

Propiedades psicométricas de un instrumento para evaluar resolución de problemas matemáticos e implementación pedagógica de IA

Psychometric properties of an instrument for assessing mathematical problem solving and the pedagogical implementation of Gauth AI

Herrera-Ordoñez, Carlos Gabriel¹

ORCID: <https://orcid.org/0000-0002-4794-9825>

Email: bioecol@gmail.com

Mora-Cifuentes, Paola Milena²

ORCID: <https://orcid.org/0000-0003-0739-4746>

Email: paola.mora@uniminuto.edu

Resumen

El estudio tuvo como propósito diseñar y examinar una batería para valorar la resolución de problemas matemáticos y la implementación de IA; se conformaron dos componentes de diez ítems, una prueba de desempeño puntuada de cero a dos y una escala de implementación con respuesta de uno a cinco; se analizaron 180 registros entre grados décimo y undécimo, valoraciones emitidas por cinco jueces; la validez de contenido alcanzó coeficientes V de Aiken globales de 0,937 y 0,938; la consistencia interna presentó alfa de Cronbach de 0,843 y 0,890, además de omega de McDonald de 0,882 y 0,914; los índices KMO fueron 0,844 y 0,901, mientras las pruebas de Bartlett resultaron significativas; la solución bifactorial explicó 50,9% de la varianza en la prueba y 56,7% en la escala; el acuerdo entre calificadores registró CCI de 0,977 y kappa ponderado de 0,931; los hallazgos respaldaron la coherencia de la batería, la diferenciación de sus dimensiones y la utilidad de la rúbrica; se recomendó ampliar la evidencia mediante análisis confirmatorio, invariancia, sensibilidad al cambio y comparación con criterios externos antes de generalizar la interpretación de las puntuaciones a comunidades educativas.

Palabras clave: resolución de problemas matemáticos; evaluación educativa; propiedades psicométricas; inteligencia artificial.

¹ Posdoctor en Educación y Desarrollo Humano Sostenible, Universidad Pedagógica Experimental Libertador. Doctor en Ciencias de la Educación, Universidad Pedagógica Experimental Libertador. Director Ejecutivo Homework LATAM. Director Ejecutivo Casa de la Poesía ALHIDA. Director Científico del Centro de Estudios Bioecol.

² Doctora en Ciencias de la Educación, Universidad Cuauhtémoc-México. Docente Investigador Full Time, Corporación Universitaria Minuto de Dios (UNIMINUTO). Profesora Líder Saber Pro, Corporación Universitaria Minuto de Dios (UNIMINUTO).

Abstract

This study aimed to design and examine an instrument for assessing mathematical problem solving and the pedagogical implementation of AI; two ten-item components were assembled, comprising a performance test scored from zero to two and an implementation scale rated from one to five; the analysis covered 180 records from grades ten and eleven and ratings provided by five judges; global Aiken's V coefficients reached 0.937 and 0.938; internal consistency yielded Cronbach's alpha values of 0.843 and 0.890 together with McDonald's omega values of 0.882 and 0.914; KMO indices were 0.844 and 0.901, while Bartlett's tests were significant; the two-factor solutions explained 50.9% of variance in the test and 56.7% in the scale; interrater agreement produced an ICC of 0.977 and an average weighted kappa of 0.931; the findings supported the coherence of the instrument, the differentiation of its dimensions, and the usefulness of its scoring rubric; further evidence through confirmatory analysis, measurement invariance, sensitivity to change, and external-criterion comparisons is recommended before score interpretations are extended to additional educational communities.

Keywords: mathematical problem solving; educational assessment; psychometric properties; artificial intelligence.

Introducción

La resolución de problemas matemáticos constituye una práctica intelectual que articula comprensión, selección de estrategias, ejecución, comprobación y argumentación, por lo cual su valoración demanda tareas capaces de registrar el razonamiento empleado y no únicamente la respuesta obtenida; en correspondencia con esta orientación, la OECD (2023) incorpora la formulación, el empleo y la interpretación de las matemáticas como procesos vinculados con contextos auténticos, decisiones justificadas y uso reflexivo de herramientas digitales.

En el ámbito escolar, las aplicaciones sustentadas en inteligencia artificial generativa amplían las posibilidades de acompañamiento mediante explicaciones inmediatas, secuencias adaptadas y retroalimentación conversacional, aunque su valor educativo depende del propósito didáctico, la calidad de las indicaciones y la capacidad del estudiante para contrastar los procedimientos recibidos; al respecto, Miao y Holmes (2023) proponen una integración

centrada en la agencia humana, la protección de datos, la adecuación por edades y la validación pedagógica previa.

Dentro de la educación matemática, Wardat, Tashtoush, AlAli y Jarrah (2023) reconocen que los asistentes generativos pueden facilitar explicaciones y apoyar el estudio de contenidos, a la vez que identifican variaciones en la precisión cuando aumenta la complejidad de las tareas; de manera complementaria, Bulut y Borromeo Ferri (2026) muestran que la interacción estudiantil con un asistente conversacional durante problemas de modelación requiere regulación metacognitiva para evaluar la información, reformular solicitudes y sostener decisiones matemáticas propias.

La evidencia acumulada tampoco autoriza una interpretación uniforme de la eficacia educativa de la inteligencia artificial, pues Eti, Mosia y Egara (2026) encontraron beneficios asociados con retroalimentación adaptada, andamiaje y secuenciación personalizada, junto con riesgos vinculados con dependencia de ayudas, preparación docente insuficiente y desigualdad

tecnológica; en una dirección convergente, Kuo, Bai y Lin (2026) documentan mejores resultados cuando el acompañamiento automatizado se integra con diálogo socrático y estrategias explícitas de resolución.

En relación con IA, la posibilidad de consultar procedimientos paso a paso plantea una tensión pedagógica entre apoyo y sustitución cognitiva, dado que una respuesta técnicamente correcta puede coexistir con comprensión limitada, copia mecánica o ausencia de verificación; esta tensión obliga a distinguir el desempeño matemático alcanzado por el estudiante del grado en que la herramienta fue incorporada a las actividades, de la interacción desarrollada y de la orientación ofrecida por el docente.

La medición de ambos componentes exige argumentos sustentados acerca de las interpretaciones propuestas para las puntuaciones, debido a que la validez no corresponde a una cualidad absoluta adherida al cuestionario y requiere evidencia acumulativa acerca del contenido, la estructura interna, la confiabilidad, la relación con variables externas y la equidad de uso; esta concepción coincide con los estándares formulados por AERA, APA y NCME (2014), los cuales vinculan calidad técnica, finalidad evaluativa y población destinataria.

Bajo tal exigencia, la valoración experta permite examinar claridad, coherencia, relevancia y suficiencia, mientras la consistencia interna informa la precisión de las puntuaciones y el análisis factorial contrasta su organización dimensional; asimismo, las tareas abiertas demandan criterios explícitos de calificación y estimaciones de acuerdo, porque una rúbrica poco delimitada puede introducir variación atribuible al evaluador aun cuando los reactivos representen adecuadamente el dominio matemático.

Con base en esta necesidad, el estudio tuvo como objetivo diseñar y examinar las propiedades psicométricas de una batería

orientada a valorar la resolución de problemas matemáticos y la implementación pedagógica de IA en estudiantes de educación media, mediante evidencia de contenido, consistencia interna, estructura factorial, comportamiento de los ítems y acuerdo entre calificadores, con una organización que permitiera separar los resultados de aprendizaje de la fidelidad de la estrategia educativa.

Metodología

Tipo de estudio

La investigación respondió a un estudio instrumental de orientación cuantitativa, enfocado en la construcción de reactivos, la definición de reglas de puntuación y la obtención de evidencia técnica para las interpretaciones previstas, conforme a una secuencia que integró delimitación del atributo, revisión de contenido, análisis de ítems, estimación de confiabilidad y exploración dimensional; el diseño fue transversal pretest–posttest.

La unidad central de análisis estuvo constituida por las respuestas registradas en cada reactivo, mientras los puntajes totales cumplieron una función complementaria para estudiar precisión y acuerdo; esta decisión evitó confundir el nivel alcanzado en resolución de problemas con la frecuencia de uso de IA, dado que ambos indicadores responden a atributos diferentes y no deben sumarse en una puntuación global.

Participantes

El conjunto analítico reunió 180 registros de estudiantes de educación media pertenecientes a una institución educativa de Bucaramanga, con distribución equilibrada de 90 casos en grado décimo y 90 en grado undécimo; las edades oscilaron entre 15 y 18 años, con media de 16,39 y desviación estándar de 0,79, mientras la codificación alfanumérica impidió incorporar nombres u otros identificadores personales en la matriz de trabajo.

Para la evidencia de contenido se dispuso de cinco valoraciones expertas aplicadas a los veinte reactivos mediante cuatro criterios y una escala ordinal de uno a cuatro; la revisión consideró claridad lingüística, correspondencia con la dimensión, relevancia para el atributo y suficiencia de la cobertura, con cálculo independiente por componente, criterio e ítem.

Materiales e instrumentos

La batería quedó integrada por un componente A de resolución de problemas matemáticos y un componente B de implementación de la estrategia basada en IA; el primero incluyó diez tareas de respuesta construida calificadas con una rúbrica de cero, uno o dos puntos, para un rango total de cero a veinte, mientras el segundo reunió diez enunciados con alternativas Nunca, Casi nunca, Algunas veces, Casi siempre y Siempre, codificadas de uno a cinco, para un rango de diez a cincuenta.

El componente A organizó los reactivos A1–A5 en comprensión del problema y los reactivos A6–A10 en ejecución y verificación del procedimiento; la primera dimensión incluyó identificación de datos relevantes y reconocimiento de la pregunta, mientras la segunda incorporó selección de estrategias, desarrollo de operaciones, comprobación y justificación de la respuesta.

El componente B agrupó los reactivos B1–B5 en uso pedagógico de IA y los reactivos B6–B10 en mediación didáctica con inteligencia artificial; la primera dimensión registró frecuencia e interacción del estudiante con la herramienta, mientras la segunda valoró integración en las actividades matemáticas y orientación docente para revisar, explicar y utilizar responsablemente los resultados.

Tabla 1.

Arquitectura de la batería

Comp.	Dimensión	Ítems	Rango
A	Comprensión del problema	A1–A5	0–10
A	Ejecución y verificación	A6–A10	0–10
B	Uso pedagógico de IA	B1–B5	5–25
B	Mediación didáctica con IA	B6–B10	5–25

Procedimiento

La preparación técnica comenzó con la correspondencia entre variables, dimensiones, subdimensiones, reactivos, formatos de respuesta y rangos de puntuación; después se consolidaron las valoraciones de los jueces, se organizó la matriz de 180 registros, se comprobó la integridad de los rangos permitidos y se calcularon puntuaciones por dimensión y componente, manteniendo separadas las calificaciones de los dos evaluadores para el análisis de objetividad.

El protocolo de uso estableció la aplicación del componente A antes y después de la estrategia, con cincuenta minutos, instrucciones uniformes y ausencia de asistentes digitales durante la prueba; al cierre de la intervención se administró el componente B, se registraron las actividades realizadas con IA y se conservaron por separado las puntuaciones de desempeño, ganancia e implementación.

Análisis de datos

La evidencia de contenido se estimó con el coeficiente V de Aiken, calculado mediante la suma de las puntuaciones ajustadas dividida entre el producto del número de jueces y la amplitud efectiva de la escala, conforme a la formulación de Aiken (1985); se obtuvo el índice por criterio, por reactivo y de manera global para cada componente, tomando 0,80 como referencia descriptiva para examinar la suficiencia inicial.

La consistencia interna se evaluó con alfa de Cronbach y omega de McDonald, dado que el primer coeficiente facilita la comparación con estudios previos y el segundo reduce la dependencia del supuesto de cargas equivalentes, tal como argumentan Hayes y Coutts (2020); la adecuación para el análisis factorial se examinó con KMO y esfericidad de Bartlett, seguida por extracción de ejes principales, dos factores y rotación Varimax.

El comportamiento de los reactivos del componente A se examinó mediante media, desviación estándar, dificultad expresada como proporción del puntaje máximo y correlación ítem-total corregida, mientras el componente B se revisó mediante media, dispersión y correlación corregida; para la objetividad de la rúbrica se calcularon kappa ponderado cuadrático por reactivo y un coeficiente de correlación intraclase de dos vías, efectos aleatorios, acuerdo absoluto y medida individual, siguiendo la nomenclatura recomendada por Koo y Li (2016).

Resultados

La inspección de la matriz confirmó 180 registros completos, ausencia de valores fuera de los rangos previstos y equilibrio entre los grados décimo y undécimo; el puntaje total del primer calificador en el componente A presentó media de 10,78, desviación estándar de 5,00 y recorrido de cero a veinte, mientras el componente B alcanzó media de 32,54, desviación estándar de 8,92 y recorrido de doce a cincuenta.

En la validez de contenido, el componente A obtuvo V global de 0,937 y el componente B alcanzó 0,938; los valores por criterio oscilaron entre 0,920 y 0,967 en la prueba, así como entre 0,920 y 0,960 en la escala, mientras el menor índice por reactivo fue 0,833, resultado que respaldó la conservación de los veinte ítems para el examen de su estructura interna.

La consistencia interna fue favorable en ambos componentes, con alfa de 0,843 y

omega de 0,882 para la prueba, junto con alfa de 0,89 y omega de 0,914 para la escala; los coeficientes por dimensión permanecieron entre 0,822 y 0,87 para alfa, además de 0,875 y 0,906 para omega, lo cual mostró precisión comparable entre las agrupaciones previstas.

Tabla 2.

Síntesis de la evidencia psicométrica por componente

Indicador	Comp. A	Comp. B
V de Aiken global	0,937	0,938
Alfa de Cronbach	0,843	0,890
Omega de McDonald	0,882	0,914
KMO	0,844	0,901
Bartlett $\chi^2(45)$	684,93*	850,48*
Varianza explicada	50,9 %	56,7 %

* $p < 0,001$

La matriz del componente A presentó KMO de 0,844 y una prueba de Bartlett significativa, $\chi^2(45) = 684,93$, $p < 0,001$; la extracción de dos factores explicó 50,9 % de la varianza, con cargas de A1–A5 entre 0,649 y 0,692 en comprensión del problema, cargas de A6–A10 entre 0,659 y 0,817 en ejecución y verificación, además de comunalidades entre 0,445 y 0,683.

En el componente B, el KMO llegó a 0,901 y Bartlett también resultó significativo, $\chi^2(45) = 850,48$, $p < 0,001$; los dos factores explicaron 56,7 % de la varianza, con cargas de B1–B5 entre 0,659 y 0,730 en uso pedagógico, cargas de B6–B10 entre 0,679 y 0,788 en mediación didáctica, además de comunalidades entre 0,493 y 0,675.

Tabla 3.

Cargas factoriales del componente A

Ítem	Dimensión	F1	F2	h ²
A1	Comprensión	0,68	0,11	0,48
		6	5	3
A2	Comprensión	0,65	0,17	0,46
		8	3	3
A3	Comprensión	0,67	0,24	0,52
		9	4	0
A4	Comprensión	0,64	0,18	0,45
		9	4	5
A5	Comprensión	0,69	0,11	0,49
		2	8	3

A6	Ejecución/verificación	0,12 3	0,81 7	0,68 3
A7	Ejecución/verificación	0,19 9	0,67 1	0,49 0
A8	Ejecución/verificación	0,14 8	0,73 8	0,56 7
A9	Ejecución/verificación	0,22 6	0,66 4	0,49 2
A10	Ejecución/verificación	0,10 7	0,65 9	0,44 5

Tabla 4.*Cargas factoriales del componente B*

Ítem	Dimensión	F1	F2	h ²
B1	Uso pedagógico	0,73 0	0,22 8	0,58 5
B2	Uso pedagógico	0,70 7	0,26 0	0,56 7
B3	Uso pedagógico	0,71 8	0,26 2	0,58 5
B4	Uso pedagógico	0,70 2	0,27 8	0,57 0
B5	Uso pedagógico	0,65 9	0,24 1	0,49 3
B6	Mediación didáctica	0,29 1	0,67 9	0,54 6
B7	Mediación didáctica	0,23 4	0,78 8	0,67 5
B8	Mediación didáctica	0,23 8	0,69 4	0,53 8
B9	Mediación didáctica	0,19 7	0,72 9	0,57 0
B10	Mediación didáctica	0,27 7	0,68 4	0,54 4

Los índices de dificultad del componente A se ubicaron entre 0,464 y 0,631, sin concentraciones cercanas a los extremos de la escala, mientras las correlaciones ítem-total corregidas variaron entre 0,482 y 0,591; en el componente B, las correlaciones corregidas se mantuvieron entre 0,589 y 0,663, patrón compatible con una contribución homogénea de los enunciados a sus puntuaciones.

La doble calificación de la prueba produjo un CCI de acuerdo absoluto de 0,977, valor que corresponde a un nivel excelente según Koo y Li (2016), mientras los kappas ponderados por reactivo oscilaron entre 0,885 y 0,964, con promedio de 0,931; estos resultados mostraron una aplicación estable

de la rúbrica y baja variación atribuible al evaluador.

Tabla 5.*Análisis de ítems y acuerdo de calificación*

Indicador	Resultado	Lectura
Dificultad A	0,464– 0,631	Rango intermedio
r corregida A	0,482– 0,591	Discriminación positiva
r corregida B	0,589– 0,663	Aporte homogéneo
CCI absoluto	0,977	Acuerdo excelente
Kappa promedio	0,931	Acuerdo excelente

Discusión

Los resultados respaldaron una arquitectura de medición capaz de diferenciar el desempeño matemático y la implementación pedagógica de IA, con evidencia convergente en contenido, precisión, estructura interna y acuerdo de calificación; esta convergencia resulta relevante porque evita atribuir los cambios de aprendizaje a la mera frecuencia de uso de la herramienta y permite examinar la intervención a partir de dos fuentes complementarias que conservan autonomía interpretativa.

En primer término, los coeficientes V de Aiken superiores a 0,90 en ambos componentes mostraron correspondencia favorable entre los reactivos y las dimensiones formuladas, aunque la evaluación del contenido debe comprenderse como una fase de una cadena probatoria más amplia; Aiken (1985) aporta una medida útil para sistematizar el juicio experto, mientras AERA, APA y NCME (2014) recuerdan que ninguna evidencia aislada basta para justificar todas las decisiones derivadas de una puntuación.

En segundo término, la combinación de alfa y omega ofreció una lectura más sólida de la consistencia interna que el empleo exclusivo de alfa, debido a que omega admite cargas factoriales desiguales y representa con

mayor flexibilidad la contribución de los reactivos; los valores obtenidos superaron 0,80 en la prueba, la escala y sus dimensiones, hallazgo que coincide con la recomendación de Hayes y Coutts (2020) acerca de informar estimadores complementarios y evitar conclusiones apoyadas en un coeficiente único.

La solución de dos factores por componente reprodujo la organización conceptual prevista, con cargas primarias superiores a las secundarias y comunalidades adecuadas, lo cual aportó evidencia inicial para distinguir comprensión de ejecución y verificación en el componente A, además de uso pedagógico y mediación didáctica en el componente B; no obstante, Costello y Osborne (2005) señalan que las decisiones factoriales requieren atención a la extracción, la rotación y la replicación, por lo cual una aplicación posterior deberá contrastar esta estructura mediante análisis confirmatorio en una muestra independiente.

La estructura del componente A dialoga con la tradición de Pólya (1945), al reconocer que comprender el problema, formular un plan, ejecutar y revisar constituyen momentos relacionados que no pueden reducirse a una operación correcta; por ello, la rúbrica valora identificación de información pertinente, modelación, desarrollo, comprobación y justificación, dimensiones que permiten analizar cómo el estudiante procesa una tarea aun cuando el resultado numérico sea incompleto.

En cuanto al componente B, la diferenciación entre uso pedagógico y mediación didáctica resulta coherente con la evidencia reciente acerca de inteligencia artificial en matemáticas, debido a que los beneficios dependen de la calidad del andamiaje, la retroalimentación y la regulación docente; Eti, Mosia y Egara (2026) describen mejoras vinculadas con personalización y apoyo oportuno, mientras Kuo, Bai y Lin (2026) muestran que el

diálogo guiado y las estrategias de resolución fortalecen el valor formativo del asistente.

El alto acuerdo entre calificadores indicó que los criterios de cero, uno y dos puntos delimitaron con precisión los niveles de respuesta, resultado especialmente importante en tareas construidas donde la variación interpretativa puede afectar la comparación pretest–posttest; de acuerdo con Koo y Li (2016), la especificación del modelo de CCI, el tipo de medida y la definición de acuerdo debe acompañar el coeficiente, por lo cual el estudio informó una modalidad de dos vías, efectos aleatorios, acuerdo absoluto y medida individual.

En el plano pedagógico, el instrumento permite examinar una estrategia de IA sin equiparar interacción tecnológica con aprendizaje, pues la escala registra cuándo se consulta, cómo se reformulan los datos, si se comparan procedimientos y qué orientación ofrece el docente, mientras la prueba exige resolver sin apoyo digital; esta separación favorece interpretaciones prudentes acerca de la autonomía matemática, la verificación de resultados y el uso responsable de explicaciones automatizadas, en coherencia con las orientaciones de Miao y Holmes (2023).

Entre las limitaciones técnicas se reconoce que el análisis factorial exploratorio impuso dos factores y utilizó rotación ortogonal, aunque las dimensiones educativas podrían mantener correlaciones; por tal motivo, futuras aplicaciones deberán comparar soluciones oblicuas, trabajar con correlaciones policóricas para los reactivos ordinales, estimar intervalos de confianza para V de Aiken y los coeficientes de confiabilidad, evaluar funcionamiento diferencial e invariancia por grado, edad y sexo, además de relacionar las puntuaciones con criterios externos de desempeño matemático.

Asimismo, la utilidad del componente A para medir cambio requiere examinar

equivalencia entre momentos, posibles efectos de práctica, sensibilidad a la intervención y estabilidad de la rúbrica; el componente B necesita evidencia adicional acerca de deseabilidad social, comprensión de las categorías de respuesta y correspondencia con registros de uso, observaciones de aula o bitácoras docentes, de manera que la implementación percibida pueda contrastarse con información conductual y pedagógica.

Conclusiones

La batería integró dos componentes diferenciados y complementarios para estudiar resolución de problemas matemáticos e implementación pedagógica de IA, con reactivos que abarcaron comprensión, estrategia, ejecución, comprobación, interacción tecnológica y orientación docente, además de reglas de puntuación compatibles con comparaciones por dimensión y momento de aplicación.

La evidencia analizada mostró validez de contenido favorable, consistencia interna superior a 0,80, matrices adecuadas para análisis factorial, soluciones bifactoriales interpretables, discriminación positiva de los reactivos y acuerdo excelente entre calificadores, conjunto de hallazgos que respaldó la coherencia técnica inicial del instrumento y la permanencia de sus veinte ítems.

La interpretación educativa deberá fortalecerse con nuevas aplicaciones que incorporen análisis confirmatorio, invariancia, sensibilidad al cambio, evidencia convergente y comparación con criterios externos, además de protocolos que aseguren acompañamiento docente, verificación matemática y uso responsable de la inteligencia artificial durante la intervención.

Referencias

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological*

Measurement, 45(1), 131–142.
<https://doi.org/10.1177/0013164485451012>

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bartlett, M. S. (1954). A note on the multiplying factors for various chi square approximations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2), 296–298.

Bulut, N., & Borromeo Ferri, R. (2026). AI-powered learning: Exploring secondary school students' use of ChatGPT for solving mathematical modelling problems. *ZDM–Mathematics Education*. Advance online publication.
<https://doi.org/10.1007/s11858-026-01803-6>

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220.
<https://doi.org/10.1037/h0026256>

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, 10, Article 7.
<https://doi.org/10.7275/jyj1-4868>

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
<https://doi.org/10.1007/BF02310555>

Eti, N., Mosia, M., & Egara, F. O. (2026). The role of AI-driven personalised learning in enhancing mathematics problem-solving skills: A systematic review. *Frontiers in Computer Science*, 8, 1813431.
<https://doi.org/10.3389/fcomp.2026.1813431>

- Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability, but.... *Communication Methods and Measures*, 14(1), 1–24. <https://doi.org/10.1080/19312458.2020.1718629>
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31–36. <https://doi.org/10.1007/BF02291575>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kuo, B.-C., Bai, Z.-E., & Lin, C.-H. (2026). Developing an AI learning companion for mathematics problem solving in elementary schools. *Computers & Education*, 240, 105463. <https://doi.org/10.1016/j.compedu.2025.105463>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
- OECD. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. <https://doi.org/10.1787/dfe0bf9c-en>
- Pólya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.
- Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), em2286. <https://doi.org/10.29333/ejmste/13272>