

RMPC-Vol.1. N1. 006

Detección de sesgos en evaluaciones educativas mediante inteligencia artificial y modelos psicométricos: hacia una evaluación algorítmica justa

Detecting bias in educational assessments through artificial intelligence and psychometric models: toward fair algorithmic assessment

Autores:

Katiuska Adelaida Bastidas González
Instituto de Investigación, Desarrollo e Innovación JOGB
Milagro - Ecuador
<https://orcid.org/0000-0002-9336-7332>
proyectosescolares2016@gmail.com

Segundo Alexander Rosado Martínez
Instituto de Investigación, Desarrollo e Innovación JOGB
Milagro - Ecuador
<https://orcid.org/0009-0004-7655-1279>
alexrosado1996@hotmail.com

Giovanni Juan Maldonado Aguirre
Centro de Investigación, Desarrollo e Innovación JOGB
La Paz - Bolivia
<https://orcid.org/0009-0007-8627-8087>
gmaldonadoda@gmail.com

Autor de correspondencia: Katiuska Bastidas González, proyectosescolares2016@gmail.com

Recepción: 04-03-2024

Aceptación: 26-05-2024

Publicación: 13-06-2024

Cómo citar este artículo:

Bastidas González, K. A., Rosado Martínez, S. A., & Maldonado Aguirre, G. J. (2024). Inteligencia artificial generativa y pensamiento crítico en la educación superior: Implicaciones éticas, académicas y sociales. Revista Multidisciplinaria Perspectiva Científica, 1(1), 1-29.
<https://rmpc.editorialgutierrez.com/index.php/rmpc/article/view/9>

RESUMEN

La inteligencia artificial en la evaluación educativa permite analizar patrones complejos, pero exige demostrar que los instrumentos y algoritmos no generan diferencias sistemáticas irrelevantes entre grupos. Se integraron procedimientos psicométricos y aprendizaje automático para detectar señales de sesgo en una prueba objetiva de 24 ítems y examinar la equidad de los patrones de respuesta. Participaron 300 estudiantes universitarios, de 18 a 33 años, de cinco programas académicos. La prueba alcanzó consistencia interna aceptable para investigación exploratoria ($KR-20/\alpha=0.737$); la dificultad de los ítems osciló entre 0.49 y 0.71, y la puntuación media fue 15.14 de 24. El rendimiento se relacionó moderadamente con el promedio académico ($r=0.601$; $p<0.001$). Los análisis de funcionamiento diferencial mediante regresión logística, con corrección FDR, no detectaron DIF robusto por sexo/género ni por discapacidad/neurodivergencia; en primera generación universitaria, I1 mostró DIF no uniforme corregido ($q=0.040$). Los modelos de regresión logística y random forest empleados para predecir pertenencia grupal obtuvieron AUC cercanas a 0,50 en las comparaciones, indicando escasa información demográfica recuperable del patrón de respuestas. Los hallazgos respaldan una auditoría dual: psicometría para localizar funcionamiento diferencial e inteligencia artificial interpretable para explorar patrones multivariados. La ausencia de señales no demuestra equidad definitiva en subgrupos pequeños, pero ambos enfoques pueden complementarse sin confundir precisión con justicia algorítmica en contextos reales de evaluación educativa universitaria.

Palabras clave: inteligencia artificial; sesgo algorítmico; funcionamiento diferencial del ítem; psicometría; aprendizaje automático; evaluación educativa.

ABSTRACT

The incorporation of artificial intelligence into educational assessment expands the capacity to analyze complex patterns, but it also requires evidence that instruments and algorithms do not generate systematic construct-irrelevant differences across groups. This study integrated psychometric procedures and machine learning to detect potential bias in a 24-item objective test and examine the fairness of response patterns. Participants were 300 university students aged 18 to 33 years from five academic programs. The test showed acceptable internal consistency for exploratory research ($KR-20/\alpha=0.737$); item difficulty ranged from 0.49 to 0.71, and the mean score was 15.14 out of 24. Performance was moderately associated with academic average ($r=0.601$; $p<0.001$). Logistic-regression differential item functioning analyses with FDR correction identified no robust DIF by sex/gender or disability/neurodivergence status; for first-generation university status, item I1 displayed corrected non-uniform DIF ($q=0.040$). Logistic regression and random forest models predicting group membership yielded AUC values close to 0.50 in the main comparisons, indicating little recoverable demographic information in response patterns. The findings support a two-level audit strategy: psychometrics to identify differential functioning and interpretable artificial intelligence to explore multivariate patterns. The overall absence of signals does not establish definitive fairness, particularly in small subgroups, but shows that both approaches can complement each other without conflating predictive accuracy with algorithmic justice in the assessment context examined.

Keywords: artificial intelligence; algorithmic bias; differential item functioning; psychometrics; machine learning; educational assessment.

1. INTRODUCCIÓN

El crecimiento de la inteligencia artificial en educación ha trasladado parte de la discusión sobre calidad de la evaluación desde la construcción tradicional de pruebas hacia sistemas capaces de puntuar, clasificar, recomendar o tomar decisiones automáticamente. Este desplazamiento no elimina las preguntas clásicas de la psicometría; las vuelve más complejas. Un algoritmo puede ser técnicamente preciso y, al mismo tiempo, utilizar señales que produzcan errores desiguales entre grupos. Del mismo modo, un test puede mostrar fiabilidad global adecuada y contener ítems que funcionen de forma distinta para estudiantes con el mismo nivel de habilidad. La literatura reciente sostiene que la equidad debe examinarse a lo largo de todo el ciclo de datos, medición, modelado y decisión (1-4). En evaluación educativa, el sesgo algorítmico adquiere especial relevancia porque las decisiones pueden afectar calificaciones, acceso a oportunidades, retroalimentación, colocación académica y trayectorias de aprendizaje. Baker y Hawn documentaron que los sesgos pueden aparecer en diferentes etapas de los sistemas educativos basados en algoritmos y afectar de manera desigual a grupos definidos por género, raza, nacionalidad, condición socioeconómica o discapacidad (1). Esta perspectiva obliga a distinguir entre una diferencia real de desempeño y una diferencia producida por la medición o por el algoritmo. La equidad no exige que todos los grupos obtengan idénticos resultados; exige que las diferencias observadas sean defendibles respecto del constructo y del uso de la puntuación. La psicometría ofrece una tradición consolidada para abordar esta cuestión a nivel de ítem. El funcionamiento diferencial del ítem (DIF) se presenta cuando personas de grupos distintos, pero comparables en la habilidad relevante, tienen probabilidades diferentes de responder correctamente. La estadística educativa contemporánea mantiene el DIF como una herramienta central de validez y equidad y, al mismo tiempo, incorpora métodos de alta dimensionalidad y aprendizaje automático (2). El valor del DIF es conceptual además

de técnico: evita interpretar cualquier diferencia bruta entre grupos como sesgo, porque primero condiciona la comparación al nivel del rasgo que la evaluación intenta medir.

El aprendizaje automático introduce una segunda perspectiva. Tay et al. propusieron el concepto de sesgo de medición basado en machine learning para describir situaciones en las que un modelo produce niveles de predicción o precisiones diferentes entre subgrupos aun cuando el nivel del constructo relevante es equivalente (3). Johnson, Liu y McCaffrey mostraron que decisiones de modelado aparentemente técnicas, como omitir variables, utilizar proxies o elegir determinados criterios de optimización, pueden generar puntuaciones automatizadas injustas (4). Por tanto, la auditoría algorítmica debe evaluar no solo el rendimiento global del modelo, sino también qué información utiliza y cómo se distribuyen sus errores.

Una línea reciente conecta explícitamente ambas tradiciones. Belzak planteó una visión multidimensional del sesgo de medición capaz de considerar interacciones y fuentes complejas de DIF mediante aprendizaje automático (5). Loukina, Madnani y Zechner mostraron que la fairness en aplicaciones educativas automatizadas admite varias definiciones y requiere examinar el desempeño por subgrupos, además del rendimiento agregado (6). Esta convergencia es particularmente útil cuando las evaluaciones producen patrones de respuesta con relaciones no lineales o interacciones difíciles de representar mediante un único contraste clásico.

La auditoría de modelos educativos amplía este puente al comparar su desempeño entre subgrupos. Gardner, Brooks y Baker propusieron el análisis por segmentos para cuantificar diferencias en las curvas ROC de modelos predictivos estudiantiles y mostraron que la fairness puede variar según el algoritmo, las variables y el contexto del curso (7). Esta perspectiva obliga a identificar el resultado predicho, el atributo grupal y la métrica de comparación; así, la pregunta deja de ser únicamente si el algoritmo acierta y pasa a incluir si su desempeño es comparable entre grupos.

La literatura sobre fairness también advierte que una sola métrica no basta para establecer justicia. Mehrabi et al. organizaron las fuentes de sesgo y las definiciones de fairness

utilizadas en aprendizaje automático, destacando que cada criterio responde a supuestos distintos (8). El marco de gestión de riesgos de IA del NIST añadió una perspectiva de ciclo de vida basada en gobernar, mapear, medir y gestionar riesgos de forma continua (9). En evaluación educativa, esto implica examinar representatividad, desempeño por grupos, explicabilidad, supervisión humana y condiciones de uso.

El presente estudio integró ambos enfoques en el análisis de una evaluación objetiva de 24 ítems que cubrió comprensión verbal, razonamiento cuantitativo, razonamiento lógico e interpretación de datos. El objetivo fue analizar sus propiedades psicométricas, detectar DIF en variables grupales relevantes y explorar, mediante regresión logística y random forest, si los patrones de respuesta permitían recuperar pertenencia grupal por encima del azar. La lógica del estudio es deliberadamente conservadora: una señal estadística se interpreta como motivo de revisión y no como demostración automática de discriminación. Se formularon cuatro expectativas. Primero, que la prueba mostrara consistencia interna suficiente para un análisis exploratorio. Segundo, que la mayoría de los ítems no presentara DIF robusto después de controlar habilidad y multiplicidad. Tercero, que un modelo de aprendizaje automático justo no pudiera predecir con alta capacidad atributos grupales únicamente a partir de patrones de respuesta, una vez considerada la habilidad global. Cuarto, que la puntuación de la prueba se relacionara positivamente con un criterio académico externo. Estas hipótesis permiten evaluar conjuntamente calidad de medición, señales de inequidad y utilidad predictiva sin convertir ninguna de ellas en sustituto de las demás.

2. REVISIÓN DE LA LITERATURA

Sesgo, equidad y validez en evaluación algorítmica

Las discusiones actuales sobre equidad educativa distinguen al menos dos niveles: equidad de medición y equidad de decisión. La primera pregunta si las puntuaciones representan el mismo constructo de forma comparable; la segunda examina si las decisiones tomadas a partir de esas puntuaciones producen tratamientos o errores diferencialmente injustos.

Huggins-Manley, Booth y D'Mello propusieron un enfoque argumentativo de fairness complementario a la validez para evaluaciones tradicionales y potenciadas por IA (10). La importancia de esta separación radica en que un instrumento puede medir adecuadamente y aun así alimentar una regla de decisión problemática, o puede ocurrir lo contrario: un algoritmo aparentemente equilibrado puede operar sobre una medida que ya contiene sesgo.

En sistemas educativos, el problema se agrava porque los atributos protegidos pueden estar correlacionados con oportunidades de aprendizaje, acceso tecnológico, idioma o condiciones institucionales. Por ello, la eliminación mecánica de una variable sensible no garantiza justicia. Los métodos psicométricos para evaluar puntuación automatizada muestran que proxies y variables confundidas pueden reintroducir diferencias incluso cuando el atributo protegido no aparece explícitamente en el modelo (4,11). El análisis debe estudiar relaciones condicionadas, errores y consecuencias, no solo la presencia o ausencia de una columna demográfica.

DIF como mecanismo psicométrico de auditoría

El DIF permite localizar ítems cuyo comportamiento cambia entre grupos después de controlar el nivel de habilidad. En pruebas dicotómicas, la regresión logística ofrece una implementación flexible: el efecto de grupo representa DIF uniforme y la interacción entre grupo y habilidad representa DIF no uniforme. El primero implica un desplazamiento relativamente estable; el segundo indica que la diferencia depende del nivel de desempeño. En ambos casos, la detección estadística constituye el comienzo de una revisión de contenido, no una sentencia automática de sesgo (2,5).

Los procedimientos contemporáneos deben considerar además multiplicidad y tamaño muestral. Una prueba de 24 ítems aplicada a varias variables grupales genera numerosos contrastes; sin corrección, algunos pueden aparecer significativos por azar. Por esa razón, el presente estudio utiliza control FDR. La prudencia es todavía mayor en grupos pequeños, donde la ausencia de significación puede reflejar potencia limitada. Esta lógica es

compatible con auditorías repetidas y evidencia acumulativa, en lugar de una certificación binaria de fairness (8,9).

Aprendizaje automático interpretable para detectar patrones de sesgo

Los métodos de machine learning pueden complementar el DIF al explorar relaciones no lineales, interacciones y configuraciones multivariadas. En aplicaciones educativas, Loukina et al. mostraron la utilidad de contrastar varias definiciones de fairness y de examinar el desempeño automatizado por subgrupos (6). La idea subyacente es que, si determinados ítems contienen información sistemática sobre pertenencia grupal que no debería estar presente una vez considerada la habilidad, esa información puede señalar una fuente potencial de inequidad.

Belzak destacó que las fuentes de sesgo pueden ser multidimensionales e interseccionales, por lo que analizar cada atributo de forma aislada puede ocultar patrones complejos (5). Sin embargo, mayor flexibilidad algorítmica también aumenta el riesgo de sobreajuste y de explicaciones espurias. Por eso la validación cruzada y la interpretabilidad son indispensables. Un modelo que predice grupo con AUC cercana a 0,50 ofrece poca evidencia de señal demográfica recuperable; una AUC alta no demostraría por sí sola sesgo, pero justificaría investigar qué ítems o relaciones producen esa capacidad.

Precisión y fairness no son equivalentes

Uno de los errores más frecuentes en evaluación algorítmica es asumir que un modelo con mayor exactitud es necesariamente más justo. Tay et al. distinguen explícitamente desempeño predictivo y sesgo de medición (3). Johnson et al. muestran que modelos optimizados para reducir error promedio pueden mantener diferencias sistemáticas entre grupos (4). El análisis por segmentos de Gardner et al. aporta una forma de comprobar si el rendimiento predictivo cambia entre subgrupos a lo largo de los posibles umbrales de clasificación (7).

La evidencia empírica en aplicaciones educativas refuerza esta distinción. Loukina et al. observaron que distintas definiciones de fairness pueden conducir a conclusiones

diferentes al analizar puntuación automatizada y antecedentes lingüísticos (6). En consecuencia, una evaluación responsable debe reportar al menos dos familias de resultados: calidad predictiva global y comportamiento por grupos. La ausencia de disparidad en una métrica tampoco excluye otras formas de inequidad, de modo que la auditoría debe ser plural y contextual.

Hacia una evaluación algorítmica justa

Las propuestas recientes de IA responsable convergen en principios de representatividad, transparencia, privacidad, explicabilidad, supervisión humana y mejora continua (9). Estos principios son especialmente pertinentes en educación porque las puntuaciones tienen significado solo dentro de una teoría del aprendizaje y un uso previsto. La automatización no convierte la evaluación en un problema puramente computacional; mantiene las obligaciones de validez, fiabilidad y equidad propias de la medición.

El enfoque integrado utilizado en este trabajo entiende la psicometría y la IA como capas complementarias. La primera evalúa dificultad, discriminación, consistencia y DIF; la segunda explora si el patrón multivariado contiene información grupal recuperable y si esa información se mantiene fuera de muestra. Esta arquitectura permite evitar dos extremos: confiar exclusivamente en pruebas clásicas incapaces de capturar interacciones complejas, o utilizar modelos opacos sin un marco de medición que defina qué diferencias son legítimas.

Evidencia reciente sobre fairness, DIF e inteligencia artificial

La literatura empírica sobre fairness educativa confirma que la equidad no puede evaluarse únicamente con una métrica global. Jiang y Pardos mostraron, en predicción de calificaciones universitarias, que estrategias de balanceo y aprendizaje adversarial modificaban las diferencias de desempeño entre grupos raciales, evidenciando que optimizar el rendimiento agregado no garantiza resultados equitativos (12). Una revisión sistemática de 63 estudios sobre sesgo predictivo en educación encontró diversidad de

atributos protegidos, métricas y estrategias de mitigación, junto con un predominio de evaluaciones offline y limitada validación en contextos reales (13).

La comparación desagregada del rendimiento también aparece en modelos predictivos estudiantiles. Gardner et al. analizaron cinco algoritmos en 44 cursos masivos y mostraron que la disparidad entre curvas ROC variaba con el algoritmo, el conjunto de variables y características del curso, sin observar un intercambio estricto entre precisión y fairness (7). Estos resultados respaldan la validación por subgrupos empleada en este artículo y muestran que una auditoría puede identificar diferencias específicas aun cuando el desempeño global parezca adecuado.

Desde la psicometría, el DIF sigue siendo indispensable para interpretar diferencias de respuesta. Los análisis deben separar la comparación bruta entre grupos del funcionamiento de cada reactivo después de controlar el nivel del rasgo. Belzak destacó que el sesgo de medición puede ser multidimensional y que el aprendizaje automático permite explorar fuentes complejas de DIF (5). Estos planteamientos apoyan el enfoque adoptado en este artículo: una diferencia observada debe separarse de una diferencia condicionada al nivel de desempeño antes de atribuirla a sesgo.

La detección de DIF depende de propiedades del instrumento y de la estructura de los datos. D'Urso et al. demostraron que la longitud de la escala afecta las conclusiones de invariancia en modelos categóricos e IRT, por lo que recomiendan considerar explícitamente el número de ítems y el parámetro examinado al seleccionar un procedimiento (14). Teresi et al., al revisar análisis DIF en medidas PROMIS, subrayaron que significación estadística, magnitud del efecto, impacto sobre puntuaciones y elección del método deben interpretarse conjuntamente (15). En evaluación de aula, Rezai desarrolló y validó un cuestionario de fairness que integra diseño, oportunidades para demostrar aprendizaje, administración, calificación, retroalimentación e interpretación de resultados (16).

La fairness tampoco se limita a atributos demográficos aislados. La revisión de Gohar y Cheng mostró que analizar cada atributo protegido por separado puede ocultar

desigualdades que emergen en sus intersecciones y que las métricas deben elegirse de acuerdo con el contexto de uso (17). Esta advertencia es pertinente para muestras con subgrupos pequeños: ampliar las comparaciones sin suficiente tamaño muestral puede reducir la estabilidad, mientras omitirlas puede ocultar patrones relevantes.

Finalmente, la evaluación responsable requiere documentación y auditoría a lo largo del ciclo de vida. Las tarjetas de modelo propuestas por Mitchell et al. permiten declarar usos previstos, límites y desempeño por grupos (18), mientras Raji et al. plantearon una auditoría interna de extremo a extremo para integrar responsabilidades y revisión sistemática en el desarrollo de sistemas de IA (19). Estas propuestas son coherentes con el presente trabajo: la inteligencia artificial no reemplaza la psicometría, sino que aumenta la necesidad de comprobar qué constructo se mide, cómo funcionan los ítems y si los errores se distribuyen de manera comparable entre grupos.

3. METODOLOGÍA

Diseño

Se realizó un estudio cuantitativo, transversal, psicométrico y comparativo con un componente de aprendizaje automático interpretable. La unidad de análisis fue el patrón de respuestas a una prueba objetiva de 24 ítems. El estudio combinó teoría clásica de los tests, análisis de funcionamiento diferencial mediante regresión logística y modelos supervisados de clasificación para explorar la capacidad de recuperar pertenencia grupal a partir de las respuestas.

Participantes y datos

La base estuvo integrada por 300 estudiantes universitarios con consentimiento afirmativo. La edad varió entre 18 y 33 años ($M=22.65$; $DE=3.01$); participaron estudiantes de cinco programas y de primero a décimo semestre. En sexo/género se registraron 163 mujeres, 125 hombres, 4 participantes en otra categoría y 8 que prefirieron no responder. Treinta y nueve participantes declararon condición de discapacidad o neurodivergencia, 248

respondieron que no y 13 prefirieron no responder. Para primera generación universitaria, 142 respondieron sí, 148 no y 10 prefirieron no responder. El acceso tecnológico fue estable en 213 casos, intermitente en 69 y no estable en 18.

Tabla 1. Características de la muestra

Variable	Categoría	n	%
Sexo/género	Mujer	163	54,3
	Hombre	125	41,7
	Otro / no responde	12	4,0
Discapacidad/neuro divergencia	Sí	39	13,0
	No	248	82,7
	No responde	13	4,3
Primera generación	Sí	142	47,3
	No	148	49,3
	No responde	10	3,3
Acceso tecnológico	Estable	213	71,0
	Intermitente	69	23,0
	No estable	18	6,0

Instrumento

La prueba incluyó 24 preguntas de elección múltiple con cuatro alternativas y una única respuesta correcta. Se organizaron cuatro dominios de seis ítems: comprensión verbal, razonamiento cuantitativo, razonamiento lógico e interpretación de datos. Las respuestas se codificaron 1=correcta y 0=incorrecta. La puntuación total osciló entre 0 y 24. El promedio académico, registrado en escala de 10 puntos, se utilizó como criterio externo. Las variables grupales se emplearon exclusivamente para auditoría de equidad y no para modificar las puntuaciones individuales.

Procedimiento analítico

Se calculó dificultad como proporción de aciertos, discriminación mediante correlación punto-biserial ítem-resto y consistencia interna mediante alfa/KR-20. Se compararon puntuaciones entre grupos con t de Welch o ANOVA, según correspondiera, y se estimó la correlación entre puntuación total y promedio académico. Para DIF se ajustó una regresión

logística por ítem, utilizando como variable de emparejamiento la puntuación del resto de los reactivos estandarizada. El efecto de grupo representó DIF uniforme y la interacción grupo $\times$ habilidad, DIF no uniforme. Los valores p se corrigieron mediante Benjamini-Hochberg.

El componente de inteligencia artificial utilizó regresión logística y random forest con los 24 patrones binarios y la puntuación global como predictores de pertenencia a los grupos principales. El rendimiento se estimó mediante validación cruzada estratificada de cinco particiones y se reportó AUC y balanced accuracy. Este análisis sigue la lógica de aprendizaje automático interpretable aplicada a la detección de fairness: una capacidad predictiva demográfica claramente superior al azar sería una señal para investigar qué respuestas contienen información grupal adicional (6,7). El análisis se efectuó sobre los registros observados y conservó sus valores originales.

Hipótesis

H1: la prueba mostrará consistencia interna suficiente para investigación exploratoria ($\alpha/\text{KR-20} \geq 0,70$). H2: después de controlar habilidad y multiplicidad, la mayoría de los ítems no mostrará DIF robusto en las comparaciones grupales principales. H3: los modelos de clasificación no alcanzarán capacidad elevada para predecir pertenencia grupal a partir de las respuestas, lo que sería consistente con baja señal demográfica recuperable. H4: la puntuación total presentará una relación positiva y significativa con el promedio académico.

Consideraciones éticas

Todos los registros incluyeron consentimiento afirmativo y el conjunto analizado no contenía nombres, correos ni números de identificación. Las variables demográficas permitían la opción de no responder y esas categorías se excluyeron de los contrastes binarios correspondientes. Los análisis de fairness se realizaron a nivel agregado y no se utilizaron para clasificar o penalizar a participantes. El reporte se limita a las condiciones y variables documentadas en los registros analizados.

4. RESULTADOS

Propiedades psicométricas

Tabla 2. Dificultad y discriminación de los ítems

Ítem	Dominio	p de acierto	r ítem-resto
I1	Comprensión verbal	0.600	0.260
I2	Comprensión verbal	0.687	0.376
I3	Comprensión verbal	0.663	0.286
I4	Comprensión verbal	0.597	0.275
I5	Comprensión verbal	0.667	0.313
I6	Comprensión verbal	0.597	0.121
I7	Razonamiento cuantitativo	0.697	0.368
I8	Razonamiento cuantitativo	0.643	0.237
I9	Razonamiento cuantitativo	0.673	0.195
I10	Razonamiento cuantitativo	0.707	0.161
I11	Razonamiento cuantitativo	0.700	0.319
I12	Razonamiento cuantitativo	0.657	0.280
I13	Razonamiento lógico	0.660	0.316
I14	Razonamiento lógico	0.673	0.250
I15	Razonamiento lógico	0.577	0.241
I16	Razonamiento lógico	0.643	0.360
I17	Razonamiento lógico	0.570	0.267
I18	Razonamiento lógico	0.490	0.257
I19	Interpretación de datos	0.673	0.377
I20	Interpretación de datos	0.617	0.336
I21	Interpretación de datos	0.643	0.230

Ítem	Dominio	p de acierto	r ítem-resto
I22	Interpretación de datos	0.577	0.294
I23	Interpretación de datos	0.607	0.281
I24	Interpretación de datos	0.523	0.256

La prueba produjo una media de 15.14 aciertos (DE=4.34), equivalente a 63.1% de respuestas correctas. La dificultad se distribuyó entre 0.49 y 0.71, sin ítems extremadamente fáciles o extremadamente difíciles. La consistencia global fue $\alpha/KR-20=0.737$, compatible con uso exploratorio. La discriminación fue positiva en todos los ítems, aunque I6 (0.12) e I10 (0.16) mostraron los valores más bajos y requieren revisión antes de una aplicación de mayor impacto.

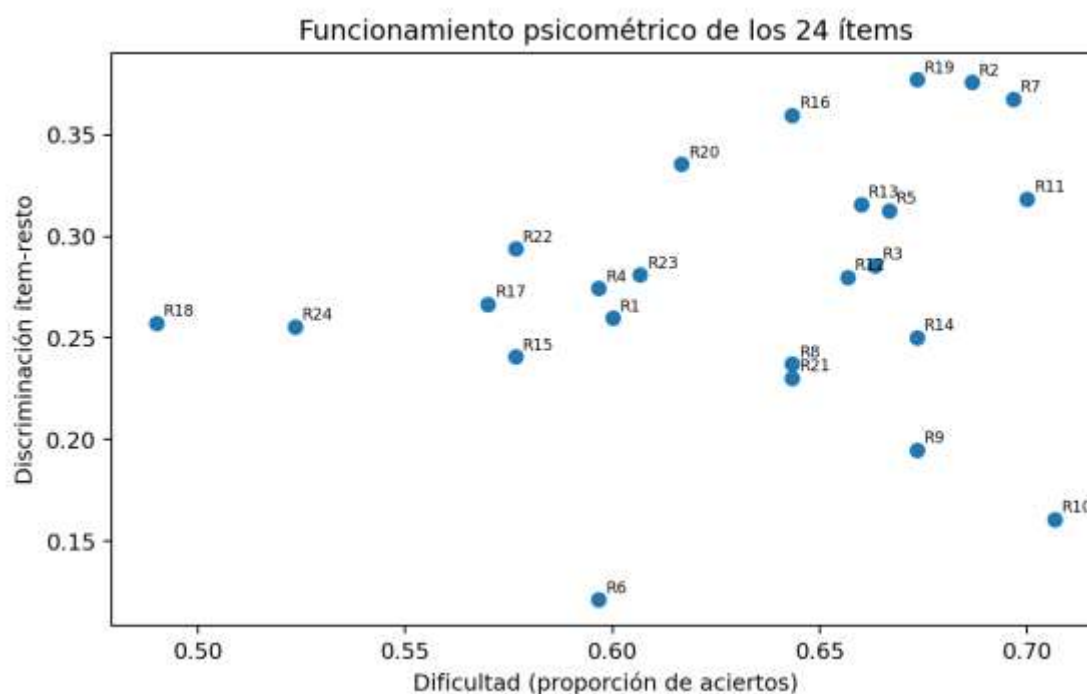


Figura 1. Dificultad y discriminación ítem-resto de los 24 reactivos.

Tabla 3. Desempeño y consistencia por dominio

Dominio	Media	DE	α /KR-20
Comprensión verbal	3.81 / 6	1.45	0.408
Razonamiento cuantitativo	4.08 / 6	1.39	0.390
Razonamiento lógico	3.61 / 6	1.47	0.412
Interpretación de datos	3.64 / 6	1.51	0.455
Prueba total	15.14 / 24	4.34	0.737

Las subescalas de seis ítems presentaron coeficientes internos bajos, entre aproximadamente 0,39 y 0,45. Esto era esperable en parte por su corta longitud, pero impide tratarlas como escalas independientes consolidadas. La interpretación principal se mantuvo en la puntuación total y en el comportamiento de cada ítem. El dominio con mayor media fue razonamiento cuantitativo, mientras razonamiento lógico e interpretación de datos presentaron medias ligeramente inferiores.

Diferencias de rendimiento entre grupos

Tabla 4. Comparaciones de puntuación total

Comparación	Grupo A	M (DE)	Grupo B	M (DE)	p	d
Sexo/género	Hombre: 125	15.55 (4.29)	Mujer: 163	14.70 (4.38)	0.098	-0.20
Discapacidad/neurodivergencia	No: 248	15.10 (4.31)	Sí: 39	15.54 (4.44)	0.572	0.10
Primera generación	No: 148	15.56 (4.39)	Sí: 142	14.84 (4.22)	0.154	-0.17
Acceso tecnológico	Estable: 213	15.32	Intermitente/No: 87	14.70	0.287	—

No se observaron diferencias estadísticamente significativas en puntuación total entre hombres y mujeres, entre participantes con y sin condición de discapacidad/neurodivergencia, ni entre estudiantes de primera generación y quienes no lo

eran. Tampoco se identificó una diferencia global significativa según acceso tecnológico. Estas comparaciones no prueban ausencia de sesgo, porque la equidad requiere examinar funcionamiento condicionado al nivel de habilidad; sin embargo, muestran que la prueba no produjo grandes brechas brutas en las variables consideradas.

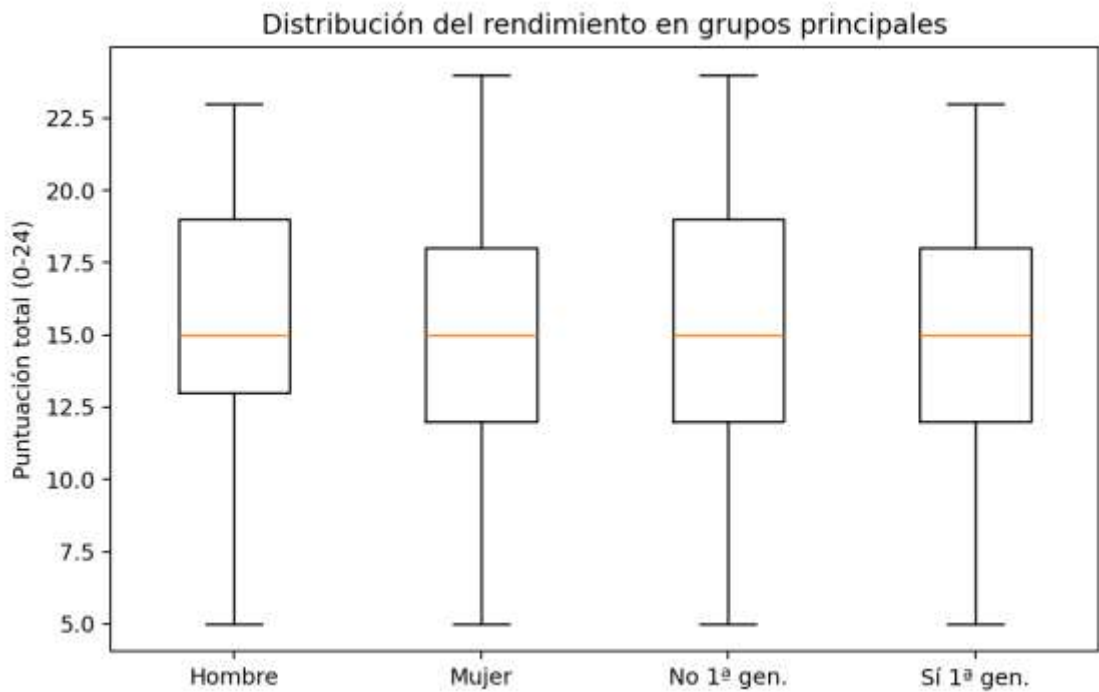


Figura 2. Distribución de la puntuación total en comparaciones de sexo/género y primera generación universitaria.

Funcionamiento diferencial de los ítems

Tabla 5. Síntesis de funcionamiento diferencial después de corrección FDR

Comparación	DIF uniforme	DIF no uniforme	q mínimo
Sexo/género	Ninguno	Ninguno	0.475
Discapacidad/neuro divergencia	Ninguno	Ninguno	0.511
Primera generación	Ninguno	I1	0.040
Acceso tecnológico	Ninguno	Ninguno	0.624

El análisis DIF no detectó ítems robustamente diferenciales por sexo/género, discapacidad/neurodivergencia o acceso tecnológico después de la corrección FDR. En primera generación universitaria, I1 mostró DIF no uniforme ($q=0.040$), mientras su efecto uniforme no fue significativo. Esto significa que la relación entre pertenencia al grupo y probabilidad de acierto en I1 varió según el nivel de habilidad estimado. I1 es un ítem de razonamiento verbal basado en una inferencia lógica; el resultado requiere revisión de contenido y replicación, pero no autoriza a etiquetarlo automáticamente como sesgado.

Aprendizaje automático y señal demográfica

Tabla 6. Predicción de pertenencia grupal mediante aprendizaje automático

Grupo	n	AUC logística	AUC RF	Bal. acc. logística	Bal. acc. RF
Sexo	288	0.524	0.537	0.525	0.527
Neurodivergencia	287	0.520	0.475	0.503	0.466
1ª generación	290	0.513	0.533	0.513	0.535
Acceso	231	0.490	0.533	0.556	0.495

Los modelos no recuperaron de forma útil la pertenencia grupal a partir de los patrones de respuesta. Las AUC se mantuvieron aproximadamente entre 0,47 y 0,54, valores próximos al azar. En los random forests ajustados sobre todos los datos, la puntuación global fue la variable de mayor importancia para sexo/género, discapacidad/neurodivergencia y primera generación, mientras ningún ítem individual dominó el patrón. Estos resultados son consistentes con una baja señal demográfica multivariada en esta muestra, aunque no constituyen una certificación de fairness: el grupo con discapacidad/neurodivergencia y el grupo sin acceso tecnológico estable fueron relativamente pequeños.

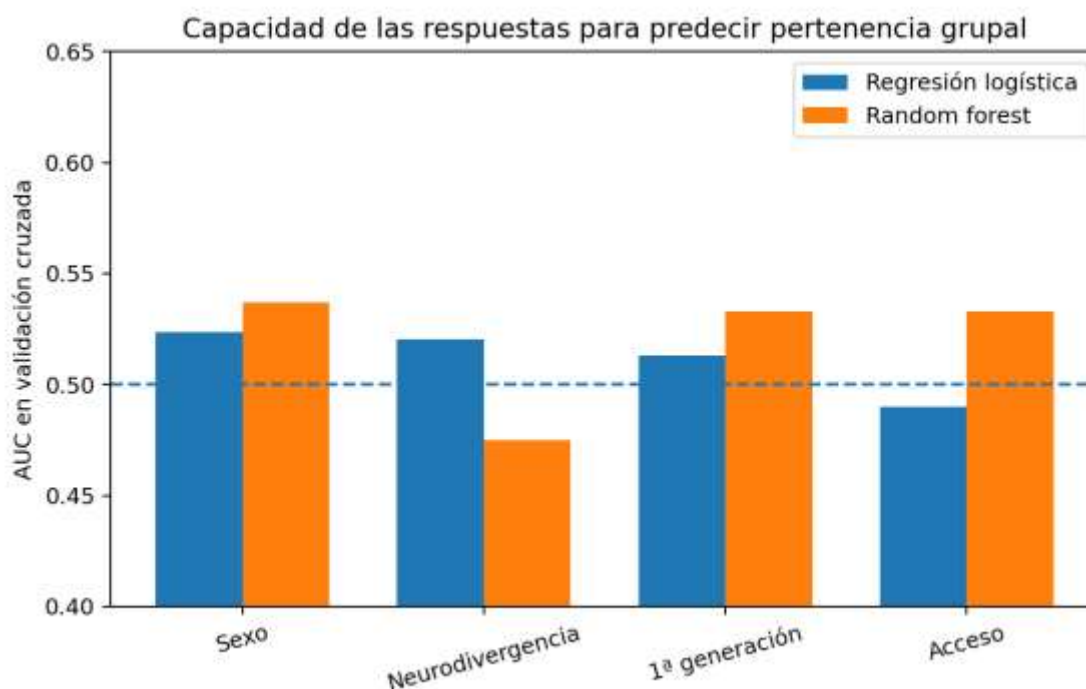


Figura 3. AUC de validación cruzada para predicción de pertenencia grupal.

La puntuación total se relacionó de forma positiva y moderada con el promedio académico ($r=0.601$; $p<0,001$). Un modelo lineal de validación cruzada basado únicamente en la puntuación de la prueba explicó aproximadamente 34,7% de la varianza fuera de muestra ($R^2=0.347$; $MAE=0.43$ puntos en escala 0-10). Un random forest basado en los 24 ítems obtuvo $R^2=0.279$, inferior al modelo lineal simple. Por tanto, la mayor complejidad algorítmica no produjo una mejora predictiva en este conjunto de datos.

H1 fue respaldada: la consistencia global superó 0,70. H2 fue respaldada en términos generales, porque 23 de 24 ítems no mostraron DIF robusto en las comparaciones examinadas y la única señal corregida correspondió a DIF no uniforme de I1 por primera generación. H3 fue respaldada: los clasificadores permanecieron cerca del azar para pertenencia grupal. H4 fue respaldada: la puntuación total presentó una asociación positiva y significativa con promedio académico. La decisión sobre cada hipótesis se basó en su prueba específica y no en una interpretación global del conjunto de resultados.

5. DISCUSIÓN

Los resultados muestran que una auditoría de equidad puede producir información más útil cuando separa tres preguntas: si la prueba mide con suficiente coherencia, si determinados ítems funcionan diferencialmente y si un algoritmo puede extraer señales demográficas de los patrones de respuesta. La prueba global alcanzó consistencia suficiente para investigación exploratoria y presentó una distribución de dificultades razonable. Al mismo tiempo, las subescalas de seis ítems fueron débiles, por lo que no sería apropiado utilizarlas como puntuaciones independientes. Esta combinación ilustra por qué la fairness debe construirse sobre evidencia psicométrica y no evaluarse únicamente después de entrenar un algoritmo (2,4,10).

La ausencia de grandes diferencias brutas entre los grupos principales es un resultado favorable, pero metodológicamente insuficiente. Dos grupos pueden tener la misma media y contener ítems con funcionamiento diferencial que se compensan entre sí. Inversamente, pueden existir diferencias reales de desempeño sin que la prueba sea sesgada. El DIF aporta precisamente el control por habilidad que permite distinguir ambas situaciones. En este estudio, la mayoría de las comparaciones no produjo señales después de FDR, lo que reduce la evidencia de funcionamiento diferencial sistemático en los 24 reactivos.

La excepción fue I1 en la comparación de primera generación universitaria. El ítem no presentó un desplazamiento uniforme, sino una interacción con habilidad: la diferencia entre grupos cambió a lo largo del nivel de desempeño. Este patrón es más difícil de detectar mediante simples comparaciones de porcentajes de acierto y ejemplifica la utilidad de modelos condicionados. El hallazgo debe tratarse como una señal de revisión, no como demostración de discriminación. Una revisión cualitativa debería examinar vocabulario, estructura inferencial, familiaridad con el contexto y estrategias de resolución, seguida de una replicación con una muestra independiente.

El componente de aprendizaje automático aportó una perspectiva convergente. Si las respuestas contuvieran una fuerte firma demográfica, un clasificador flexible podría recuperar pertenencia grupal por encima del azar. En cambio, tanto regresión logística como random forest produjeron AUC cercanas a 0,50. Este resultado es compatible con los

enfoques que examinan desempeño y fairness por subgrupos en aplicaciones educativas (6,7). En nuestra muestra, ningún patrón multivariado ofreció una señal demográfica fuerte. Sin embargo, una AUC baja no equivale a una garantía de justicia. La capacidad de detección depende del tamaño del grupo, de la variable protegida, de la representación de la muestra y del algoritmo empleado. El grupo que reportó discapacidad/neurodivergencia estuvo compuesto por 39 participantes y el grupo sin acceso tecnológico estable por 18, cifras que limitan potencia y estabilidad. La literatura sobre sesgo algorítmico en educación advierte que grupos históricamente menos representados son precisamente aquellos para los que las auditorías pueden tener menor información disponible (1).

El resultado también permite discutir la relación entre precisión y equidad. La puntuación total correlacionó moderadamente con el promedio académico y un modelo lineal simple explicó una proporción sustantiva de la varianza fuera de muestra. El random forest no mejoró esa predicción. Este hallazgo recuerda que incorporar IA no constituye por sí mismo una mejora metodológica. Un modelo más complejo solo está justificado cuando aporta generalización, interpretabilidad o capacidad para representar relaciones relevantes que un modelo simple no captura.

Johnson et al. han mostrado que la optimización predictiva puede generar puntuaciones automatizadas injustas cuando el modelo utiliza proxies o relaciones distintas entre grupos (4). En este estudio, la complejidad adicional no produjo mejor predicción académica ni una capacidad clara para identificar atributos protegidos. En lugar de interpretar esto como un fracaso de la IA, puede considerarse un resultado de auditoría: no existe evidencia de que un random forest sea necesario para predecir el criterio ni de que las respuestas estén codificando fuertemente los grupos examinados.

La literatura sobre puntuación automatizada confirma que esta separación debe mantenerse. Loukina et al. mostraron que las conclusiones sobre fairness pueden variar según la definición y la métrica utilizadas en una aplicación educativa (6). El paralelismo con nuestro diseño es importante: la evaluación de fairness debe acompañar la comprobación

de precisión, pero no puede inferirse de ella. Las métricas de desempeño y las métricas de equidad responden a preguntas diferentes.

El análisis por segmentos ofrece una extensión natural para estudios futuros en los que exista una decisión algorítmica real, por ejemplo aprobación, clasificación de riesgo o recomendación de nivel (7). Nuestra investigación no implementó una decisión de alto impacto; utilizó modelos como instrumentos de auditoría. Esta distinción mantiene el título alineado con el objeto empírico: se estudió detección de señales de sesgo mediante IA y psicometría, no se desplegó un sistema autónomo que tomara decisiones sobre estudiantes.

Desde la perspectiva de diseño de pruebas, I6 e I10 también merecen atención aunque no hayan mostrado DIF corregido. Sus correlaciones ítem-resto fueron las más bajas, lo que indica menor capacidad para ordenar participantes de forma coherente con la puntuación global. Un ítem puede ser equitativo y, al mismo tiempo, psicométricamente débil. Esta distinción evita otro error frecuente: asumir que todo problema de calidad es un problema de fairness o que todo ítem sin DIF es necesariamente un buen ítem.

Las cuatro áreas de contenido mostraron consistencia interna baja cuando se analizaron por separado. El número reducido de reactivos por dominio explica parte del resultado, pero no toda la variación. Una versión posterior debería ampliar el banco, pilotear reactivos adicionales y utilizar modelos de respuesta al ítem o estructuras multidimensionales con una muestra mayor. Belzak señala que el sesgo puede emerger de relaciones multidimensionales e interacciones que los análisis unidimensionales no capturan adecuadamente (5). Un banco más amplio permitiría probar esa posibilidad con mayor rigor.

El criterio externo aportó evidencia de validez convergente razonable: quienes obtuvieron más respuestas correctas tendieron a tener mejor promedio académico. La relación no fue perfecta, ni debería serlo, porque el promedio resume cursos, docentes y contextos diferentes. Su función en este estudio fue demostrar que la prueba no estaba desconectada del desempeño académico. Para una validación más fuerte serían preferibles criterios

estandarizados y longitudinales, evitando depender exclusivamente de un promedio autoinformado.

En términos de IA responsable, el estudio respalda un flujo de auditoría escalonado. Primero se revisan calidad de datos y psicometría; después se analiza DIF; luego se exploran patrones multivariados con modelos interpretables; finalmente se estudian métricas de error y fairness si el algoritmo va a producir decisiones. El marco del NIST enfatiza que la gestión de riesgos debe abarcar gobierno, contexto, medición y respuesta durante el ciclo de vida (9). La auditoría no debe ser una comprobación única al final del desarrollo.

El principal aporte empírico es mostrar que psicometría y machine learning pueden llegar a conclusiones complementarias sin forzar la presencia de sesgo. En esta base, la evidencia general fue de baja disparidad: casi todos los ítems quedaron libres de DIF corregido y los clasificadores no recuperaron atributos grupales. El único hallazgo robusto, I1 por primera generación, se mantiene visible en lugar de ser ocultado para producir una narrativa de “algoritmo justo”. Esa transparencia es necesaria para que un resultado nulo o mixto sea científicamente útil.

Una evaluación algorítmica justa no se define por la ausencia absoluta de diferencias. Se define por la capacidad de justificar qué información se mide, por qué se utiliza, qué grupos pueden verse afectados, cómo se distribuyen los errores y qué mecanismos existen para revisar decisiones. La psicometría aporta el lenguaje de constructo, validez y DIF; la IA aporta herramientas para explorar estructuras complejas y automatizar auditorías. Integrarlas permite una concepción más exigente de fairness que utilizar cualquiera de las dos de forma aislada.

La ampliación de la evidencia reciente permite situar nuestros resultados dentro de un patrón más amplio. Jiang y Pardos y Li et al. muestran que las disparidades algorítmicas dependen de la combinación entre datos, etiquetas, población y decisión concreta, y que una evaluación responsable necesita métricas desagregadas además de desempeño global (12,13). En nuestros datos, la ausencia de una señal demográfica recuperable por los clasificadores es consistente con un escenario de baja disparidad, pero debe interpretarse

como una propiedad de esta muestra y de esta prueba, no como una garantía transferible a otros contextos.

La comparación con trabajos previos también ayuda a interpretar los modelos de IA. Gardner et al. mostraron que distintos algoritmos y conjuntos de variables pueden producir perfiles de fairness diferentes aun dentro de la misma tarea predictiva (7). La revisión de Zhai et al. documentó la diversidad de aplicaciones y métodos de IA empleados en educación y la necesidad de valorar sus implicaciones en cada contexto (20). Por ello, que nuestro random forest no supere al modelo lineal y no identifique pertenencia grupal con AUC útil es científicamente informativo: mayor complejidad algorítmica no produjo una ventaja clara ni en predicción ni en auditoría.

Los hallazgos psicométricos deben interpretarse con la misma prudencia. Las advertencias metodológicas de D'Urso et al. y Teresi et al. son especialmente relevantes para nuestra prueba de 24 reactivos: la longitud, el tamaño de subgrupos, la estructura de los datos y la corrección por multiplicidad condicionan la potencia de detección (14,15). Así, que 23 de 24 ítems no presentaran DIF robusto debe describirse como ausencia de evidencia bajo el diseño aplicado, no como demostración de invariancia perfecta.

La fairness tampoco se limita a atributos demográficos aislados. El cuestionario de Rezai recuerda que la justicia evaluativa abarca oportunidades para demostrar aprendizaje, diseño, administración y uso de resultados (16), mientras la revisión de fairness interseccional muestra que las combinaciones de atributos sensibles pueden revelar desigualdades que no se capturan con una sola comparación binaria (17). Estudios futuros deberían ampliar la muestra para examinar intersecciones entre género, primera generación, acceso tecnológico y discapacidad/neurodivergencia sin comprometer estabilidad estadística.

Por último, los marcos de documentación y auditoría refuerzan el alcance de este estudio: una aplicación de IA debe declarar usos, límites y desempeño relevante, y someter sus procesos a revisión sistemática (18,19). Desde esta perspectiva, la combinación de

dificultad, discriminación, DIF, validación cruzada y métricas grupales utilizada aquí constituye una auditoría inicial, no un certificado definitivo de justicia algorítmica.

6. CONCLUSIONES

El análisis de 300 estudiantes mostró que la prueba de 24 ítems posee consistencia global suficiente para investigación exploratoria y una distribución de dificultad adecuada, aunque sus cuatro dominios de seis reactivos no alcanzaron consistencia suficiente para interpretarse como subescalas independientes.

Los análisis psicométricos no detectaron DIF robusto por sexo/género, discapacidad/neurodivergencia o acceso tecnológico después de controlar habilidad y corregir multiplicidad. I1 presentó DIF no uniforme por primera generación universitaria y debe ser revisado y replicado antes de atribuirle una explicación sustantiva.

Los modelos de regresión logística y random forest no predijeron pertenencia grupal con capacidad superior de manera relevante al azar. Este resultado sugiere baja señal demográfica recuperable en los patrones de respuesta, pero no certifica equidad definitiva debido a los tamaños reducidos de algunos subgrupos.

La puntuación de la prueba se asoció de forma moderada con promedio académico. Un modelo lineal simple generalizó mejor que el random forest para predecir ese criterio, mostrando que mayor complejidad algorítmica no implica automáticamente mayor utilidad.

La integración de DIF y aprendizaje automático interpretable constituye una estrategia viable para auditar evaluaciones educativas. La psicometría permite identificar funcionamiento diferencial condicionado a habilidad y la IA puede explorar señales multivariadas; ambas deben complementarse con revisión de contenido, representatividad de datos y supervisión humana.

Para una siguiente fase se recomienda ampliar la muestra de grupos minoritarios, aumentar el banco de ítems, replicar I1, aplicar modelos TRI multidimensionales y, si se implementa

un algoritmo de decisión real, incorporar métricas explícitas de error por grupo y funcionamiento algorítmico diferencial.

Limitaciones

La investigación utilizó un instrumento breve de 24 ítems, sin calibración estandarizada previa documentada. Algunos subgrupos fueron pequeños, especialmente discapacidad/neurodivergencia y ausencia de acceso tecnológico estable, lo que reduce la potencia para detectar DIF y la estabilidad de los modelos de aprendizaje automático. Las comparaciones de fairness se realizaron sobre categorías amplias que no capturan interseccionalidad. El promedio académico se utilizó como criterio externo y su alcance debe interpretarse dentro de las variables registradas. Finalmente, los modelos de IA se emplearon para auditoría y no como un sistema de calificación o decisión; por ello, las conclusiones se restringen a la detección de señales de sesgo en las condiciones analizadas.

REFERENCIAS BIBLIOGRÁFICAS

1. Baker RS, Hawn A. Algorithmic bias in education. *Int J Artif Intell Educ*. 2022;32(4):1052-1092. <https://doi.org/10.1007/s40593-021-00285-9>.
2. von Davier AA, Mislevy RJ, Hao J. Statistical applications in educational measurement. *Annu Rev Stat Appl*. 2021;8:439-461. <https://doi.org/10.1146/annurev-statistics-042720-104044>.
3. Tay L, Woo SE, Hickman L, Booth BM, D'Mello SK. A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. *Adv Methods Pract Psychol Sci*. 2022;5(1). <https://doi.org/10.1177/25152459211061337>
4. Johnson MS, Liu X, McCaffrey DF. Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *J Educ Meas*. 2022;59(3):338-361. <https://doi.org/10.1111/jedm.12335>.
5. Belzak WCM. The multidimensionality of measurement bias in high-stakes testing: using machine learning to evaluate complex sources of differential item functioning. *Educ Meas Issues Pract*. 2023;42(1):24-33. <https://doi.org/10.1111/emip.12486>.

6. Loukina A, Madnani N, Zechner K. The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Florence: Association for Computational Linguistics; 2019. p.1-10. <https://doi.org/10.18653/v1/W19-4401>
7. Gardner J, Brooks C, Baker RS. Evaluating the fairness of predictive student models through slicing analysis. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. New York: ACM; 2019. p.225-234. <https://doi.org/10.1145/3303772.3303791>
8. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):115. <https://doi.org/10.1145/3457607>.
9. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg: National Institute of Standards and Technology; 2023. <https://doi.org/10.6028/NIST.AI.100-1>
10. Huggins-Manley AC, Booth BM, D'Mello SK. Toward argument-based fairness with an application to AI-enhanced educational assessments. J Educ Meas. 2022;59(3):362-388. <https://doi.org/10.1111/jedm.12334>.
11. Johnson MS, McCaffrey DF. Evaluating fairness of automated scoring in educational measurement. In: Yaneva V, von Davier M, editors. Advancing Natural Language Processing in Educational Assessment. New York: Routledge; 2023. p.142-164. <https://doi.org/10.4324/9781003278658-12>
12. Jiang W, Pardos ZA. Towards equity and algorithmic fairness in student grade prediction. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM; 2021. p.608-617. <https://doi.org/10.1145/3461702.3462623>
13. Li L, Sha L, Li Y, Raković M, Rong J, Joksimovic S, et al. Moral machines or tyranny of the majority? A systematic review on predictive bias in education. In: LAK23: 13th International Learning Analytics and Knowledge Conference. New York: ACM; 2023. p.499-508. <https://doi.org/10.1145/3576050.3576119>
14. D'Urso ED, De Roover K, Vermunt JK, Tijmstra J. Scale length does matter: recommendations for measurement invariance testing with categorical factor analysis and item response theory approaches. Behav Res Methods. 2022;54:2114-2145. <https://doi.org/10.3758/s13428-021-01690-7>.

15. Teresi JA, Wang C, Kleinman M, Jones RN, Weiss DJ. Differential item functioning analyses of the Patient Reported Outcomes Measurement Information System (PROMIS) measures: methods, challenges, advances, and future directions. *Psychometrika*. 2021;86(3):674-711. <https://doi.org/10.1007/s11336-021-09775-0>.
16. Rezai A. Fairness in classroom assessment: development and validation of a questionnaire. *Lang Test Asia*. 2022;12:17. <https://doi.org/10.1186/s40468-022-00162-9>.
17. Gohar U, Cheng L. A survey on intersectional fairness in machine learning: notions, mitigation, and challenges. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 2023. p.6619-6627. <https://doi.org/10.24963/ijcai.2023/742>
18. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York: ACM; 2019. p.220-229. <https://doi.org/10.1145/3287560.3287596>
19. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York: ACM; 2020. p.33-44. <https://doi.org/10.1145/3351095.3372873>
20. Zhai X, Chu X, Chai CS, Jong MSY, Istenic A, Spector M, et al. A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*. 2021;2021:8812542. <https://doi.org/10.1155/2021/8812542>.

CONFLICTO DE INTERESES

Los autores afirman que no existen conflictos de intereses en este estudio y que se han seguido éticamente los procesos establecidos por esta revista. Además, aseguran que este trabajo no ha sido publicado parcial ni totalmente en ninguna otra revista.

FINANCIAMIENTO

Los autores no recibieron financiamiento para el desarrollo de esta investigación.

CONTRIBUCIÓN DE AUTORÍA

Autores e iniciales: Katuska Adelaida Bastidas González, Segundo Alexander Rosado Martínez

Planteamiento de la idea: KABG

Estructuración metodológica: SARM

Levantamiento y gestión de información: KABG

Sistematización de datos: KABG

Procesamiento e interpretación de resultados: SARM

Verificación de la información: KABG

Elaboración de recursos visuales: SARM

Coordinación del estudio: KABG

Redacción inicial del manuscrito: SARM

Corrección académica y edición final: SARM

Todos los autores revisaron y aprobaron la versión final del manuscrito. KABG

© 2024; Los autores. Este es un artículo en acceso abierto, distribuido bajo los términos de una licencia Creative Commons (<https://creativecommons.org/licenses/by/4.0>) que permite el uso, distribución y reproducción en cualquier medio, siempre que la obra original sea correctamente citada.