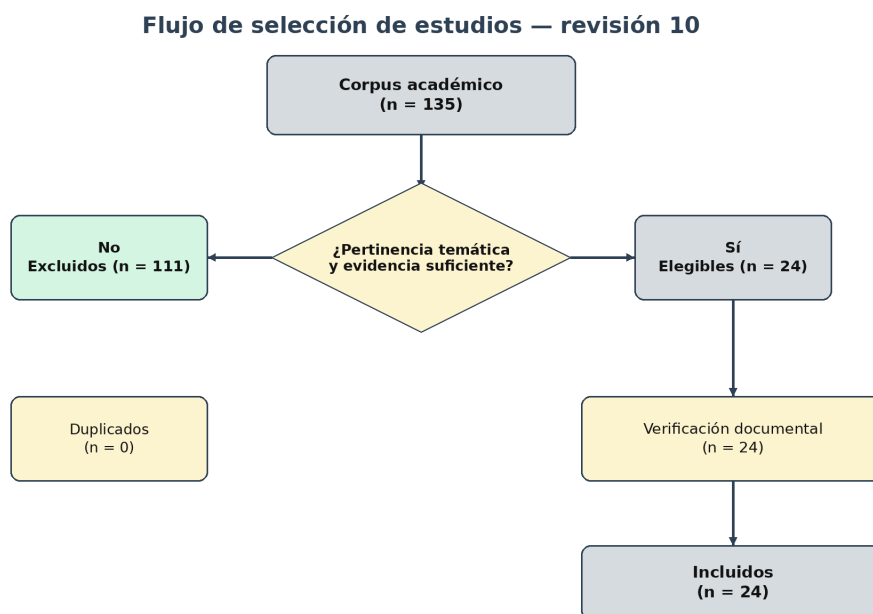
por tratar poblaciones o desenlaces periféricos, o por no aportar evidencia metodológica suficiente para la síntesis focal. El corpus final quedó constituido por 24 estudios. La selección se efectuó sobre la información documentada en las matrices y se evitó atribuir a una base bibliográfica particular 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. 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.

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



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.

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

Modelos TRI y supuestos

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 uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales 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.

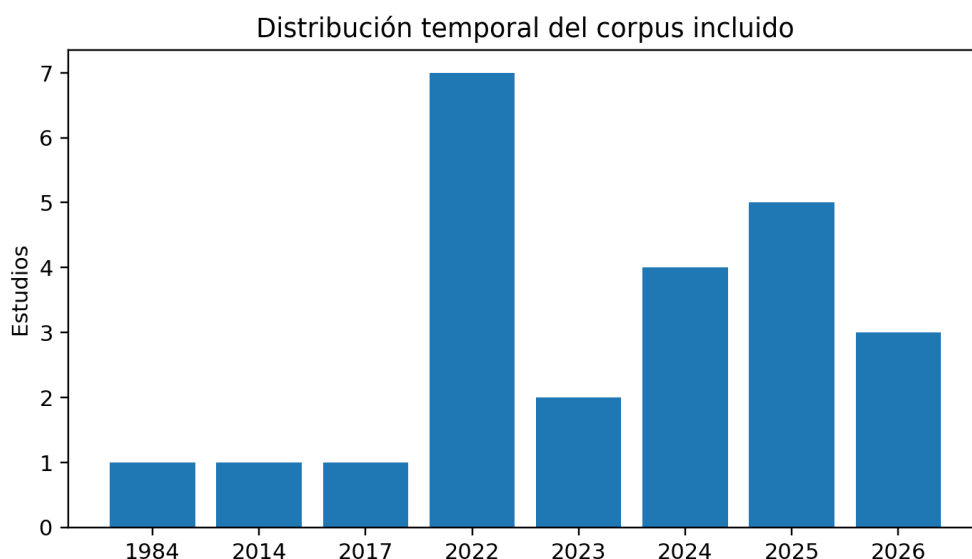
La síntesis de resultados revela un patrón central vinculado con TRI y validación digital: los efectos o propiedades favorables no aparecen como atributos automáticos de la tecnología o del instrumento, sino como resultados condicionados por diseño, implementación, calidad de datos y alineación pedagógica. 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.

Tabla 1. Caracterización metodológica del corpus incluido

Año	Diseño/metodología	Muestra o corpus	Evidencia
2026	IRT 2PL en examen computarizado	500 estudiantes universitarios de primer año; 10 ítems de álgebra	Alta
2017	Psicométrico IRT	735 estudiantes universitarios	Alta
2014	Validación con teoría de respuesta al ítem	560 estudiantes; 56 ítems	Alta
2023	Psicométrico IRT	450 estudiantes universitarios	Alta
2022	Validación IRT	13,670 estudiantes de primaria	Alta
2024	Desarrollo y validación Rasch	Estudiantes de secundaria	Alta
2024	Desarrollo/evaluación Rasch	Estudiantes en aprendizaje de ciencias	Alta
2026	Revisión/diseño educativo	Corpus documental	Media
2025	Cuantitativo Rasch	251 estudiantes	Alta
2025	Estudio piloto Rasch	261 futuros docentes universitarios	Alta
2024	Desarrollo y validación de escala	Estudiantes de una universidad venezolana	Alta
1984	Artículo metodológico fundacional	Aplicaciones de CAT en problemas educativos	Alta

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

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.

Evidencias de calibración y funcionamiento de ítems

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: Evalúa equidad y fiabilidad de selección aleatoria de ítems en pruebas computarizadas. Confirmó estructura unifactorial y evaluó utilidad de la escala de aprendizaje activo mediante IRT. La medida final de 27 ítems fue especialmente precisa en niveles bajo y medio de habilidad. 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: Evaluó propiedades psicométricas de Grit-S con IRT y estructura de consistencia de interés/perseverancia. Evidenció validez rigurosa; identificó déficits conceptuales, crecimiento por grado y ventaja pequeña de los varones. Produjo 20 ítems válidos y fiables con valores estadísticos adecuados. 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: Propone un formato de

prueba eficiente y psicométricamente analizado para alfabetización digital. Aporta criterios de segmentación, carga cognitiva y autonomía relevantes para diseño de microcontenidos, aunque no es exclusivamente microlearning. Halló diferencias por género, educación parental y uso de Internet; el instrumento permitió mapear competencias digitales. 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: Todos los ítems cumplieron criterios de ajuste; aportó mapeo de dificultad por área de competencia digital. Se obtuvo un modelo de 22 variables en cuatro dimensiones con propiedades psicométricas satisfactorias. 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: Ofrece sustento contextual para secuenciar microcontenidos según progreso y desempeño; pertinencia indirecta al microlearning móvil. Reporta mejoras en comprensión, motivación, habilidades prácticas y retención; barreras: costo, infraestructura, formación docente y resistencia. La literatura muestra expansión rápida y beneficios potenciales, pero exige terminología consistente, mejor diseño y estudios comparables. 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: Alta motivación y emociones positivas, especialmente con interactividad; se reportaron cibermareo y problemas técnicos en algunos contextos. Identifica barreras económicas, técnicas, pedagógicas y sociales y plantea accesibilidad como condición para adopción sostenible. Los beneficios fueron mayores en participantes con alta tendencia de absorción; características individuales moderan efectos de la inmersión. 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 experiencia VR apoyó aprendizaje activo y pensamiento computacional, con relevancia del diseño de interacción y feedback. Se observaron beneficios en motivación e implicación, junto con barreras de infraestructura y diseño instruccional. VR mejoró experiencia de aprendizaje y algunas medidas de desempeño/conocimiento en educación clínica. 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 VR se usa ampliamente en anatomía/procedimientos y suele mejorar engagement; faltan estudios longitudinales y comparadores estandarizados. La aceptación es favorable, pero capacitación, soporte técnico, costo y alineación curricular condicionan la integración. La escala mostró ajuste, cargas, validez interna y fiabilidad adecuadas; edad, experiencia informática y área de estudio se asociaron con dimensiones, no el género. 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
Item Response Theory for trait assessment in randomized item pool for computer based test	Evalúa equidad y fiabilidad de selección aleatoria de ítems en pruebas computarizadas.	Incluido por pertinencia directa con la
The Analysis of Active Learning Scale by Item Response Theory (IRT)	Confirmó estructura unifactorial y evaluó utilidad de la escala de aprendizaje activo mediante IRT.	Incluido por pertinencia directa con la
Direct measures of digital information processing and communication skills in primary educ	La medida final de 27 ítems fue especialmente precisa en niveles bajo y medio de habilidad.	Incluido por pertinencia directa con la
Psychometric properties of Grit-S scale in university students: An application of the grad	Evaluó propiedades psicométricas de Grit-S con IRT y estructura de consistencia de interés/perseverancia.	Incluido por pertinencia directa con la
Validating a computational thinking concepts test for primary education using item respons	Evidenció validez rigurosa; identificó déficits conceptuales, crecimiento por grado y ventaja pequeña de los varones.	Incluido por pertinencia directa con la
Digital literacy scale: Validity and reliability study with the Rasch model	Produjo 20 ítems válidos y fiables con valores estadísticos adecuados.	Incluido por pertinencia directa con la
Innovations in assessing students' digital literacy skills in learning science: effective	Propone un formato de prueba eficiente y psicométricamente analizado para alfabetización digital.	Incluido por pertinencia directa con la
Diseño y evaluación de entornos virtuales centrados en el estudiante desde la neuroeducaci	Aporta criterios de segmentación, carga cognitiva y autonomía relevantes para diseño de microcontenidos, aunque no es exclusivamente microlearning.	Incluido por pertinencia directa con la
Assessing Digital Competence in Indonesian students: demographic and Internet usage factor	Halló diferencias por género, educación parental y uso de Internet; el instrumento permitió mapear competencias digitales.	Incluido por pertinencia directa con la
Pilot Study of Digital Competency Mapping of Indonesian Preservice Teachers: Rasch Model A	Todos los ítems cumplieron criterios de ajuste; aportó mapeo de dificultad por área de competencia digital.	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.

Aplicaciones adaptativas y bancos de ítems

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 uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales 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
Fiabilidad	4
Validez	4
competencia	2
rendimiento	1
Equidad	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 uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales debe interpretarse desde una lógica de TRI y validación digital 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 exceder los alcances de la evidencia disponible. 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.

NoIofrece un punto de contraste útil para esta interpretación porque su estudio documenta evalúa equidad y fiabilidad de selección aleatoria de ítems en pruebas computarizadas. 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.

Kai et al.2ofrece un punto de contraste útil para esta interpretación porque su estudio documenta confirmó estructura unifactorial y evaluó utilidad de la escala de aprendizaje activo mediante IRT. 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.

Koen et al.3ofrece un punto de contraste útil para esta interpretación porque su estudio documenta la medida final de 27 ítems fue especialmente precisa en niveles bajo y medio de habilidad. 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.

Amal4ofrece un punto de contraste útil para esta interpretación porque su estudio documenta evaluó propiedades psicométricas de Grit-S con IRT y estructura de consistencia de interés/perseverancia. 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.

Siu-Cheung et al.5ofrece un punto de contraste útil para esta interpretación porque su estudio documenta evidenció validez rigurosa; identificó déficits conceptuales, crecimiento por grado y ventaja pequeña de los varones. 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.

Ece et al.6ofrece un punto de contraste útil para esta interpretación porque su estudio documenta produjo 20 ítems válidos y fiables con valores estadísticos adecuados. 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 validación

Las implicaciones prácticas derivadas de la revisión deben traducirse en decisiones graduales. Para uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales, 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 uso de teoría de respuesta al ítem en la validación de instrumentos educativos digitales 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 TRI y validación digital 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. Theory for trait assessment in randomized item pool for computer based test. *Scientific African*. 2026. <https://doi.org/10.1016/j.sciaf.2026.e03226>
2. Kai Hatano; Yugo Saito. The Analysis of Active Learning Scale by Item Response Theory (IRT). *Japan Journal of Educational Technology*. 2017. <https://doi.org/10.15077/jjet.40056>
3. Koen Aesaert; Daniël van Nijlen; Ruben Vanderlinde; Johan van Braak. Direct measures of digital information processing and communication skills in primary education: Using item response theory for the development and validation of an ICT competence scale. *Computers & Education*. 2014. <https://doi.org/10.1016/j.compedu.2014.03.013>
4. Amal Alhadabi. Psychometric properties of Grit-S scale in university students: An application of the graded response model. *Cogent Education*. 2023. <https://doi.org/10.1080/2331186X.2022.2162699>
5. Siu-Cheung Kong; Ming Lai. Validating a computational thinking concepts test for primary education using item response theory: An analysis of students' responses. *Computers & Education*. 2022. <https://doi.org/10.1016/j.compedu.2022.104562>
6. Ece Avinç; Fatih Doğan. Digital literacy scale: Validity and reliability study with the Rasch model. *Education and Information Technologies*. 2024. <https://doi.org/10.1007/s10639-024-12662-7>
7. Fitria Lafifa; Dadan Rosana. Innovations in assessing students' digital literacy skills in learning science: Effective multiple choice closed-ended tests using Rasch model. *Turkish Online Journal of Distance Education*. 2024. <https://doi.org/10.17718/tojde.1364528>
8. Delia Consuegra; Daniel Alejandro Rodríguez Estrella; Braulio Alejandro Olaya Mieles. Diseño y evaluación de entornos virtuales centrados en el estudiante: un enfoque desde la neuroeducación digital. *Imperium Académico Multidisciplinary Journal*. 2025;2(4):1-15. <https://doi.org/10.63969/pypf5d63>
9. Wibowo Heru Prasetyo; Beti Indah Sari; Rizky Novia Saputri; Noor Banu Mahadir Naidu; Triyanto; Jagad Aditya Dewantara. Assessing Digital Competence in Indonesian students: demographic and Internet usage factors through the Rasch Model. *Journal of e-Learning and Knowledge Society*. 2025. <https://doi.org/10.20368/1971-8829/1135991>
10. Muhammad Luthfi Hidayat; Shemsu Gulta Abdurahman; Dwi Setyo Astuti; et al. Pilot Study of Digital Competency Mapping of Indonesian Preservice Teachers: Rasch Model Analysis. *Indonesian Journal on Learning and Advanced Education*. 2025. <https://doi.org/10.23917/ijolae.v7i1.23935>
11. 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>
12. David J. Weiss; G. Gage Kingsbury. Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*. 1984. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>
13. Erick Daniel Rivera Quiñonez; Blanca Romina Barcia Rivera; Amy Linett Triviño Díaz; María Gabriela Zambrano Álvarez. Modelos matemáticos para la personalización del aprendizaje en entornos virtuales. *Multidisciplinary Journal of Sciences, Discoveries, and Society*. 2026;3(3):1-12. <https://doi.org/10.63969/18qqe109>
14. Evolución de la realidad virtual en la educación superior latinoamericana: una revisión sistemática. *Formación Universitaria*. 2025. <https://doi.org/10.4067/S0718-50062025000600035>
15. Immersive virtual reality in higher education: a systematic review of the scientific literature. *Virtual Reality*. 2025. <https://doi.org/10.1007/s10055-025-01136-x>
16. Immersive virtual reality learning environments for higher education: A student acceptance study. *Computers & Education: X Reality*. 2025. <https://doi.org/10.1016/j.cexr.2025.100105>
17. Michael Antonio Pinargote Castro; Cecibel Verónica Solórzano Ortega; María José Lozano Torres; Flavio Jesús Jiménez Mejía. Accesibilidad de la Realidad Virtual Aumentada en la Educación Universitaria: Estrategias, Desafíos y Beneficios. *Revista Científica*. 2024;9(33):252-275. <https://doi.org/10.29394/Scientific.issn.2542-2987.2024.9.33.12.252-275>
18. Learning by teaching in immersive virtual reality – Absorption tendency increases learning outcomes. *Learning and Instruction*. 2023. <https://doi.org/10.1016/j.learninstruc.2022.101716>
19. Felix J. Agbo; Solomon S. Oyelere; Jarkko Suhonen; Markku Tukiainen. Design, development, and evaluation of a virtual reality game-based application to support computational thinking. *Educational Technology Research and Development*. 2023. <https://doi.org/10.1007/s11423-022-10161-5>
20. Eduardo Campos; Irving Hidrogo; Genaro Zavala. Impact of virtual reality use on the teaching and learning of vectors. *Frontiers in Education*. 2022. <https://doi.org/10.3389/feduc.2022.965640>
21. Yue Huang; Xin Cheng; Unman Chan; Liyang Zheng; Yingwen Hu; Yueting Sun; Pengyu Lai; Jun Dai; Xuesong Yang. Virtual reality approach for orthodontic education at School of Stomatology, Jinan University. *Journal of Dental Education*. 2022;86(8):1025-1035. <https://doi.org/10.1002/jdd.12915>
22. Hanyi Jiang; Shalini Vimalasvaran; Jason K. Wang; Keng B. Lim; Shalini R. Mogali; Lorainne Tudor Car. Virtual Reality in Medical Students' Education: Scoping Review. *JMIR Medical Education*. 2022. <https://doi.org/10.2196/34860>
23. Current State and General Perceptions of the Use of Extended Reality (XR) Technology at the University of Newcastle: Interviews and Surveys From Staff and Students. *SAGE Open*. 2022. <https://doi.org/10.1177/21582440221093348>
24. Katerina Tzafilkou; Maria Perifanou; Anastasios A. Economides. Development and validation of students' digital competence scale (SDiCoS). *International Journal of Educational Technology in Higher Education*. 2022. <https://doi.org/10.1186/s41239-022-00330-0>

Financiación

Ninguna.

Conflictos de interés

No existen.

Contribución de autoría

Conceptualización: Karla Andrea Paredes Rosado

Curación de datos: Karla Andrea Paredes Rosado

Análisis formal: Karla Andrea Paredes Rosado

Investigación: Karla Andrea Paredes Rosado

Metodología: Karla Andrea Paredes Rosado

Supervisión: Karla Andrea Paredes Rosado

Validación: Karla Andrea Paredes Rosado

Visualización: Karla Andrea Paredes Rosado

Redacción – borrador original: Karla Andrea Paredes Rosado

Redacción – revisión y edición: Karla Andrea Paredes Rosado