Comparativa entre IAGen y humanos: ítems sobre lengua escrita para un examen de alto impacto

KARLA KARINA RUIZ MENDOZA[I]
http://dx.doi.org/10.22347/2175-2753v17i57.5108

Resumen
El presente estudio se enfoca en la emergente interacción entre la Inteligencia Artificial Generativa y la educación, destacando el uso de tecnologías como el Procesamiento de Lenguaje Natural y modelos específicos como *ChatGPT* de *OpenAI* para su inclusión en Exámenes de Alto Impacto. El objetivo es valorar el uso de *ChatGPT*, en su versión 4.0. para generar ítems de Lengua Escrita y compararlos con los ítems creados por humanos. Los ítems, piloto, pertenecen al Examen de Ingreso a la Educación Superior (ExIES) de la Universidad Autónoma de Baja California. Se aplicaron análisis a partir de la Teoría de Respuesta al Ítem (TRI) a las respuestas de 2,263 sustentantes. Los resultados indicaron que, aunque los ítems generados por *ChatGPT* tienden a ser de mayor dificultad, ambos tipos de ítems son comparables en términos de ajuste al modelo Rasch y capacidad para discriminar entre diferentes niveles de habilidad de los estudiantes. Este hallazgo sugiere que la IAG puede complementar eficazmente la labor de los evaluadores en la elaboración de exámenes a gran escala. Asimismo, *ChatGPT* 4.0 muestra una capacidad superior para discriminar entre diferentes niveles de habilidad de los estudiantes. En conclusión, se subraya la importancia de seguir explorando el uso de la IAG en procesos de evaluación, así como valorar las posibilidades para enriquecer las prácticas pedagógicas de los educadores.
Palabras clave: inteligencia artificial; *ChatGPT*; evaluación de la educación; prueba; digitalización.

Submetido em: 07/10/2024
Aprovado em: 18/12/2025

[I] Universidad Autónoma de Baja California (UABC), Mexicali, México; https://orcid.org/0000-0001-8978-8364; e-mail: ruiz.karla32@uabc.edu.mx.

Comparison between GenAI and humans: creating written language items for a high-impact exam

Abstract
This study focuses on the emerging interaction between Generative Artificial Intelligence (AI) and education, highlighting the use of technologies such as Natural Language Processing and specific models like OpenAI's *ChatGPT* for inclusion in High-Stakes Examinations. The objective is to evaluate the use of *ChatGPT*, version 4.0, for generating Written Language items and to compare them with human-created items. The pilot items belong to the Higher Education Entrance Examination (ExIES) at the Autonomous University of Baja California. Item Response Theory (IRT) analyses were applied to the responses of 2,263 examinees. The results indicated that, although the items generated by *ChatGPT* tend to be more challenging, both types of items are comparable in terms of fit to the Rasch model and their ability to discriminate between different levels of student ability. This finding suggests that Generative AI can effectively complement the work of evaluators in the development of large-scale examinations. Furthermore, *ChatGPT* 4.0 demonstrates superior capability in discriminating between different levels of student ability. In conclusion, the study underscores the importance of continuing to explore the use of Generative AI in assessment processes and evaluating the potential to enhance educators' pedagogical practices.
Keywords: artificial intelligence; *ChatGPT*; educational evaluation; test; digital technology.

Comparação entre IAGen e humanos: itens de linguagem escrita para um exame de alto impacto

Resumo
Este estudo centra-se na interação emergente entre Inteligência Artificial Generativa e educação, destacando a utilização de tecnologias como o Processamento de Linguagem Natural e modelos específicos como o *ChatGPT* da *OpenAI* para inclusão em exames de alto impacto. O objetivo é avaliar a utilização do *ChatGPT*, versão 4.0, para gerar itens de linguagem escrita e compará-los com itens criados por humanos. Os itens piloto pertencem ao Exame de Acesso ao Ensino Superior (ExIES) da Universidade Autônoma da Baja California. A análise da Teoria de Resposta ao Item (TRI) foi aplicada às respostas de 2.263 participantes do exame. Os resultados indicaram que, embora os itens gerados pelo *ChatGPT* tendam a ser mais difíceis, ambos os tipos de itens são comparáveis em termos de ajuste ao modelo de Rasch e capacidade de discriminar entre diferentes níveis de habilidade dos alunos. Esta descoberta sugere que a IA generativa pode complementar eficazmente o trabalho dos avaliadores no desenvolvimento de exames de larga escala. Além disso, o *ChatGPT* 4.0 demonstra uma capacidade superior de discriminar entre diferentes níveis de habilidade entre os alunos. Em conclusão, enfatiza-se a importância de explorar mais a fundo o uso da IAG nos processos de avaliação, bem como avaliar seu potencial para enriquecer as práticas pedagógicas dos educadores.
Palavras-chave: inteligência artificial; *ChatGPT*; avaliação educacional; teste; tecnologia digital.

Introducción

La aparición de Inteligencia Artificial Generativa (IAGen), como mencionan Bozkurt, Karadeniz, Baneres, Guerrero-Roldán y Rodríguez (2021) y Dimitriadou y Lanitis (2023), está en auge y de alguna manera, tanto consciente como inconscientemente, estamos interactuando ya con estas tecnologías; tanto así que desde 2021 se ha observado un aumento en la publicación de artículos sobre la relación IAGen-educación (Bozkurt; Karadeniz; Baneres; Guerrero-Roldán; Rodríguez, 2021). Desde esta perspectiva posthumanista, como lo indican Bozkurt, Karadeniz, Baneres, Guerrero-Roldán y Rodríguez (2021), esta relación humano-no humano es omnipresente, lo que sugiere una interacción humano-tecnología, la cual requiere y requerirá un constante proceso de análisis, no sólo en aspectos prácticos y pragmáticos, si no en torno al uso ético. No obstante, observaremos cada vez más la adopción de este tipo de herramientas y las aulas también se ambientaron a partir de nuevas tecnologías (Dimitriadou; Lanitis, 2023).

Asimismo, la tecnología de Procesamiento de Lenguaje Natural (NLP), que forma parte del campo de la ciencia computacional, impulsada por el aprendizaje automático como *ChatGPT* (de *OpenAI*), siendo un Modelo de Lenguaje Grande (LLM por sus siglas en inglés), ofrece una interacción avanzada capaz de generar respuestas humanas a preguntas en lenguaje natural. *ChatGPT*, como IAGen, se entrena con extensos datos, lo que le permite comprender y responder preguntas de manera coherente con las políticas de aplicación y la disponibilidad de datos (Susnjak, 2022; OpenAI, 2023). La tokenización es un paso esencial en el NLP para organizar información no estructurada en texto apto para el procesamiento informático (Hosseini; Ramussen; Resnik, 2024). Así, *ChatGPT*, al ser interactivo, puede comprender solicitudes y generar respuestas específicas, diferenciándose de motores de búsqueda tradicionales que solo proporcionan enlaces a información pertinente, por ejemplo, *Google*.

En la literatura reciente (2024–2025) se ha intensificado el análisis sobre el papel de la IAGen en el diseño de ítems y en la sustentación de inferencias de validez mediante enfoques argumentativos. En particular, se ha propuesto que los modelos de lenguaje pueden integrarse al desarrollo de instrumentos siempre que se acompañen de especificaciones claras del constructo, revisión experta, y evidencia empírica del

funcionamiento psicométrico antes de su uso operativo (Laverghetta Jr.; Luchini; Linell; reiter-Palmon; Beaty, 2024).

De manera consistente, se ha documentado que los ítems producidos con apoyo de IAGen pueden alcanzar niveles aceptables de calidad, pero su desempeño puede variar por dominio y exige control editorial y verificación sistemática (Küchemann; Rau; Schmidt; Kuhn, 2024). Asimismo, en el ámbito de la evaluación educativa, se ha enfatizado que la incorporación de IAGen debe vincularse con decisiones explícitas de gobernanza del proceso, criterios de calidad y consecuencias de uso, especialmente en contextos de evaluación de alto impacto (Weng; Xia; Gu; Rajaram; Chiu, 2024).

Aunque actualmente son escasos los artículos sobre comparaciones entre el uso de IAGen como *ChatGPT* 3.5 o 4.0, e incluso versiones más recientes, Barrot (2023) ofrece sugerencias para los profesores de inglés en el área de escritura (*writing*) en L2, sobre cómo integrar *ChatGPT* en la práctica pedagógica para capitalizar su uso, mientras se resuelven diversas consideraciones éticas o de procedimientos no abordados; por ejemplo, en los Estándares para Pruebas Educativas y Psicológicas de la *American Educational Research Association* (AERA), *American Psychological Association* (APA), y la *National Council on Measurement in Education* (NCME) (2018).

Estas sugerencias de Barrot (2023) incluyen enfatizar el valor del proceso de escritura, fomentar una voz e identidad de escritura distintivas y utilizar las capacidades de edición o corrección de *ChatGPT* para enseñar formas y estilos del lenguaje correctos. En este sentido, la Tabla 1 sintetiza un balance pedagógico, donde las ventajas se concentran en apoyo formativo (retroalimentación, generación/organización de ideas, corrección), mientras que las desventajas se relacionan con riesgos de validez educativa (respuestas inexactas), agencia del estudiante (dependencia), evaluación auténtica (distinguir autoría) e integridad académica (plagio/ética).

Tabla 1 - *Ventajas y desventajas del uso de ChatGPT en la enseñanza de la escritura*

Ventajas	Desventajas
Retroalimentación adaptativa y oportuna: *ChatGPT* puede ofrecer retroalimentación personalizada y práctica de escritura en cualquier momento.	**Respuestas inexactas:** El *chatbot* puede producir respuestas que no son precisas o relevantes para la consulta.
Es una base de datos e información: Con acceso a una vasta base de conocimiento, *ChatGPT* puede ser una fuente de entrada de lenguaje valiosa.	**Dependencia de los estudiantes:** Existe la preocupación de que los estudiantes puedan depender demasiado de *ChatGPT*, potencialmente afectando su creatividad y pensamiento crítico.
Generación de contenido coherente y gramaticalmente correcto: *ChatGPT* puede ayudar a los usuarios a refinar su escritura y mejorar su uso de formas de lenguaje.	**Dificultad para distinguir entre el trabajo del estudiante y el texto generado por *ChatGPT*:** Esto podría complicar la evaluación de la escritura por parte de los docentes.
Asistencia en la generación de temas y organización de ideas: Puede generar temas de ensayos y crear esquemas en varios formatos.	**Rigidez de plantilla y limitación en la verificación de plagio:** *ChatGPT* sigue estructuras específicas y puede tener dificultades para ajustar textos a un grupo específico de audiencia o detectar plagio.
Herramienta de corrección de escritura automatizada: Ofrece funciones útiles vinculadas a la evaluación de la escritura, incluida la calificación automática y la retroalimentación específica.	**Cuestiones éticas y de integridad académica:** El uso de *ChatGPT* plantea desafíos para la integridad académica y la pedagogía de la escritura.

Fuente: La autora (2025) basado en Barrot (2023).

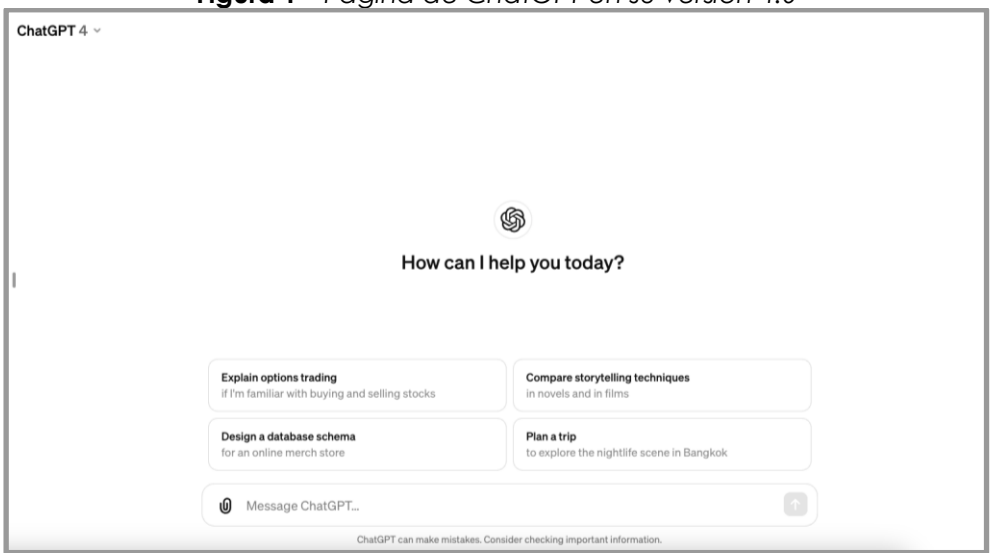
Por lo dicho con anterioridad, hoy en día sabemos que *ChatGPT* puede ser utilizado de diversas maneras en la educación, desde formular preguntas de información hasta generar opciones de múltiples respuestas para exámenes. Nasution (2023) menciona que, aunque *ChatGPT* actualmente requiere instrucciones explícitas para crear instrumentos de evaluación precisos, no se descarta que en el futuro pueda generar preguntas complejas de forma autónoma con suficientes datos y entrenamiento.

Nasution (2023) realizó un estudio con IAG, específicamente con *ChatGPT* pero no declaró cuál versión utilizó, donde evaluó la validez y confiabilidad de 21 preguntas generadas por *ChatGPT*. Las respuestas fueron de opción múltiple de temas básicos de biología y se administró a 272 estudiantes en una universidad en Indonesia. El análisis lo

realizó utilizando la correlación de Pearson para la validez y alfa de Cronbach para la confiabilidad. De los 21 ítems, sólo uno fue descartado, con un alfa de Cronbach de 0.655, donde la mayoría eran de dificultad fácil o media, con un poder de discriminación de adecuado a bueno. En este sentido, este tipo de estudio demuestra el potencial de *ChatGPT* en la creación de preguntas de elección múltiple confiables para fines educativos. Según la versión que utilizó, Nasution (2023) destaca posibles inexactitudes y sesgos, no obstante, también puede depender del *prompt* utilizado para su elaboración y la versión de *ChatGPT*, como lo indica Ruiz Mendoza (2023).

Ante este panorama, este estudio se centra en valorar la calidad de los ítems generados por *ChatGPT* en su versión 4.0 (véase Figura 1), del área de Lengua Escrita considerando temas como uso de palabras en oraciones, economía del lenguaje, uso efectivo de la semántica, concordancia entre sujeto y verbo, y convenciones de puntuación, a través de los resultados del modelo Rasch y comparándolos con ítems realizados por humanos para comprender mejor los alcances y limitaciones de la IAG. Este tipo de evidencia forma parte de la inferencia de generalización según Chapelle (2021), sobre la recopilación de evidencias de validez, el cual determina que las puntuaciones observadas en el examen ofrecen una estimación confiable de las habilidades del estudiante en las áreas evaluadas, que serían similares en contextos o versiones paralelas del examen.

Figura 1 - *Página de ChatGPT en su versión 4.0*



Fuente: OpenAI (2023).

Metodología

Enfoque Metodológico

Este estudio se llevó a cabo desde un enfoque cuantitativo, como recomienda Creswell (2009), donde se presta especial atención a la validez y confiabilidad de los instrumentos utilizados y a la interpretación de los resultados derivados.

Participantes

La muestra estuvo compuesta por 2,263 sustentantes del Examen de Ingreso a la Educación Superior (ExIES), con una distribución equitativa de género (50.06% mujeres y 49.93% hombres) (véase Tabla 2). Este examen se realizó en noviembre de 2023 bajo el contexto de un pilotaje especial del Instituto de Investigación y Desarrollo Educativo (IIDE) de la Universidad Autónoma de Baja California. Cabe mencionar que el ExIES es un Examen de Alto Impacto (EAI) o de altas consecuencias (American Education Research Association; American Psychological Association; National Council on Measurement in Education, 2018; México, 2017), el cual implica una revisión constante por las implicaciones y consecuencias en la toma de decisiones (Shepard, 2006). Así, este tipo de análisis se vinculan directamente con la recopilación de evidencias para fundamentar si es pertinente o no el uso del mismo (American Education Research Association; American Psychological Association; National Council on Measurement in Education, 2018; Chapelle, 2021); en este caso, de los ítems.

Tabla 2 - *Prueba t para muestras independientes de la variable Sexo para confirmar la inexistencia de sesgos*

Variable	Respuesta	N	%	Media	*DS*	*p* valor
Sexo	Mujer	1,133	50.06	997.8	58.23	.250
	Hombre	1,130	49.93	1,000.7	60.46	
	Total	2,263				

Fuente: Instituto de Investigación y Desarrollo Educativo (2024).

Instrumentos

- ExIES: Este examen evalúa competencias en Lengua Escrita y está diseñado para aplicarse a gran escala; además incluye las áreas de Comprensión Lectora y Matemáticas, sin embargo, en estas dos áreas no se elaboraron ítems con

ChatGPT. Incluye ítems ancla y piloto distribuidos en varias subversiones del examen. Es importante resaltar que esta versión se contó con seis subversiones (véase Figura 2) con 36 ítems ancla por versión y 14 ítems piloto[II] para todas las áreas, donde en cada subversión se colocaron de cuatro a cinco ítems elaborados con *ChatGPT* 4.0 como parte de los ítems piloto.

Figura 2 - *Subversiones por forma consolidada del ExIES*

Forma A	**Subversión 1** (5 ítems con *ChatGPT* 4.0)	**Forma C**	**Subversión 4** (5 ítems con *ChatGPT* 4.0)
	Subversión 2 (4 ítems con *ChatGPT* 4.0)		**Subversión 5** (5 ítems con *ChatGPT* 4.0)
	Subversión 3 (5 ítems con *ChatGPT* 4.0)		**Subversión 6** (4 ítems con *ChatGPT* 4.0)

Fuente: Instituto de Investigación y Desarrollo Educativo (2024).

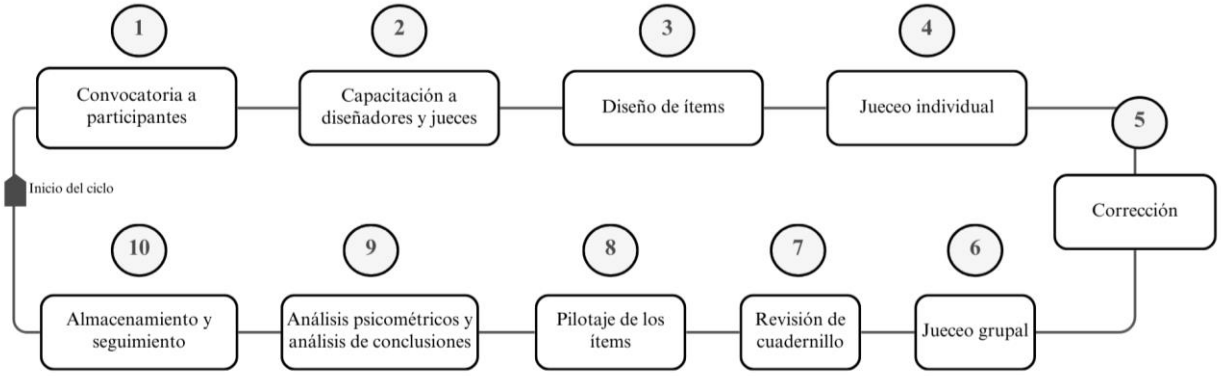
- Generación de ítems con IAG: Se utilizó la versión 4.0 de *ChatGPT* para generar 28 ítems piloto, utilizando especificaciones detalladas en el manual de Lengua Escrita del ExIES (Instituto de Investigación y Desarrollo Educativo, 2024). Se siguieron criterios basados en la Taxonomía de Anderson y Krathwohl (2001) para asegurar que los ítems reflejen el nivel adecuado de demanda cognitiva.

Procedimiento

El proceso del examen es riguroso, por lo que se cuenta con bases sólidas para el diseño de su proceso, tanto para la parte aplicando la IAG como la Humana (Jornet Meliá; González Such; Suárez Rodríguez, 2010; Kolen; Brennan, 2014; Lane; Raymond; Haladyna, 2015). En este estudio se implementó un proceso híbrido, donde se integró a *ChatGPT* 4.0 como diseñador y como juez. Como se observa en la Figura 3, este proceso contempló diez pasos fundamentales para su implementación.

[II]Según el Instituto de Investigación y Desarrollo Educativo (2024), los "Ítems piloto. Son ítems recién diseñados cuyas evidencias de validez de contenido han sido probadas, pero no sus propiedades métricas. No son tomados en cuenta para medir el desempeño de los sustentantes, su único objetivo es ser probados en campo en condiciones iguales o similares a una aplicación común para valorar si sus características respaldan su inclusión en el banco de ítems general del instrumento."

Figura 3 - *Proceso del desarrollo de la prueba ExIES*



Fuente: La autora (2025).

Después de la convocatoria y capacitación, de los diseñadores y jueces, se pasa al diseño de los ítems. Para el desarrollo de ítems se hizo uso de la IAG *ChatGPT* (chat.openai.com) en su versión de paga nombrada versión 4.0. Para cada uno de los ítems se hicieron conversaciones separadas, atendiendo a problemáticas y equivocaciones cuando el *chat* se satura; en este mes no estaba disponible la opción de crear tu propio *chat* con características específicas. En este sentido, cada ítem fue solicitado con las características del manual de Lengua Escrita (Instituto de Investigación y Desarrollo Educativo, 2024), comenzando por el uso de la Taxonomía de Anderson y Krathwohl (2001) para el nivel de demanda cognitiva según la tabla de especificaciones. En cada *prompt* creado se especificó:

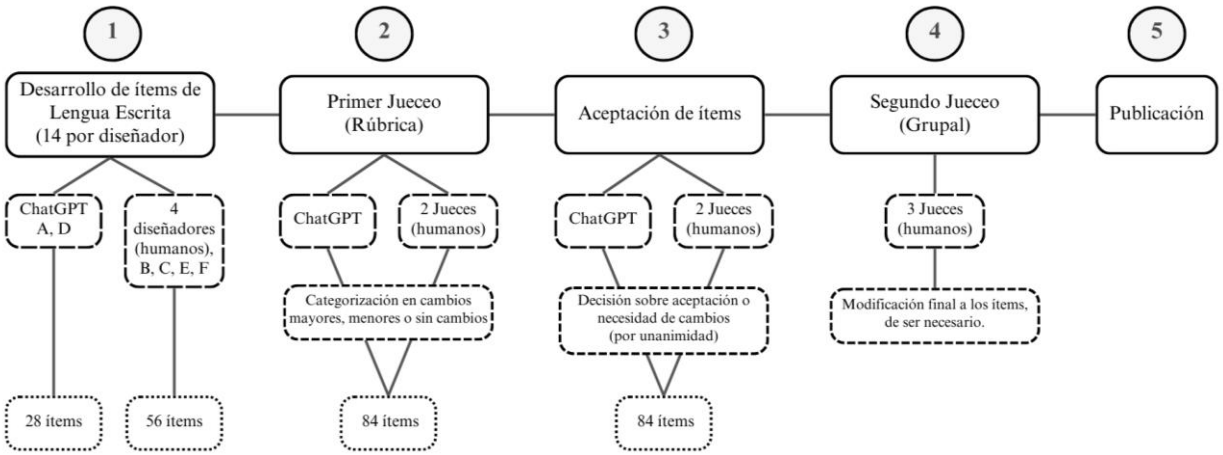
1. Identificación del contenido a evaluar;
2. Descripción del contenido a evaluar:
    a. Interpretación;
    b. Ejemplos;
    c. Delimitación del contenido;
    d. Conocimientos y habilidades previas;
    e. Actividades cognoscitivas.
3. Plantilla del ítem:
    a. Estructura base del ítem;
    b. Características del texto;
    c. Estructura y descripción de respuesta correcta y distractores.
4. Peculiaridades de la plantilla:

a. Base del ítem;
b. Vocabulario empleado;
c. Edición;
d. Peculiaridades de los distractores;

5. Bibliografía consultada.

Posterior a la elaboración de ítems se pasa a un proceso de jueceo independiente, donde *ChatGPT* 4.0 realizó la revisión y corrección de los ítems. Después, se pasó a un proceso de jueceo grupal donde se sometieron a dos jueces humanos y a *ChatGPT* 4.0. En este jueceo grupal no se observaron diferencias significativas entre humanos y el uso de *ChatGPT* en el jueceo; sin embargo, la implementación final de los cambios fue realizada por el propio *chat*. Por lo tanto, los resultados de este estudio se interpretan como un proceso híbrido de validación/edición (Bozkurt; Karadeniz; Baneres; Guerrero-Roldán; Rodríguez, 2021) (véase Figura 3 y 4).

Figura 4 - *Proceso de evaluación y revisión de los ítems*



Fuente: La autora (2025).

Los temas que se abordaron por *ChatGPT* 4.0 se visualizan en la Tabla 4, con el fin de realizar una comparación con ítems de la misma temática, se seleccionaron dos temas para poder realizar las comparaciones pertinentes entre humanos e IAG, estos se seleccionaron de forma aleatoria. En total, se compararon 18 ítems, 9 humanos y 9 con *ChatGPT*, a través de la técnica *t-student* (Field, 2013), así como del análisis Rasch

expresados en el reporte técnico del ExIES (Instituto de Investigación y Desarrollo Educativo, 2024).

Tabla 4 - *Ítems retomados para el estudio*

Tema	Ítems elaborados con *ChatGPT*	Ítems elaborados con humanos para comparar
Uso efectivo de la semántica: sinónimos	5	
Concordancia entre sujeto y verbo	5	5
Convenciones de puntuación: coma	5	
Convenciones de puntuación: interrogación	5	
Economía del lenguaje	4	
Uso de frases o palabras en oraciones	4	4
Total	28	9

Fuente: La autora (2025).

Análisis de los datos

El estudio de las métricas de cada aplicación se basó en el modelo Rasch. Este modelo probabilístico predice la probabilidad de que una persona seleccione la respuesta correcta de un ítem en función de la diferencia entre la dificultad del ítem y el nivel de habilidad del sustentante (Tristán López, 1998). Asimismo, se aplicaron la prueba t de Student y la prueba U de Mann–Whitney en SPSS para comparar los ítems elaborados por *ChatGPT* y por humanos.

Los parámetros estimados de los ítems del examen se basaron en las recomendaciones de Jurado-Núñez, Flores-Hernández, Delgado-Maldonado, Sommer-cervantes, Martínez-González y Sánchez-Mediola (2013) y Ghio, Morán, Garrido, Azpilicueta, Córtez y Cupani (2020). La discriminación de los ítems se evaluó como excelente para valores iguales o superiores a 0.40, bueno con posibilidad de mejora para valores entre 0.30 y 0.39, marginal requiriendo revisión para valores entre 0.20 y 0.29, y pobre sugiriendo una revisión profunda o descarte para valores inferiores a 0.20. Los índices Outfit e Infit t (Zstd) se consideraron razonables si estaban dentro del rango de -2

a 2, muy impredecibles si eran menores a -2 y muy predecibles si excedían 2. Además, los índices Infit y Outfit MSQ se interpretaron como ideales entre 0.8 y 1.2, aceptables entre 0.7 y 0.8 o entre 1.2 y 1.3, y el coeficiente Ptbis se consideró ideal si era mayor a 0.2, aceptable entre 0.1 y 0.2, y no aceptable si era menor a 0.1 (Jurado-Núñez; Flores-Hernámdez; Delgado-Maldonado; Sommer-Cervantes; Martínez-González; Sánchez-Mendiola, 2013; Ghio; Morán; Garrido; Azpilicueta; Córtez; Cupani, 2020).

A continuación, se presentan los resultados de los ítems generados por *ChatGPT* 4.0, después una comparativa entre los elaborados por humanos vs *ChatGPT*, y terminando con los resultados t-student para confirmar si hay o no diferencia significativa.

Resultados

A. *Resultados de los ítems elaborados por ChatGPT 4.0*

Según los resultados de la Tabla 6 sobre los 28 ítems generados por *ChatGPT* 4.0 para evaluar distintos temas del área de Lengua Escrita, han demostrado en su mayoría una dificultad que oscila entre el rango de medio a difícil, lo cual es indicativo de un desafío apropiado para la evaluación en el nivel educativo correspondiente; en este caso como un examen para el ingreso a la educación superior. Un valor de dificultad promedio de 0.5 es deseable en evaluaciones (Jurado-Núñez; Flores-Hernámdez; Delgado-Maldonado; Sommer-Cervantes; Martínez-González; Sánches-Mendiola, 2013), y varios ítems se acercan a este puntaje con excepción del 1 y 10, proporcionando un equilibrio entre preguntas que todos los estudiantes pueden responder y aquellas que solo pueden responder quienes desarrollan altas capacidades.

En términos de ajuste, la mayoría de los ítems se mantienen cercanos al valor óptimo de 1.0 para Infit y Outfit MNSQ, lo cual sugiere que las respuestas de los sustentantes estuvieron en línea con las expectativas del modelo estadístico. Los valores ZSTD dentro de un rango de -2 a 2 para la mayoría de los ítems indican un buen ajuste. El porcentaje de ítems creados por *ChatGPT* que se encuentran dentro de los parámetros establecidos es del 92.86%. Esto incluye las métricas de Infit y Outfit (MNSQ y ZSTD) (véase Tabla 5).

Tabla 6 - *Resultados de los 28 ítems elaborados por ChatGPT 4.0*

No.	Tema	Demanda cognitiva	Dificultad	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Corr. Punto-biserial	Discriminación
1	Uso de palabras en oraciones	Evaluación	0.39	0.93	-0.8	0.85	-1.1	0.35	1.16
2	Uso de palabras en oraciones	Evaluación	0.65	1.13	0.9	1.27	1.4	0.01	0.81
3	Uso de palabras en oraciones	Evaluación	0.68	1.07	0.5	1.26	1.1	0.04	0.9
4	Uso de palabras en oraciones	Evaluación	0.56	0.96	-0.6	0.96	-0.4	0.32	1.12
5	Economía del lenguaje	Evaluación	0.56	1.05	0.7	1.07	0.8	0.19	0.82
6	Economía del lenguaje	Evaluación	0.58	1.07	0.9	1.1	0.9	0.14	0.79
7	Economía del lenguaje	Evaluación	0.55	1.12	1.8	1.17	1.9	0.05	0.52
8	Economía del lenguaje	Evaluación	0.53	0.95	-0.9	0.93	-1	0.34	1.23
9	Uso efectivo de la semántica	Evaluación	0.6	1.08	1	1.1	0.7	0.13	0.82
10	Uso efectivo de la semántica	Evaluación	0.28	0.96	-0.2	0.82	-0.6	0.25	1.05
11	Uso efectivo de la semántica	Evaluación	0.77	1.07	0.3	1.3	0.8	0	0.93
12	Uso efectivo de la semántica	Evaluación	0.81	1.05	0.2	1.87	1.6	-0.06	0.92
13	Uso efectivo de la semántica	Evaluación	0.6	1	0	1.04	0.3	0.23	0.98
14	Concordancia entre sujeto y verbo	Aplicación	0.68	1.05	0.3	1.19	0.8	0.1	0.93
15	Concordancia entre sujeto y verbo	Aplicación	0.51	1.05	0.9	1.06	0.9	0.18	0.75
16	Concordancia entre sujeto y verbo	Aplicación	0.51	0.98	-0.4	0.96	-0.5	0.28	1.13

Continua

Conclusão

17	Concordancia entre sujeto y verbo	Aplicación	0.49	0.89	-2.3	0.87	-2	0.42	1.64
18	Concordancia entre sujeto y verbo	Aplicación	0.48	0.87	-2.4	0.84	-2.1	0.45	1.63
19	Convenciones de puntuación: coma	Aplicación	0.62	1.01	0.1	1.1	0.7	0.2	0.95
20	Convenciones de puntuación: coma	Aplicación	0.42	0.94	-0.8	0.9	-1	0.33	1.17
21	Convenciones de puntuación: coma	Aplicación	0.51	0.94	-1.1	0.93	-1	0.34	1.31
22	Convenciones de puntuación: coma	Aplicación	0.57	1.07	0.9	1.11	1	0.16	0.79
23	Convenciones de puntuación: coma	Aplicación	0.58	1.03	0.4	1.08	0.7	0.19	0.88
24	Convenciones de puntuación: interrogación	Aplicación	0.43	0.92	-1.1	0.87	-1.2	0.37	1.26
25	Convenciones de puntuación: interrogación	Aplicación	0.51	1.01	0.2	1	0.1	0.24	0.96
26	Convenciones de puntuación: interrogación	Aplicación	0.44	0.88	-1.9	0.85	-1.8	0.42	1.44
27	Convenciones de puntuación: interrogación	Aplicación	0.42	0.88	-1.6	0.84	-1.5	0.42	1.32
28	Convenciones de puntuación: interrogación	Aplicación	0.63	1.13	1.2	1.26	1.4	0.03	0.77
Promedio general			**0.55**	**1.00**	**-0.14**	**1.06**	**0.03**	**0.22**	**1.04**

Fuente: La autora (2025).

Los ítems generados por *ChatGPT* mostraron cumplimiento con los parámetros de Infit MNSQ, donde el 100% de los ítems se situaron dentro del rango óptimo (0.8 a 1.2), indicando un ajuste muy adecuado al modelo esperado para estas evaluaciones. Para Outfit MNSQ, aunque la mayoría (82.14%) de los ítems también se ajustaron bien, un

17.86% no cumplió con este criterio, lo que podría sugerir cierta variabilidad en cómo los ítems se comportan con respecto a las respuestas atípicas de los examinandos. En cuanto a las medidas ZSTD, tanto para Infit (92.86%) como para Outfit (96.43%), la gran mayoría de los ítems estuvieron dentro de los límites aceptados de -2 a 2, indicando una normalidad en la dispersión de las respuestas; sin embargo, un pequeño porcentaje quedó fuera de estos rangos, lo que podría reflejar problemas potenciales de sobreajuste o subajuste en ciertos ítems.

Por otro lado, también se realizó un ejercicio de promediar por cada una de las temáticas y según el nivel cognitivo, esto con el fin de hacer un análisis más detallado, obsérvese la Tabla 7. La variabilidad en la dificultad de los ítems destaca la complejidad de ajustarlos a un amplio espectro de habilidades entre los participantes. Siguiendo el análisis Rasch (1960), comprendemos que la dificultad de un ítem refleja la interacción entre la habilidad del participante y el ítem en sí, lo que resalta la importancia de una calibración precisa para evaluaciones válidas y fiables. Por ejemplo, con el tema "Interrogación" presentando una dificultad más baja (promedio de 0.49) comparado con "Semántica" (promedio de 0.61), vemos cómo el contenido y enfoque de los ítems influyen en la dificultad percibida, subrayando la necesidad de un diseño equilibrado de ítems.

Tabla 7 - *Resultados por temas y demanda cognitiva*

Promedios	Dificultad	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Corr. Punto-biserial	Discriminación
Por tema: Uso de palabras	0.57	1.02	0.00	1.09	0.25	0.18	1.00
Por tema: Economía del lenguaje	0.56	1.05	0.63	1.07	0.65	0.18	0.84
Por tema: Semántica	0.61	1.03	0.26	1.23	0.56	0.11	0.94
Por tema: Sujeto y verbo	0.53	0.97	-0.78	0.98	-0.58	0.29	1.22
Por tema: Uso de la coma	0.54	1.00	-0.10	1.02	0.08	0.24	1.02
Por tema: Interrogación	0.49	0.96	-0.64	0.96	-0.60	0.30	1.15
Demanda cognitiva: Evaluación	0.52	0.99	-0.32	1.02	-0.20	0.25	1.08

Continua

Conclusão

Demanda cognitiva: Aplicación	0.52	0.98	-0.51	0.99	-0.37	0.28	1.13
Promedio general	**0.55**	**1.00**	**-0.13**	**1.06**	**0.04**	**0.22**	**1.03**

Fuente: La autora (2025).

Al considerar el ajuste de los ítems a través de las métricas de Infit y Outfit, se observa cómo estos valores reflejan la alineación de los ítems con las expectativas teóricas. Los ítems cercanos a 1.00 en Infit MNSQ sugieren un buen ajuste, indicando predictibilidad y coherencia con las habilidades de los participantes. Sin embargo, valores altos en estas métricas, lo cual implica la necesidad de un análisis detallado para identificar necesidades de ajuste o reemplazo, manteniendo así la precisión de las evaluaciones.

Asimismo, la capacidad de los ítems para discriminar entre diferentes niveles de habilidad, evidenciada tanto en la correlación punto-biserial como en los índices de discriminación, es crucial. Esta variabilidad en la discriminación, con ejemplos que van desde 0.11 en "Semántica" hasta 0.30 en "Interrogación", ilustra la importancia de alinear los ítems con objetivos educativos claros y relevantes, además de principios psicométricos sólidos.

B. *Resultados comparativos entre ChatGPT 4.0 y Humanos*

Como se observa en la Tabla 8, la diferencia en la dificultad entre los ítems generados por humanos y por *ChatGPT* 4.0 se puede notar, siendo los ítems de *ChatGPT* en promedio más difíciles. Esto podría sugerir que *ChatGPT* tiende a generar preguntas que requieren un nivel más alto de comprensión o habilidad para responder correctamente, lo cual puede ser deseable dependiendo del objetivo de la evaluación. Los resultados generales por ítem se pueden consultar en el Anexo 1. Adicionalmente, realicé una prueba t para comparar los promedios de Infit MNSQ entre los ítems elaborados por humanos y por *ChatGPT* 4.0. Los resultados de esta prueba muestran un valor t de 0.550 y un valor p de 0.590, lo que indica que no hay una diferencia estadísticamente significativa entre los grupos en cuanto a la calidad del ajuste de los ítems, según el Infit MNSQ.t=0.550, p=0.590 (No hay diferencia significativa).

Los valores de Infit ZSTD reflejan la desviación estándar del ajuste del ítem al modelo. Los ítems de humanos muestran una ligera sobre-ajuste, mientras que los de *ChatGPT* muestran un bajo-ajuste. Idealmente, los valores deberían estar cercanos a 0. La diferencia sugiere variaciones en cómo los ítems se ajustan al modelo esperado, pero ninguno de los grupos muestra una desviación muy impredecible o muy predecible. Al igual que con el Infit MNSQ, los valores de Outfit MNSQ cerca de 1.0 son deseables y aquí vemos que ambos grupos están casi igualmente ajustados, indicando que los ítems de ambos grupos muestran un buen ajuste global al modelo. Asimismo, se observa que el 86.11% (Tabla 8), tanto de los ítems elaborados por humanos como por *ChatGPT*, se encuentran dentro de los parámetros señalados.

Tabla 8 - *Resultados comparativos entre humanos y ChatGPT 4.0*

Métrica	Humanos (Promedio General)	*ChatGPT* 4.0 (Promedio General)	Promedio
Dificultad	0.458	0.550	0.5039
Correlación Punto-biserial	0.229	0.233	0.2339
Discriminación	0.919	1.107	1.0189
% de ítems dentro de los parámetros	86.11%	86.11%	86.11%

Fuente: La autora (2025).

En cuanto a la correlación Punto-biserial que indica cómo los ítems discriminan entre participantes de diferentes habilidades, la Tabla 8 apunta a que los resultados son similares entre humanos y *ChatGPT* 4.0, es decir los ítems de ambos grupos tienen una capacidad similar para discriminar entre participantes de diferentes niveles de habilidad. Esta es una métrica clave para la calidad de los ítems, indicando que ambos grupos producen ítems efectivos. Y en general, los ítems generados por *ChatGPT* 4.0 muestran una capacidad superior para discriminar entre estos grupos, lo que sugiere que pueden ser particularmente útiles en evaluaciones.

Y, como se muestra en la Tabla 9, también se optó por comparar los resultados de humanos vs *ChatGPT*, recordando que fueron un total de nueve ítems por cada uno. Al comparar la dificultad de los ítems en las categorías de "Uso de Palabras" y "Sujeto y Verbo", se encontró que los ítems generados por *ChatGPT* 4.0 presentan un mayor nivel de dificultad en comparación con aquellos generados por humanos. Esto sugiere que, desde una perspectiva de diseño de ítems, *ChatGPT* 4.0 tiende a crear preguntas que requieren un mayor nivel de comprensión y análisis por parte de los sustentantes en estas dos temáticas; sin olvidar el proceso que pasan los ítems antes de ser publicados.

Tabla 9 - *Comparativa entre temas: humanos vs ChatGPT*

Métrica	Humanos (Uso de palabras) | Evaluación	*ChatGPT* 4.0 (Uso de palabras) | Evaluación	Humanos (Sujeto y verbo) | Aplicación	*ChatGPT* 4.0 (Sujeto y verbo) | Aplicación
Dificultad	0.5000	0.5700	0.4240	0.5340
Infit MNSQ	1.0150	1.0225	1.0220	0.9680
Infit ZSTD	-0.2000	0.0000	0.6000	-0.7800
Outfit MNSQ	1.0725	1.0850	1.0000	0.9840
Outfit ZSTD	0.0750	0.2500	0.3000	-0.5800
Correlación Punto-biserial	0.2250	0.1800	0.2320	0.2860
Discriminación	1.0725	0.9975	0.7960	1.2160

Fuente: La autora (2025).

En cuanto a la calidad de los ítems basada en los parámetros de ajuste del modelo Rasch, como son Infit y Outfit MNSQ, los ítems de "Sujeto y Verbo" generados por humanos muestran un ajuste más cercano al ideal, indicando que estos ítems se alinean de manera más eficiente con las expectativas del modelo. Esto contrasta con los ítems de "Uso de Palabras", donde la diferencia en el ajuste entre ítems humanos y de *ChatGPT* es menos pronunciada, sugiriendo una similaridad en la calidad de los ítems entre ambas fuentes.

Otro aspecto es la variabilidad en el ajuste, medida a través de Infit y Outfit ZSTD, que resulta ser mayor en los ítems generados por humanos, especialmente en la categoría de "Sujeto y Verbo". Esto podría indicar que, aunque los ítems de *ChatGPT* 4.0 en general se ajustan bien al modelo, hay una mayor inconsistencia en cómo estos ítems se comportan entre diferentes grupos de estudiantes.

La capacidad de discriminación, evaluada a través de la correlación punto-biserial y el coeficiente de discriminación, es generalmente superior en los ítems generados por humanos, resaltando una ventaja notable en "Sujeto y Verbo". Esto significa que los ítems elaborados por humanos son más efectivos al diferenciar entre estudiantes de diferentes niveles de habilidad en esta categoría específica.

C. *Resultados de la prueba t-student*

En el estudio comparativo de los ítems generados por humanos frente a los generados por *ChatGPT* 4.0, se utilizó una prueba *t-student* complementada por la prueba *U de Mann-Whitney* para examinar diferencias en varias métricas de calidad de los ítems. Los resultados indican que la única métrica que mostró una diferencia estadísticamente significativa fue la dificultad, donde los ítems generados por *ChatGPT* 4.0 resultaron ser más difíciles en comparación con aquellos creados por humanos (t = -2.144, U = 0.019, p = 0.037), sugiriendo una mayor capacidad de *ChatGPT* para formular preguntas que plantean un mayor desafío a los evaluados (véase Tabla 10).

Tabla 10 - *Resultados de la prueba t-student*

Métrica	Media (Humanos)	Media (*ChatGPT* 4.0)	Estadístico t	U de Mann-Whitney	Valor p (Levene)	Valor p (Dos colas)	Interpretación
Dificultad	0.458	0.550	-2.144	0.019	0.363	0.037	*ChatGPT* genera ítems más difíciles
Infit MNSQ	1.019	0.992	-0.618	0.489	0.323	0.273	No hay diferencia significativa

Continua

Conclusão

Infit ZSTD	0.244	-0.433	1.024	0.436	0.395	0.161	No hay diferencia significativa
Outfit MNSQ	1.032	1.029	-1.086	1.00	0.185	0.147	No hay diferencia significativa
Outfit ZSTD	0.200	-0.211	0.632	0.489	0.791	0.268	No hay diferencia significativa
Corr. Punto-biserial	0.229	0.239	-0.126	0.931	0.953	0.451	No hay diferencia significativa
Discriminación	0.919	1.119	-1.145	0.436	0.362	0.135	Tendencia mínima a mejor discriminación en *ChatGPT*

Fuente: La autora (2025).

Para las demás métricas evaluadas - Infit MNSQ, Infit ZSTD, Outfit MNSQ, Outfit ZSTD, Correlación Punto-biserial y Discriminación - los resultados no mostraron diferencias estadísticamente significativas. Esto sugiere que tanto los ítems generados por humanos como los de *ChatGPT* 4.0 son comparables en términos de ajuste al modelo y capacidad para discriminar entre diferentes niveles de habilidad de los evaluados. Específicamente, en el caso de la discriminación, aunque no se encontró una diferencia significativa (t = -1.145, U = 0.436, p = 0.135), los resultados apuntan a una ligera tendencia hacia una mejor discriminación por parte de los ítems de *ChatGPT*, lo cual podría implicar un potencial marginalmente superior de *ChatGPT* en la diferenciación entre respuestas de evaluados de diversos niveles de competencia.

Discusión y conclusiones

Los hallazgos sugieren que, bajo un esquema híbrido (diseño–revisión–jueceo), los ítems generados con IAGen pueden comportarse de manera comparable a los ítems humanos en métricas centrales de ajuste y discriminación, mientras que la diferencia más consistente se ubica en la dificultad. Este patrón es coherente con planteamientos

recientes que sostienen que el valor de la IAGen no radica únicamente en "producir texto", sino en su integración dentro de un sistema de diseño y control de calidad (p. ej., especificaciones, revisión experta, pilotaje y monitoreo psicométrico) que permita sostener inferencias de uso (Laverghetta Jr.; Luchini; Linell; reiter-Palmon; Beaty, 2024; Küchemann; Rau; Schmidt; Kuhn, 2024).

En el contexto de EAI, estos resultados refuerzan la necesidad de documentar decisiones de gobernanza: criterios de aceptación/rechazo, trazabilidad de cambios, y delimitación del rol de la IAGen (apoyo a redacción, propuesta de distractores, revisión lingüística, etc.), particularmente por las consecuencias asociadas al uso de puntajes. En esta línea, se ha señalado que la incorporación de IAGen en evaluación requiere políticas explícitas sobre calidad, transparencia y consecuencias, de modo que la innovación tecnológica no sustituya la responsabilidad técnica del programa de evaluación, sino que la fortalezca (Weng; Xia; Gu; Rajaram; Chiu, 2024).

Así, la incorporación de la IAGen en la educación, como señalan Bozkurt, Karadeniz, Baneres, Guerrero-Roldán y Rodríguez (2021) y Dimitriadou y Lanitis (2023), representa un fenómeno creciente que sugiere una interacción simbiótica entre humanos y tecnología. Al analizar ítems creados por *ChatGPT* 4.0 en el área de Lengua Escrita, se pretende abonar a un campo que promete mucho en cuanto a la evaluación educativa como lo es la inclusión de la IAG en EAI. Como se observó en la Tabla 9 y 10, se encuentran los siguientes hallazgos puntuales:

1. Podríamos encontrar diferencias significativas entre temáticas, es decir, habrá algunos temas que puedan ser más sencillos para *ChatGPT* y otros para los humanos, pero la conjunción de ambos podría disipar estas diferencias.
2. En este estudio, los ítems de *ChatGPT* 4.0 presentan un nivel de dificultad generalmente mayor en comparación con los creados por humanos, desafiando la percepción sobre los múltiples errores que causaba *ChatGPT* (Barrot, 2023), esto aunado a la constante evolución del propio modelo de *Open AI*.
3. Ajuste al Modelo Rasch: Los ítems generados por *ChatGPT* 4.0 demuestran un ajuste más cercano al ideal, indicando una alineación precisa con las expectativas teóricas del modelo y una alta calidad de los ítems; sin embargo, se

debe precisar que pasaron por un proceso de jueceo, que aunque la edición final la hizo el propio *ChatGPT* no se descarta esa mejora en los ítems.

4. Capacidad de Discriminación: La IAG, a través de *ChatGPT* 4.0, muestra una capacidad superior para discriminar entre diferentes niveles de habilidad de los estudiantes, lo que subraya su utilidad en el diseño de evaluaciones efectivas; tomando la misma consideración que el punto 3.

El estudio de Nasution (2023) complementa estos hallazgos al demostrar el potencial de *ChatGPT* en la creación de ítems de elección múltiple confiables para fines educativos, aunque también señala la importancia del diseño de los *prompts* (Ruiz Mendoza, 2023) y la versión específica de *ChatGPT* utilizada. Los resultados generales de este estudio enfatizan el valor y relevancia del uso de *ChatGPT* 4.0 como herramienta para la elaboración de ítems de evaluación educativa, ilustrando la necesidad de un enfoque equilibrado que aproveche tanto la capacidad humana para generar ítems desafiantes como la precisión y efectividad de la IAG para ajustar y discriminar adecuadamente entre las habilidades de los estudiantes. Esto resalta la importancia de seguir explorando y optimizando el uso de *ChatGPT* y otras tecnologías de IAGen en el contexto educativo, no solo para mejorar la calidad de las evaluaciones sino también para enriquecer las prácticas pedagógicas a través de la integración efectiva de herramientas innovadoras.

Finalmente, los límites de este estudio tienen relación con la versión utilizada, puesto que se realizó con una versión específica (*ChatGPT* 4.0) y con datos de noviembre de 2023; por ello, los resultados no deben extrapolarse automáticamente a versiones posteriores del modelo. A diciembre de 2025, el ecosistema de modelos ha avanzado (p. ej., versiones posteriores), lo que sugiere la necesidad de replicación y comparación longitudinal.

Referencias

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. *Estándares para pruebas educativas y psicológicas*. Washington: American Educational Research Association, 2018.

ANDERSON, L. W.; KRATHWOHL, D. R. (ed.). *Taxonomía del aprendizaje, la enseñanza y la evaluación*: La revisión de los objetivos de la educación de Bloom. Nova York: Pearson Educación, 2001.

BARROT, J. S. Using ChatGPT for second language writing: pitfalls and potentials. *Assessing Writing*, [*S. l.*], v. 57, 2023. DOI: https://doi.org/10.1016/j.asw.2023.100745. Disponible em: https://www.sciencedirect.com/science/article/abs/pii/S1075293523000533?via%3Dihub. Acceso em: 10 jul. 2025.

BOZKURT, A.; KARADENIZ, A.; BANERES, D.; GUERRERO-ROLDÁN, A.E.; RODRÍGUEZ, M.E. Artificial intelligence and reflections from educational landscape: a review of AI studies in half a century. *Sustainability*, [*S. l.*], v. 13, n. 2, 2021. DOI: https://doi.org/10.3390/su13020800. Disponible en: https://www.mdpi.com/2071-1050/13/2/800. Acceso en: 10 jul. 2025.

CHAPELLE, C. A. *Argument-based validation in testing and assessment*. Thousand Oaks, CA: Sage, 2021. DOI: https://doi.org/10.4135/9781071878811. Disponible en: http://methods.sagepub.com/book/mono/argument-based-validation-in-testing-and-assessment/toc. Acceso en: 10 jul. 2025.

CRESWELL, J. *Research Design:* Qualitative, Quantitative, and Mixed Methods Approaches. 3. ed.. Thousand Oaks, CA: Sage, 2009.

DIMITRIADOU, E.; LANITIS, A. A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learning Environments*, [*S. l.*], v. 10, 2023. DOI: https://doi.org/10.1186/s40561-023-00231-3. Disponible en: https://link.springer.com/article/10.1186/s40561-023-00231-3. Acceso en: 10 jul. 2025.

FIELD, A. *Discovering statistics using IBM SPSS statistics*. 4. ed. Thousand Oaks, CA: Sage, 2013.

GHIO, F. B.; MORÁN, V. E.; GARRIDO, S. J.; AZPILICUETA, A. E.; CÓRTEZ, F.; CUPANI, M. Calibración de un banco de ítems mediante el modelo de Rasch para medir razonamiento numérico, verbal y espacial. *Avances en Psicología Latinoamericana*, Bogotá, v. 38, n. 1, p. 157-171, 2020. DOI: https://doi.org/10.12804/revistas.urosario.edu.co/apl/a.7760. Disponible en: https://revistas.urosario.edu.co/index.php/apl/article/view/7760. Acceso en: 10 jul. 2025.

HOSSEINI, M.; RASMUSSEN, L. M.; RESNIK, D. B. Using AI to write scholarly publications. *Accountability in Research*, [*S. l.*], v. 31, n. 7, p. 715-723, 2024. DOI: https://doi.org/10.1080/08989621.2023.2168535. Disponible en: https://www.tandfonline.com/doi/full/10.1080/08989621.2023.2168535. Acceso en: 10 jul. 2025.

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO EDUCATIVO. *Reporte técnico:* Examen de Ingreso a la Educación Superior (ExIES) 2023-1. Ensenada, B. C., Mx: UABC, 2024. [documento interno].

JORNET MELIÁ, J. M; GONZÁLEZ SUCH, J.; SUÁREZ RODRÍGUEZ, J. M. Validación de los procesos de determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios Sobre Educación*, [*S. l.*], v. 19, p. 11-29, 2010. DOI: https://doi.org/10.15581/004.19.4578. Disponible en: https://revistas.unav.edu/index.php/estudios-sobre-educacion/article/view/4578. Acceso en: 10 jul. 2025.

JURADO-NÚÑEZ, A.; FLORES-HERNÁNDEZ, F.; DELGADO-MALDONADO, L.; SOMMER-CERVANTES, H.; MARTÍNEZ-GONZÁLEZ, A.; SÁNCHEZ-MENDIOLA, M. Distractores en preguntas de opción múltiple para estudiantes de medicina: ¿cuál es su comportamiento en un examen sumativo de altas consecuencias?. *Investigación en educación médica*, Ciudad de México, v. 2, n. 8, p. 202-210, 2013. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-50572013000400005&lng=es&tlng=es. Acceso en: 22 mar. 2024.

MÉXICO. INSTITUTO NACIONAL PARA LA EVALUACIÓN DE LA EDUCACIÓN. Criterios técnicos para el desarrollo, uso y mantenimiento de instrumentos de evaluación. Diario Oficial, México, 28 abr. 2017.

KOLEN, M. J.; BRENNAN, R. L. (2014). *Test equating, scaling, and linking:* Methods and practices. 3. ed. [S. l.]: SSBS, 2014. (Statistics for Social and Behavioral Sciences). DOI: https://doi.org/10.1007/978-1-4939-0317-7. Disponible en: https://link.springer.com/book/10.1007/978-1-4939-0317-7. Acceso en: 22 mar. 2024.

KÜCHEMANN, S.; RAU, M.; SCHMIDT, A.; KUHN, J. ChatGPT's quality: reliability and validity of concept inventory items. *Frontiers in Psychology*, [*S. l.*], v. 15, 2024. DOI: https://doi.org/10.3389/fpsyg.2024.1426209. Disponible en: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1426209/full. Acceso en: 22 mar. 2025.

LANE, S.; RAYMOND, M. R.; HALADYNA, T. M. (ed.). *Handbook of test development*. 2. ed. Nova York: Routledge, 2015.

LAVERGHETTA JR., A. N.; LUCHINI, S.; LINELL, A.; REITER-PALMON, R.; BEATY, R. The creative psychometric item generator: a framework for item generation and validation using

large language models. *ArXiv*. [S. l.], 2024. DOI: https://arxiv.org/abs/2409.00202. Disponible en: https://arxiv.org/abs/2409.00202. Acceso en: 10 mar. 2025.

NASUTION, N. E. A. Using artificial intelligence to create biology multiple choice questions for higher education. *Agricultural and Environmental Education*, [*S. l.*], v. 2, n. 1, 2023. DOI: https://doi.org/10.29333/agrenvedu/13071. Disponible en: https://www.agrenvedu.com/article/using-artificial-intelligence-to-create-biology-multiple-choice-questions-for-higher-education-13071. Acceso en:

OPENAI. *Chat GPT*. 2023. Disponible en: https://chat.openai.com/chat. Acceso en: 10 mar. 2023.

RUIZ MENDOZA, K. K. El uso de ChatGPT 4.0 para la elaboración de exámenes: crear el prompt adecuado. *LATAM:* Revista Latinoamericana De Ciencias Sociales Y Humanidades, [*S. l.*], v. 4, n. 2, p. 6142–6157, 2023. DOI: https://doi.org/10.56712/latam.v4i2.1040. Disponible en: https://latam.redilat.org/index.php/lt/article/view/1040. Acceso en: 10 jun. 2024.

SHEPARD, L. La evaluación en el aula. En: BRENNAN, R. L. *Educational measurement*. 4. ed. Westport: ACE, 2006. p. 623-646.

SUSNJAK, T. ChatGPT: the end of online exam integrity? arXiv, [*S. l.*], 2022. DOI: https://doi.org/10.48550/arXiv.2212.09292. Disponible en: http://arxiv.org/abs/2212.09292. Acceso em: 10 mar. 2024.

TRISTÁN LÓPEZ, A. *Análisis de Rasch para todos:* una guía simplificada para evaluadores educativos. San Luís Potosí: Instituto de Evaluación e Ingeniería Avanzada, 1998.

WENG, X.; XIA, Q.; GU, M.; RAJARAM, K.; CHIU, T. K.. F. Assessment and learning outcomes for generative AI in higher education: a scoping review on current research status and trends. *Australasian Journal of Educational Technology*, [*S. l.*], v. 40, n. 6, p. 37-55, 2024. DOI: https://doi.org/10.14742/ajet.9540. Disponible en: https://ajet.org.au/index.php/AJET/article/view/9540. Acceso en: 22 dic. 2025.

Anexo 1

Tabla 1 - Comparativa entre humanos y *ChatGPT* 4.0

Elaboración	No.	Dificultad	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Corr. Punto-biserial	Discriminación
Humano	2	0.4200	1.0600	0.7000	1.0800	0.7000	0.1800	0.8400
Humano	13	0.4800	0.9400	-1.0000	0.9300	-0.8000	0.3500	1.2900
Humano	14	0.4700	0.8900	-1.9000	0.8600	-1.7000	0.4300	1.5000
Humano	25	0.6300	1.1700	1.4000	1.4200	2.1000	-0.0600	0.6600
Humano	12	0.4900	1.1400	2.3000	1.1500	1.7000	0.0800	0.3400
Humano	16	0.3800	0.9900	-0.1000	0.9600	-0.2000	0.2700	1.0300
Humano	20	0.3800	0.8500	-1.4000	0.7800	-1.5000	0.4600	1.2400
Humano	12	0.4900	1.1400	2.3000	1.1500	1.7000	0.0800	0.3400
Humano	16	0.3800	0.9900	-0.1000	0.9600	-0.2000	0.2700	1.0300
Promedio (Humanos)		**0.458**	**1.019**	**0.244**	**1.032**	**0.200**	**0.229**	**0.919**
ChatGPT 4.0	37	0.39	0.93	-0.8	0.85	-1.1	0.35	1.16
ChatGPT 4.0	37	0.65	1.13	0.9	1.27	1.4	0.01	0.81
ChatGPT 4.0	37	0.68	1.07	0.5	1.26	1.1	0.04	0.9
ChatGPT 4.0	37	0.56	0.96	-0.6	0.96	-0.4	0.32	1.12
ChatGPT 4.0	42	0.68	1.05	0.3	1.19	0.8	0.1	0.93
ChatGPT 4.0	42	0.51	1.05	0.9	1.06	0.9	0.18	0.75
ChatGPT 4.0	45	0.51	0.98	-0.4	0.96	-0.5	0.28	1.13
ChatGPT 4.0	46	0.49	0.89	-2.3	0.87	-2	0.42	1.64
ChatGPT 4.0	42	0.48	0.87	-2.4	0.84	-2.1	0.45	1.63
Promedio (*ChatGPT*)		**0.550**	**0.992**	**-0.433**	**1.029**	**-0.211**	**0.239**	**1.119**
Promedio general		**0.5039**	**1.0056**	**-0.0944**	**1.0306**	**-0.0056**	**0.2339**	**1.0189**

Fuente: La autora (2025).