Medición y Evaluación del Aprendizaje Inclusivo. 2025; 2(2):16
doi: 10.63688/aprendizaje.v2.i2.16

REVISIÓN SISTEMÁTICA

Equity and bias in automated educational evaluation systems based on artificial intelligence

Equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial

Pamela Navarro Escobar[1]

[1] Centro de Investigación, Desarrollo e Innovación JYMNIJA, Oruro, Bolivia

Recibido: 30-05-2025 | Revisado: 07-06-2025 | Aceptado: 19-08-2025 | Publicado: 29-08-2025

Autor de correspondencia: Pamela Navarro Escobar

ABSTRACT

Introduction: Equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial 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; 7 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: Equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial 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 7 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 equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial 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.

En este escenario, la literatura sobre inteligencia artificial aplicada a la educación advierte que el sesgo algorítmico puede emerger en distintas etapas del ciclo de datos y afectar de forma desigual a grupos de estudiantes, por lo que la equidad debe examinarse como una propiedad del sistema sociotécnico y no únicamente del algoritmo[8]. En evaluación automatizada, la validez de las puntuaciones requiere además justificar la relevancia del constructo, la representación de la tarea, la transparencia del modelo y las consecuencias de uso de los resultados[9-11].

El núcleo conceptual del manuscrito se organiza alrededor de justicia algorítmica. 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.

Esta perspectiva también converge con marcos recientes de IA responsable en educación superior, que articulan equidad, rendición de cuentas, transparencia y ética como dimensiones interdependientes[15]. Las revisiones sobre riesgos éticos en inteligencia artificial educativa señalan que la no discriminación, la supervisión humana, la gestión responsable de datos y la explicabilidad son condiciones necesarias para una adopción institucional justificable[16,26].

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.

La investigación previa sobre calificación automatizada confirma que la precisión técnica, por sí sola, no resuelve problemas de validez, generalización ni comparabilidad entre tareas y poblaciones. Las revisiones de sistemas de puntuación de ensayos muestran una evolución desde modelos basados en rasgos hacia arquitecturas neuronales y modelos de lenguaje, manteniéndose desafíos asociados con coherencia, representación del constructo y transferencia entre contextos[12,13,30]. De manera complementaria, la explicabilidad se ha propuesto como requisito para que docentes, estudiantes y responsables institucionales puedan comprender y cuestionar decisiones asistidas por IA[14].

Equidad, sesgo y validez de decisiones automatizadas

No[1]aportaron evidencia publicada en 2026 mediante estudio metodológico de equidad. El estudio trabajó con Múltiples modelos de scoring y subgrupos demográficos, 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 equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales. Este antecedente es relevante para la perspectiva de justicia algorítmica 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.

No[2]aportaron evidencia publicada en 2025 mediante evaluación empírica de equidad algorítmica. El estudio trabajó con 38,722 respuestas de texto de PISA 2015, 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 no halló diferencias de género discernibles, pero

sí un sesgo pequeño por lengua de hogar; identificó factores de ítem vinculados a inequidad. Este antecedente es relevante para la perspectiva de justicia algorítmica 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.

No3aportaron evidencia publicada en 2026 mediante comparación empírica de modelos. El estudio trabajó con Ensayos con subgrupos de género, ELL y educación especial, 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 gPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados. Este antecedente es relevante para la perspectiva de justicia algorítmica 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.

Ikkyu et al. (4) aportaron evidencia publicada en 2025 mediante estudio metodológico con simulación y datos reales. El estudio trabajó con conjunto de datos real, 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 identificó correctamente características responsables del sesgo en simulaciones y halló conjuntos consistentes de variables en datos reales. Este antecedente es relevante para la perspectiva de justicia algorítmica 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.

Matthew et al.5aportaron evidencia publicada en 2022 mediante metodológico con simulación y aplicación. El estudio trabajó con Simulaciones y un ítem de evaluación de lectura a gran escala, 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 mostró cómo omisión/proxies/confusión y criterios de optimización pueden producir sesgo y cómo evaluarlo psicométricamente. Este antecedente es relevante para la perspectiva de justicia algorítmica 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.

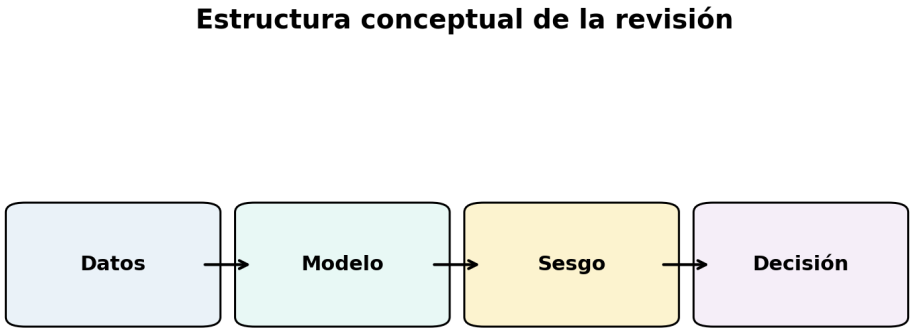
Hamida et al.6aportaron evidencia publicada en 2026 mediante evaluación técnica de accesibilidad. El estudio trabajó con Estudiantes con necesidades educativas especiales en educación superior, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria 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 justicia algorítmica 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.

A. et al.7aportaron evidencia publicada en 2022 mediante estudio metodológico/aplicación ilustrativa. El estudio trabajó con Aplicación ilustrativa a entrevistas de video automatizadas, 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 marco más flexible para incorporar perspectivas de stakeholders y equidad en evaluaciones con IA. Este antecedente es relevante para la perspectiva de justicia algorítmica 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 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.

Figura 1. Estructura conceptual de la revisión



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

Preguntas de revisión

Pregunta 1. ¿Qué patrones de evidencia permiten comprender la dimensión 1 de justicia algorítmica en relación con equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 2. ¿Qué patrones de evidencia permiten comprender la dimensión 2 de justicia algorítmica en relación con equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 3. ¿Qué patrones de evidencia permiten comprender la dimensión 3 de justicia algorítmica en relación con equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial y qué condiciones metodológicas o contextuales modifican su interpretación?

MÉTODOS

Protocolo PRISMA

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 equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial, correspondencia con educación superior o con instrumentos aplicables a ese nivel, descripción suficiente del método y disponibilidad de resultados interpretables. Se excluyeron 128 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 7 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.

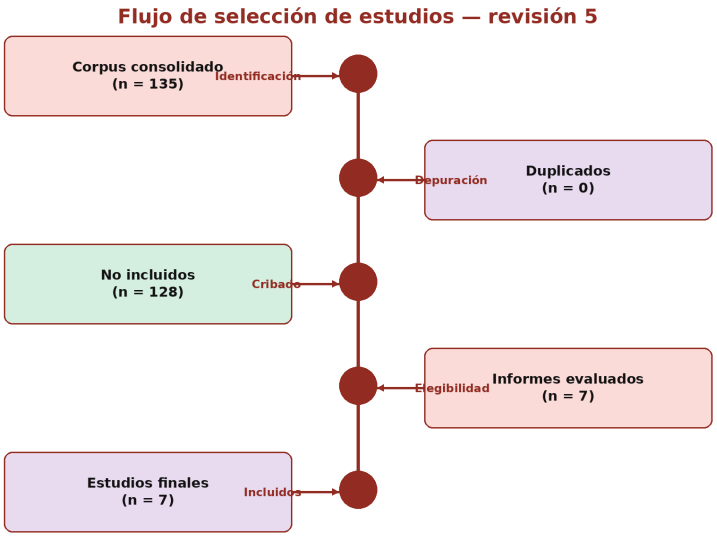
La extracción de información se estructuró mediante una matriz con autores, año, título, fuente, DOI o enlace trazable, procedencia, tipo de estudio, muestra o corpus, metodología resumida, resultados principales, nivel de evidencia y pertinencia. Esta organización permitió comparar trabajos con diseños 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 preserva 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



Nota. El flujo representa 135 registros académicos normalizados, sin duplicados en el corpus consolidado; 128 registros no cumplieron el umbral de pertinencia temática y 7 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 128 registros por insuficiente correspondencia con la pregunta específica de la revisión, mientras que 7 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

Fuentes y formas de sesgo

El corpus final reunió 7 estudios publicados durante 2022–2026. 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 equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial 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 justicia algorítmica: 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.

Tabla 1. Caracterización metodológica del corpus incluido

Año	Diseño/metodología	Muestra o corpus	Evidencia
2026	Estudio metodológico de equidad	Múltiples modelos de scoring y subgrupos demográficos	Alta
2025	Evaluación empírica de equidad algorítmica	38,722 respuestas de texto de PISA 2015	Alta
2026	Comparación empírica de modelos	Ensayos con subgrupos de género, ELL y educación especial	Alta
2025	Estudio metodológico con simulación y datos reales	conjunto de datos real	Alta
2022	Metodológico con simulación y aplicación	Simulaciones y un ítem de evaluación de lectura a gran escala	Alta
2026	Evaluación técnica de accesibilidad	Estudiantes con necesidades educativas especiales en educación superior	Alta
2022	Estudio metodológico/aplicación ilustrativa	Aplicación ilustrativa a entrevistas de video automatizadas	Alta

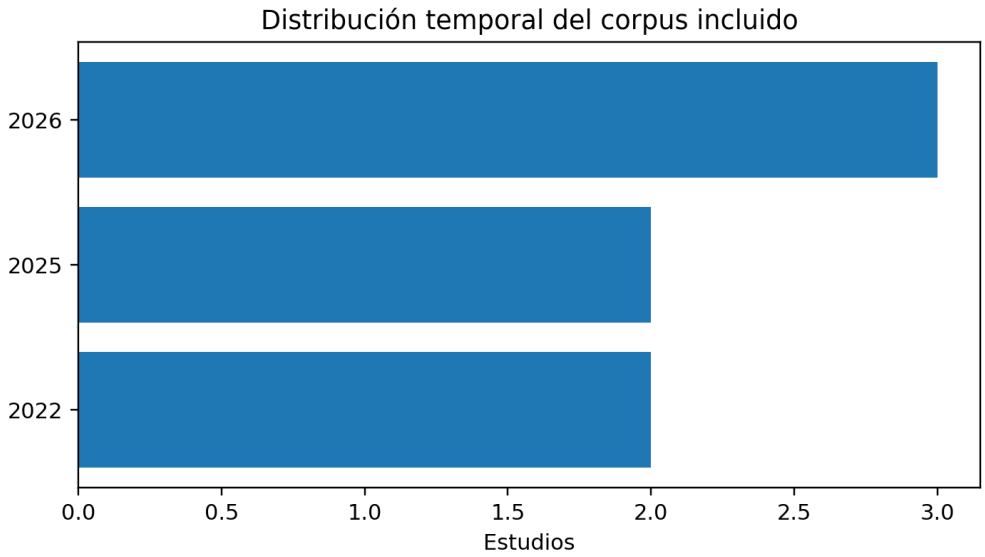
Nota. Datos extraídos de la matriz de evidencia suministrada para esta revisión.

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.

Mitigación y gobernanza

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.

Figura 3. Distribución temporal de los estudios incluidos



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

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: La equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales. No halló diferencias de género discernibles, pero sí un sesgo pequeño por lengua de hogar; identificó factores de ítem vinculados a inequidad. GPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados. 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 medida identificó correctamente características responsables del sesgo en simulaciones y halló conjuntos consistentes de variables en datos reales. Mostró cómo omisión/proxies/confusión y criterios de optimización pueden producir sesgo y cómo evaluarlo psicométricamente. Identifica que accesibilidad de evaluación digital exige compatibilidad con tecnologías de apoyo, autonomía e integridad bajo restricciones de examen. 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 marco más flexible para incorporar perspectivas de stakeholders y equidad en evaluaciones con IA. 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.

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.

Tabla 2. Síntesis de hallazgos relevantes

Estudio	Hallazgo recuperado	Pertinencia
Assessing fairness in AI-assisted writing scoring: Developing fairness measures to detect	La equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales.	Incluido por pertinencia directa con la
Algorithmic Fairness in Automatic Short Answer Scoring	No halló diferencias de género discernibles, pero sí un sesgo pequeño por lengua de hogar; identificó factores de ítem vinculados a inequidad.	Incluido por pertinencia directa con la
Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, featu	GPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados.	Incluido por pertinencia directa con la
Identifying Features Contributing to Differential Prediction Bias of Automated Scoring Sys	La medida identificó correctamente características responsables del sesgo en simulaciones y halló conjuntos consistentes de variables en datos reales.	Incluido por pertinencia directa con la
Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring	Mostró cómo omisión/proxies/confusión y criterios de optimización pueden producir sesgo y cómo evaluarlo psicométricamente.	Incluido por pertinencia directa con la
Accessible ICT-Enabled Examination Technologies for Students with Special Educational Need	Identifica que accesibilidad de evaluación digital exige compatibilidad con tecnologías de apoyo, autonomía e integridad bajo restricciones de examen.	Incluido por pertinencia directa con la
Toward Argument-Based Fairness with an Application to AI-Enhanced Educational Assessments	Propone un marco más flexible para incorporar perspectivas de stakeholders y equidad en evaluaciones con IA.	Incluido por pertinencia directa con la

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

DISCUSIÓN

La síntesis permite sostener que el problema de equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial debe interpretarse desde una lógica de justicia algorítmica 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.

Los hallazgos del corpus son consistentes con investigaciones externas que muestran que la elección de métricas de equidad puede modificar la interpretación de un mismo sistema predictivo. En aprendizaje analítico, no existe una métrica universalmente adecuada: su selección depende del contexto, de los grupos afectados y del tipo de decisión que se pretende apoyar[17,20]. Por ello, la evaluación de justicia algorítmica debe documentar explícitamente qué noción de equidad se adopta y qué consecuencias tendría priorizar una métrica sobre otra.

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 transparencia y la confianza adquieren especial relevancia cuando los resultados algorítmicos orientan intervenciones, calificaciones o decisiones de apoyo académico.

La literatura sobre aprendizaje analítico responsable destaca que equidad, confianza, transparencia y responsabilidad deben incorporarse desde el diseño y no añadirse únicamente después de detectar una disparidad. 18 Asimismo, estudios sobre poblaciones históricamente subatendidas muestran que una intervención algorítmica puede perseguir objetivos de equidad y, al mismo tiempo, introducir nuevas tensiones éticas si la clasificación de riesgo no se interpreta dentro de su contexto social.19

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 evidencia reciente también refuerza la necesidad de evaluar la robustez de los sistemas entre subgrupos. Estudios sobre puntuación automatizada han mostrado diferencias de precisión y generalización según el conjunto de datos y el grupo analizado, mientras que investigaciones sobre percepción estudiantil indican que la aceptación de calificadores basados en IA depende también de la percepción de justicia del procedimiento y del resultado21,22. En consecuencia, la equidad técnica y la equidad percibida deben considerarse dimensiones relacionadas, pero no equivalentes.

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.

La investigación sobre mitigación de sesgos muestra, además, que reducir una disparidad no garantiza mejorar simultáneamente todas las métricas de justicia y desempeño. Comparaciones recientes entre reponderación, aprendizaje adversarial y otros procedimientos evidencian compensaciones entre utilidad predictiva y equidad entre subgrupos23. Esto respalda la conveniencia de informar resultados desagregados y realizar análisis de sensibilidad antes de utilizar una predicción automatizada en decisiones educativas de consecuencias relevantes.

No1ofrece un punto de contraste útil para esta interpretación porque su estudio documenta la equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales. 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.

No2ofrece un punto de contraste útil para esta interpretación porque su estudio documenta no halló diferencias de género discernibles, pero sí un sesgo pequeño por lengua de hogar; identificó factores de ítem vinculados a inequidad. 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.

No3ofrece un punto de contraste útil para esta interpretación porque su estudio documenta gPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados. 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.

Ikkyu et al.4ofrece un punto de contraste útil para esta interpretación porque su estudio documenta la medida identificó correctamente características responsables del sesgo en simulaciones y halló conjuntos consistentes de variables en datos reales. 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.

Matthew et al.5ofrece un punto de contraste útil para esta interpretación porque su estudio documenta mostró cómo omisión/proxies/confusión y criterios de optimización pueden producir sesgo y cómo evaluarlo psicométricamente. 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.

Hamida et al.6ofrece un punto de contraste útil para esta interpretación porque su estudio documenta identifica que accesibilidad de evaluación digital exige compatibilidad con tecnologías de apoyo, autonomía e integridad bajo restricciones de examen. 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.

Salvaguardas para implementación

Las implicaciones prácticas derivadas de la revisión deben traducirse en decisiones graduales. Para equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial, 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.

La gobernanza también debe contemplar los sesgos presentes en contenidos y datos educativos, pues estos pueden propagarse a los modelos que los utilizan como insumo. El aprendizaje analítico ha comenzado a incorporar métodos específicos para identificar sesgos étnicos en materiales educativos, ampliando la evaluación de equidad más allá de la salida final del algoritmo24. En educación superior, la justicia de la evaluación algorítmica también se relaciona con la agencia del estudiante, la alfabetización de datos y la posibilidad de comprender cómo se generan las recomendaciones o clasificaciones25.

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.

Las revisiones contemporáneas sobre IA en educación superior coinciden en que la evaluación responsable debe integrar dimensiones pedagógicas y éticas durante todo el ciclo de instrucción, desde la recolección de datos hasta la interpretación y uso de resultados27. En sistemas basados en grandes modelos de lenguaje, los marcos de evaluación comienzan igualmente a incorporar criterios de efectividad educativa, veracidad, alineación social y equidad, evitando reducir la calidad del sistema a una única medida de rendimiento31.

Vacíos de evidencia

La principal limitación del corpus es su heterogeneidad. Las diferencias entre diseños, instrumentos, poblaciones y desenlaces restringen la posibilidad de producir un estimador común y obligan a interpretar los patrones desde una lógica comparativa. Esta limitación no invalida la síntesis, pero sí condiciona el alcance de las conclusiones y refuerza la necesidad de replicaciones con protocolos comparables.

También se observa una concentración de estudios en determinados contextos y tecnologías, lo que reduce la certeza sobre la transferibilidad a instituciones con infraestructura, perfiles

estudiantiles o culturas de evaluación diferentes. Las investigaciones futuras deberían documentar con mayor precisión condiciones de implementación, características de los participantes y decisiones adoptadas durante el despliegue de las herramientas.

Un vacío especialmente relevante corresponde a la representatividad de los datos de entrenamiento. Evidencia reciente en puntuación automatizada de escritura indica que la composición demográfica del conjunto de entrenamiento puede modificar los patrones de sesgo y que los conjuntos balanceados pueden reducir, aunque no necesariamente eliminar, disparidades entre grupos28. De forma paralela, comparaciones entre calificadores humanos y modelos GPT muestran que la consistencia automatizada puede coexistir con diferencias dependientes del criterio evaluado, lo que refuerza la necesidad de validación por constructo y supervisión humana29.

Finalmente, la revisión se apoya en el corpus consolidado documentado en las matrices suministradas. Las conclusiones representan fielmente ese conjunto y no se extienden a registros que no forman parte de la evidencia disponible. Esta delimitación preserva la trazabilidad y permite que futuras actualizaciones incorporen nuevos estudios sin alterar retrospectivamente la base analítica utilizada en el presente manuscrito.

CONCLUSIONES

La revisión sistemática muestra que equidad y sesgos en sistemas automatizados de evaluación educativa basados en inteligencia artificial 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 justicia algorítmica 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.

Síntesis ampliada de la evidencia

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por desarrolla dos medidas de equidad; compara ml, llm y diferencias por grupos/proficiencia. y considera Múltiples modelos de scoring y subgrupos demográficos. Entre los resultados recuperados se informa que la equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño.

Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa.

El trabajo se caracteriza por combinó representaciones semánticas y clasificadores; comparó disparidades por género y lengua del hogar. y considera 38,722 respuestas de texto de PISA 2015. Entre los resultados recuperados se informa que no halló diferencias de género discernibles, pero sí un sesgo pequeño por lengua de hogar; identificó factores de ítem vinculados a inequidad. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por comparó gpt-4o con modelos basados en características y calificadores humanos; evaluó precisión y equidad por subgrupos. y considera Ensayos con subgrupos de género, ELL y educación especial. Entre los resultados recuperados se informa que gPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por propone medida de contribución al sesgo aplicable a algoritmos no lineales; pruebas con redes neuronales y svr. y considera conjunto de datos real. Entre los resultados recuperados se informa que la medida identificó correctamente características responsables del sesgo en simulaciones y halló conjuntos consistentes de variables en datos reales.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por revisión de mecanismos de sesgo, dos métodos simples de evaluación, simulación y aplicación empírica. y considera Simulaciones y un ítem de evaluación de lectura a gran escala. Entre los resultados recuperados se informa que mostró cómo omisión/proxies/confusión y criterios de optimización pueden producir sesgo y cómo evaluarlo psicométricamente.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por evaluación de usabilidad, fiabilidad, calidad de ajustes y equidad de plataformas de examen digital. y considera Estudiantes con necesidades educativas especiales en educación superior. Entre los resultados recuperados se informa que identifica que accesibilidad de evaluación digital exige compatibilidad con tecnologías de apoyo, autonomía e integridad bajo restricciones de examen. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por desarrolla un argumento de equidad complementario a validez y lo aplica a una evaluación mejorada por ia. y considera Aplicación ilustrativa a entrevistas de video automatizadas. Entre los resultados recuperados se informa que propone un marco más flexible para incorporar perspectivas de stakeholders y equidad en

evaluaciones con IA.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por desarrolla dos medidas de equidad; compara ml, llm y diferencias por grupos/proficiencia. y considera Múltiples modelos de scoring y subgrupos demográficos. Entre los resultados recuperados se informa que la equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por combinó representaciones semánticas y clasificadores; comparó disparidades por género y lengua del hogar. y considera 38,722 respuestas de texto de PISA 2015. Entre los resultados recuperados se informa que no halló diferencias de género discernibles, pero sí un sesgo pequeño por lengua de hogar; identificó factores de ítem vinculados a inequidad.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por comparó gpt-4o con modelos basados en características y calificadores humanos; evaluó precisión y equidad por subgrupos. y considera Ensayos con subgrupos de género, ELL y educación especial. Entre los resultados recuperados se informa que gPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por propone medida de contribución al sesgo aplicable a algoritmos no lineales; pruebas con redes neuronales y svr. y considera conjunto de datos real. Entre los resultados recuperados se informa que la medida identificó correctamente características responsables del sesgo en simulaciones y halló conjuntos consistentes de variables en datos reales.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por revisión de mecanismos de sesgo, dos métodos simples de evaluación, simulación y aplicación empírica. y considera Simulaciones y un ítem de evaluación de lectura a gran escala. Entre los resultados recuperados se informa que mostró cómo omisión/proxies/confusión y criterios de optimización pueden producir sesgo y cómo evaluarlo psicométricamente.. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación

del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por evaluación de usabilidad, fiabilidad, calidad de ajustes y equidad de plataformas de examen digital. y considera Estudiantes con necesidades educativas especiales en educación superior. Entre los resultados recuperados se informa que identifica que accesibilidad de evaluación digital exige compatibilidad con tecnologías de apoyo, autonomía e integridad bajo restricciones de examen. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

La evidencia correspondiente a uno de los estudios del corpus permite profundizar la interpretación comparativa. El trabajo se caracteriza por desarrolla un argumento de equidad complementario a validez y lo aplica a una evaluación mejorada por ia. y considera Aplicación ilustrativa a entrevistas de video automatizadas. Entre los resultados recuperados se informa que propone un marco más flexible para incorporar perspectivas de stakeholders y equidad en evaluaciones con IA. En la síntesis global, este hallazgo se utiliza para contrastar patrones y no para extender una conclusión más allá de su diseño. Su lectura junto con otros estudios permite precisar la relación entre calidad de medición, condiciones de implementación y utilidad educativa, al tiempo que muestra la necesidad de distinguir entre efectos observados, percepciones de usuarios y evidencia sobre validez o desempeño. Esta diferenciación fortalece la interpretación del conjunto y evita que la innovación sea evaluada únicamente por novedad tecnológica.

Referencias

1. Chen G, Wu AD, Zhang C. Assessing fairness in AI-assisted writing scoring: Developing fairness measures to detect predictive bias in automated essay scoring. Assessing Writing. 2026;69:101066. https://doi.org/10.1016/j.asw.2026.101066
2. Andersen N, Mang J, Goldhammer F, et al. Algorithmic Fairness in Automatic Short Answer Scoring. International Journal of Artificial Intelligence in Education. 2025;35:3128–3165. https://doi.org/10.1007/s40593-025-00495-5
3. Huang Y, Palermo C, Wilson J. Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. Assessing Writing. 2026;69:101047. https://doi.org/10.1016/j.asw.2026.101047
4. Choi I, Johnson MS. Identifying Features Contributing to Differential Prediction Bias of Automated Scoring Systems. Journal of Educational Measurement. 2025;62(4):838–861. https://doi.org/10.1111/jedm.70015
5. Johnson MS, Liu X, McCaffrey DF. Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring. Journal of Educational Measurement. 2022;59(3):338–361. https://doi.org/10.1111/jedm.12335
6. Al-Maamari H, Sidhu MS, Hussain SM, Al-Ghaili AM, Sidhu KK. 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.
7. Huggins-Manley AC, Booth BM, D'Mello SK. Toward Argument-Based Fairness with an Application to AI-Enhanced Educational Assessments. Journal of Educational Measurement. 2022;59(3):362–388. https://doi.org/10.1111/jedm.12334
8. Baker RS, Hawn A. Algorithmic Bias in Education. International Journal of Artificial Intelligence in Education. 2022;32(4):1052-1092. https://doi.org/10.1007/s40593-021-00285-9
9. Ferrara S, Qunbar S. Validity Arguments for AI-Based Automated Scores: Essay Scoring as an Illustration. Journal of Educational Measurement. 2022;59(3):288-313. https://doi.org/10.1111/jedm.12333
10. Ercikan K, McCaffrey DF. Optimizing Implementation of Artificial-Intelligence-Based Automated Scoring: An Evidence Centered Design Approach for Designing Assessments for AI-based Scoring. Journal of Educational Measurement. 2022;59(3):272-287. https://doi.org/10.1111/jedm.12332
11. Shermis MD. Anchoring Validity Evidence for Automated Essay Scoring. Journal of Educational Measurement. 2022;59(3):314-337. https://doi.org/10.1111/jedm.12336
12. Ramesh D, Sanampudi SK. An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review. 2022;55(3):2495-2527. https://doi.org/10.1007/s10462-021-10068-2
13. Susanti MNI, Ramadhan A, Warnars HLHS. Automatic essay exam scoring system: a systematic literature review. Procedia Computer Science. 2023;216:531-538. https://doi.org/10.1016/j.procs.2022.12.166
14. Khosravi H, Shum SB, Chen G, Conati C, Tsai YS, Kay J. Explainable Artificial Intelligence in education. Computers and Education: Artificial Intelligence. 2022;3:100074. https://doi.org/10.1016/j.caeai.2022.100074
15. Memarian B, Doleck T. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. Computers and Education: Artificial Intelligence. 2023;5:100152. https://doi.org/10.1016/j.caeai.2023.100152
16. Zhu H, Sun Y, Yang J. Towards responsible artificial intelligence in education: a systematic review on identifying and mitigating ethical risks. Humanities and Social Sciences Communications. 2025;12:1111. https://doi.org/10.1057/s41599-025-05252-6
17. Deho OB, Zhan C, Li J, Liu J, Liu L, Le TD. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? British Journal of Educational Technology. 2022;53(4):822-843. https://doi.org/10.1111/bjet.13217
18. Khalil M, Prinsloo P, Slade S. Fairness, Trust, Transparency, Equity, and Responsibility in Learning Analytics. Journal of Learning Analytics. 2023;10(1):1-7. https://doi.org/10.18608/jla.2023.7983
19. Froehlich L, Weydner-Volkmann S. Adaptive Interventions Reducing Social Identity Threat to Increase Equity in Higher Distance Education: A Use Case and Ethical Considerations on Algorithmic Fairness. Journal of Learning Analytics. 2024. https://doi.org/10.18608/jla.2024.8301
20. Cohausz L, Kappenberger J, Stuckenschmidt H. What Fairness Metrics Can Really Tell You: A Case Study in the Educational Domain. Proceedings of the 14th Learning Analytics and Knowledge Conference. 2024:792-799. https://doi.org/10.1145/3636555.3636873
21. Yang K, Raković M, Li Y, Guan Q, Gašević D, Chen G. Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability. Proceedings of the AAAI Conference on Artificial Intelligence. 2024;38(20). https://doi.org/10.1609/aaai.v38i20.30254
22. Jones-Jang SM, Chung M, Choi J, Kim N, Lee S. Fairness perceptions of AI in grading systems: Examining how discontent with the status quo and outcome favorability reduce AI reluctance. Computers and Education: Artificial Intelligence. 2025;8:100419. https://doi.org/10.1016/j.caeai.2025.100419
23. Villegas-Ch W, Gutierrez R, Garcia-Ortiz J, Govea J. Mitigating algorithmic bias in educational prediction systems using reweighting and adversarial learning. Frontiers in Education. 2026;11:1841724. https://doi.org/10.3389/feduc.2026.1841724
24. Albuquerque J, Rienties B, Hlosta M, Holmes W. Learning Analytics to Uncover Ethnic Bias in Educational Texts: An Ensemble Learning Approach. Journal of Learning Analytics. 2026;13(1):143-162. https://doi.org/10.18608/jla.2026.8905

25. Lluch Molins L, Lindín Soriano C. The self-regulatory paradox of learning analytics: student expectations and the conditions for fair algorithmic assessment in higher education. Frontiers in Education. 2026;11:1913278. https://doi.org/10.3389/feduc.2026.1913278
26. Agarwal B, Urlings C, van Lankveld G, Klemke R. Identifying the ethical values and norms for artificial intelligence in education: A systematic literature review. International Journal of Artificial Intelligence in Education. 2026;36(1-2):100004. https://doi.org/10.1016/j.ijaied.2026.100004
27. Zhuang M, Long S, Martin F, Castellanos-Reyes D. The affordances of Artificial Intelligence (AI) and ethical considerations across the instruction cycle: A systematic review of AI in online higher education. The Internet and Higher Education. 2025;67:101039. https://doi.org/10.1016/j.iheduc.2025.101039
28. Holmes L, Morris W, Crossley S, Choi JS. Assessing fairness in finetuned scoring models with demographically restricted training data. Assessing Writing. 2026;68:101032. https://doi.org/10.1016/j.asw.2026.101032
29. Wu HN, Chu MN, Hsu JL. Comparing GPT and human raters in essay assessment: Variability, bias, and the potential of LLM-based scoring. Computers and Education Open. 2026;10:100341. https://doi.org/10.1016/j.caeo.2026.100341
30. Sun J, Song T, Weiming P, Song J. A survey of automated essay scoring: Challenges, advances, and future. Neurocomputing. 2025;650:130916. https://doi.org/10.1016/j.neucom.2025.130916
31. Lin P, Deng Q, Zhou Y. Towards responsible AI in education: A Delphi-AHP-based framework for evaluating educational large language models. Computers and Education: Artificial Intelligence. 2026;10:100534. https://doi.org/10.1016/j.caeai.2025.100534

Financiación

Ninguna.

Conflictos de interés

No existen.

Contribución de autoría

Conceptualización: Pamela Navarro Escobar
Curación de datos: Pamela Navarro Escobar
Análisis formal: Pamela Navarro Escobar
Investigación: Pamela Navarro Escobar
Metodología: Pamela Navarro Escobar
Supervisión: Pamela Navarro Escobar
Validación: Pamela Navarro Escobar
Visualización: Pamela Navarro Escobar
Redacción – borrador original: Pamela Navarro Escobar
Redacción – revisión y edición: Pamela Navarro Escobar