




Creación y jueceo de ítems: ChatGPT como diseñador y juez

Criação e julgamento de itens: ChatGPT como designer e juiz
Item creation and judging: ChatGPT as designer and judge

Karla Karina Ruiz Mendoza ^{*1}, Luis Horacio Pedroza Zúñiga ^{†1} y Alma Yadhira López García ^{‡2}

¹Universidad Autónoma de Baja California, IIIDE, Ensenada, Baja California, México.

²TheLearning Bar, Canadá.

Resumen

El fin de este estudio fue evaluar la efectividad de la inteligencia artificial (IA), representada por ChatGPT 4.0, comparada con diseñadores humanos en la creación de ítems para un examen para el ingreso a la educación superior en el área de Lengua Escrita. Se utilizó un enfoque mixto, combinando metodologías clásicas y contemporáneas en evaluación educativa, incluyendo el juicio de expertos. ChatGPT y cuatro diseñadores humanos desarrollaron 84 ítems, siguiendo la Taxonomía de Anderson y Krathwohl para establecer el nivel de demanda cognitiva. Los ítems fueron evaluados por dos jueces humanos y ChatGPT, utilizando una rúbrica detallada que incluye claridad, neutralidad, formato, alineación curricular y redacción. Los resultados mostraron una alta tasa de aceptación sin cambios tanto para ítems de ChatGPT como para los humanos, indicando una buena alineación con los estándares de evaluación. Sin embargo, se observaron diferencias en la necesidad de cambios menores y mayores propuestos por la rúbrica. El estudio concluye que tanto la IA como los diseñadores humanos son capaces de generar ítems de alta calidad, resaltando el potencial de la IA en el diseño de ítems educativos.

Palabras clave: Inteligencia Artificial. Evaluación educativa. ChatGPT. Diseño de ítems. Jueceo.

Resumo

O objetivo deste estudo foi avaliar a eficácia da inteligência artificial (IA), representada pelo ChatGPT 4.0, em comparação com designers humanos na criação de itens para um exame de ingresso ao ensino superior na área de Língua Escrita. Utilizou-se uma abordagem mista, combinando metodologias clássicas e contemporâneas em avaliação educacional, incluindo o julgamento de especialistas. O ChatGPT e quatro designers humanos desenvolveram 84 itens, seguindo a Taxonomia de Anderson e Krathwohl para estabelecer o nível de demanda cognitiva. Os itens foram avaliados por dois juízes humanos e pelo ChatGPT, utilizando uma rubrica detalhada que inclui clareza, neutralidade, formato, alinhamento curricular e redação. Os resultados mostraram uma alta taxa de aceitação sem mudanças tanto para itens do ChatGPT quanto para os humanos, indicando um bom alinhamento com os padrões de avaliação. No entanto, foram observadas diferenças na necessidade de mudanças menores e maiores propostas pela rubrica. Conclui-se que tanto a IA quanto os designers humanos são capazes de gerar itens de alta qualidade, destacando o potencial da IA no design de itens educacionais.

Palavras-chave: Inteligência Artificial. Avaliação educacional. ChatGPT. Design de itens. Julgamento.

Abstract

The purpose of this study was to evaluate the effectiveness of artificial intelligence (AI), represented by ChatGPT 4.0, compared to human designers in creating items for an exam for entry into higher education in the area of Written Language. A mixed approach was utilized, combining classic and contemporary methodologies in educational evaluation including expert judgment. ChatGPT and four human designers developed 84 items, following Anderson and Krathwohl's Taxonomy to establish the level of cognitive demand. The items were evaluated by two human judges and ChatGPT, using a detailed rubric that includes clarity, neutrality, format, curricular alignment, and writing. The results showed a high rate of acceptance without changes for both ChatGPT and human items, indicating good alignment with the evaluation standards. However, differences were observed in the need for minor and major changes proposed by the rubric. The study concludes that

*Email: ruiz.karla32@uabc.edu.mx

†Email: horacio.pedroza@uabc.edu.mx

‡Email: alma.l@thelearningbar.com

Textolivre
Linguagem e Tecnologia

DOI: 10.1590/1983-3652.2024.51222

Sección:
Artículos

Autor correspondiente:
Karla Karina Ruiz Mendoza

Editor de sección:
Hugo Heredia Ponce
Editor de maquetación:
João Mesquita

Recibido el:
11 de febrero de 2024
Aceptado el:
20 de mayo de 2024
Publicado el:
31 de mayo de 2024

Esta obra está bajo una
licencia «CC BY 4.0».



both AI and human designers are capable of generating high-quality items, highlighting the potential of AI in the design of educational items.

Keywords: Artificial Intelligence. Educational assessment. ChatGPT. Item design. Judging Process.

1 Introducción

La evolución de la inteligencia artificial (IA) ha marcado un hito significativo en la actualidad sofocada por infodemia e infoxicación. Esta transformación tecnológica, inicialmente concebida por John McCarthy en 1956, ha avanzado significativamente, incorporando aplicaciones de aprendizaje automático y procesamiento del lenguaje natural (NLP) en herramientas educativas (Sadiku *et al.*, 2021). No obstante, el gran salto se dio con la presentación de ChatGPT en 2020, y con mejor respuesta a principios del 2023, siendo la IA Generativa (IAGen) otra forma de entender a la IA; la IAGen crea contenido original a partir de datos existentes mediante algoritmos y redes neuronales avanzadas (Feuerriegel *et al.*, 2024).

En el campo de la educación, los modelos de lenguaje grandes (LLM, *Large Language Model*), como ChatGPT (Open AI, 2023), ha generado diversos debates públicos y digitales como el de la opinión emitida por el lingüista Chomsky, Roberts y Watumull (2023), donde califica a ChatGPT como una forma de plagio de alta tecnología, pudiendo socavar la educación al motivar a los estudiantes en la búsqueda de atajos para la entrega de trabajos, como los ya clásicos ensayos o resolución a preguntas cerradas, como en un cuestionario de reforzamiento, por ejemplo. Ante este tipo de reflexiones, han surgido otras como la de Yell (2023), un profesor retirado de la Universidad de Wisconsin, argumentan sobre que, si se utiliza de forma adecuada, ChatGPT puede ser un recurso valioso para fomentar el aprendizaje basado en la búsqueda e investigación, permitiendo promover el pensamiento crítico. Aunque es indiscutible que este tipo de tecnología es capaz de crear contenido nuevo en formato de texto, imágenes o audio, permitiendo hasta asistir en tareas de conocimiento y necesidades cotidianas (Feuerriegel *et al.*, 2024).

La aplicación de ChatGPT en la educación se ve reflejada en el análisis de Dimitriadou y Lanitis (2023), quienes abordan la integración de la IA en las aulas inteligentes y los desafíos éticos asociados, así como en el estudio de Tlili *et al.* (2023) abordando el uso de *chatbots*, que examinan la aplicación de ChatGPT en la elaboración de ejercicios en forma de cuestionarios. Estos enfoques resaltan el equilibrio necesario entre las capacidades de la IA y la intervención humana para garantizar la relevancia, la exactitud y la equidad en la educación.

Recientes investigaciones, como las de Nasution (2023) y Ruiz Mendoza (2023), han explorado el uso de ChatGPT 4.0 en la generación de ítems de examen, destacando no solo su capacidad para crear preguntas de elección múltiple relevantes y coherentes, sino también abordando desafíos como irregularidades y redundancias en interacciones más prolongadas. Estos estudios subrayan la importancia de la especificidad y sistematización en los prompts para generar exámenes eficientes y precisos, capitalizando las fortalezas de la IA para la educación (Nasution, 2023; Ruiz Mendoza, 2023).

La investigación de Nasution (2023) se enfocó en la validez y confiabilidad de las preguntas generadas por IA, un tema que ha suscitado tanto interés como preocupación en la comunidad educativa. Con una muestra de 272 estudiantes, Nasution emprendió la tarea de evaluar una serie de preguntas creadas por ChatGPT, obteniendo resultados que son tanto prometedores como reveladores. De las 21 preguntas generadas por la IA, 20 resultaron ser válidas, lo que indica una alta tasa de éxito. Este hallazgo es significativo, ya que subraya la capacidad de la IA para producir contenido educativo que no solo es relevante, sino también de calidad.

No obstante, también hay investigaciones en torno al uso de Machine Learning (ML) como el de Rauber *et al.* (2024) quienes desarrollaron un modelo automatizado para medir el aprendizaje de conceptos y prácticas de clasificación de imágenes mediante redes neuronales. Se basó en datos de 240 estudiantes de secundaria y bachillerato, concluyendo que la evaluación es confiable y válida. Además, destacaron la efectividad del modelo resaltando la importancia de incluir ML en la educación escolar y la capacidad del modelo para asistir en el proceso de evaluación, facilitando la carga de trabajo de

los docentes.

A medida que la tecnología de IA continúa evolucionando, con avances significativos en las versiones más recientes de ChatGPT, se presenta una oportunidad única para mejorar y sistematizar el proceso de creación de exámenes. Las investigaciones de Nasution (2023) y Ruiz Mendoza (2023) se alinean con esta visión, proponiendo un enfoque metodológico que combina la exploración y descripción detallada de las capacidades de ChatGPT 4.0 en la generación de ítems de examen, proporcionando así una perspectiva integral de su aplicabilidad y eficacia en el ámbito educativo. No obstante, todavía hacen falta estudios que comparen el comportamiento de la IAGen y si los seres humanos somos capaces de detectar esas diferencias, o bien, podrían ayudar a reducir la carga de los docentes e instituciones al momento de crear exámenes de alto impacto; como los del ingreso a la universidad.

Ante todos estos acontecimientos, y retomando estos modelos de lenguaje que podrían ayudarnos a evaluar nuestra propia forma de comunicar, el objetivo principal de esta investigación fue explorar y comparar la eficacia de la IAGen, representada por ChatGPT 4.0, y los diseñadores humanos en el desarrollo de ítems para el Examen de Ingreso a la Educación Superior (ExIES), en el área de Lengua Escrita, a través del método de juicio de expertos. Lo anterior, con el fin de determinar la calidad, relevancia y alineación de los ítems generados por ambas fuentes (IA y humanos) con los estándares establecidos para la evaluación educativa, centrándose en aspectos como claridad, neutralidad, concisión, alineación curricular y adecuación de formato y contenido.

2 Validez y jueceo de expertos

Validez y jueceo de expertos

La integración de estas nuevas tecnologías nos obliga a revisar los procesos ya planteados y utilizados por los investigadores, como lo es la validez y el juicio de expertos. Según los Estándares para pruebas psicológicas y educativas de la AERA, APA y NCME (2014), la validez es un indicador clave de la calidad de cualquier instrumento de medición debido a que forma parte del fundamento al momento de interpretar y tomar decisiones, nos provee de confianza en los resultados, haciendo que forme parte de la base de las mejoras educativas y políticas, impactando en la responsabilidad y rendición de cuentas. Esto se complementa con la noción de confiabilidad, que se refiere a la consistencia de un instrumento de medición, es decir, a la coherencia de puntajes entre repeticiones de un procedimiento de evaluación, independientemente de cómo se estime o reporte esta coherencia (AERA; APA y NCME, 2014).

Es importante señalar que, incluso en la actualidad en investigaciones como la de Galicia Alarcón *et al.* (2017), se sigue utilizando el nombre de validez de contenido al juicio de expertos, no obstante, comprendiendo las ideas de Kane (2001, 2013) y Chapelle (2021), el Enfoque Basado en Argumentos (EBA), propuesto en gran medida por Kane (2001, 2013), implica que la validez no se limita al contenido del test, sino que se expande para incluir la interpretación y el uso de los resultados del test. Según este enfoque, así como la AERA, APA y NCME (2014), la validación se convierte en un proceso integral y argumentativo, donde cada interpretación de los resultados del test se justifica a través de un argumento de validez estructurado y basado en evidencia, por lo que se debe hablar de evidencias de validez y no de validez de contenido.

En este contexto, el juicio de expertos es definido como una opinión informada proporcionada por personas reconocidas como expertas cualificadas en un tema específico, capaces de ofrecer información, evidencia, juicios, y valoraciones; este tipo de método forma parte de las evidencias de validez (AERA; APA y NCME, 2014); que antes solían llamarse validez de contenido. Se lleva a cabo mediante la evaluación de ítems de un instrumento por estos expertos, quienes juzgan su claridad, coherencia, relevancia y suficiencia. La selección cuidadosa de los jueces, basada en su conocimiento y experiencia, es fundamental para obtener evaluaciones precisas y útiles, que normalmente va acompañado de una rúbrica para apegarse a ciertos criterios y evitar sesgos (Galicia Alarcón *et al.*, 2017).

Debido a lo anterior, se entiende que las preguntas generadas por IAGen deben someterse a los mismos procesos que un ser humano; este tema podría traer de nuevo al debate la profundidad del concepto de validez ahondado por Messick (1989), Kane (2001, 2013) y Chapelle (2021). La integración de la IAGen en este proceso abre nuevas posibilidades para analizar la interacción entre los

ítems y las tecnologías emergentes, mejorando así la metodología de validación. Este avance permite una evaluación más profunda de cómo se interpretan y utilizan los ítems en variados contextos, enriqueciendo la comprensión del juicio de expertos.

Además, los Estándares también ofrecen guías claras para el juicio de expertos (AERA; APA y NCME, 2014), incluyendo la documentación de las cualificaciones de los expertos, el manejo de la interacción entre participantes y la importancia de mantener juicios objetivos y bien razonados. Estas prácticas aseguran que los procesos de evaluación sean justos, confiables y capaces de soportar un escrutinio riguroso.

La introducción de la inteligencia artificial generativa, como ChatGPT, en el contexto del juicio de expertos y el EBA, propone una transformación en la metodología de validación. La IA puede jugar un papel crucial en la mejora de los procesos de evaluación, ofreciendo nuevas perspectivas y capacidades analíticas. En el enfoque basado en argumentos, como propone Chapelle (2021), la IA generativa puede contribuir significativamente a la construcción de argumentos de validez, proporcionando análisis avanzados y simulaciones que apoyen o desafíen las interpretaciones de los datos. La IA puede también identificar patrones y correlaciones que podrían pasar desapercibidos en las revisiones tradicionales.

3 Metodología

La presente investigación adoptó un enfoque mixto para la creación y evaluación de ítems para ser aplicados en exámenes que evalúan el área de Lengua Escrita, siguiendo los principios metodológicos establecidos por autores clásicos en el campo de la educación y evaluación de aprendizaje, como Bloom (1956) y Messick (1989), y se incorporaron las perspectivas sobre diseño de ítems educativos (Haladyna; Downing y Rodríguez, 2002).

En este estudio, con el fin de diseñar ítems, se asignó a cuatro diseñadores humanos y a la versión 4.0 de ChatGPT (Open AI, 2023) el desarrollo de ítems para evaluar competencias en el área de Lengua Escrita, siguiendo las teorías de evaluación educativa propuestas por Popham (1990). Se empleó la metodología de triangulación de datos (Denzin, 1978) para enriquecer la calidad y confiabilidad de los ítems mediante la combinación de múltiples perspectivas y fuentes. Cada diseñador humano fue responsable de la creación de 14 ítems, y ChatGPT generó 28 ítems, totalizando 84 ítems en el área de Lengua Escrita. La Figura 1 ilustra este proceso metodológico.

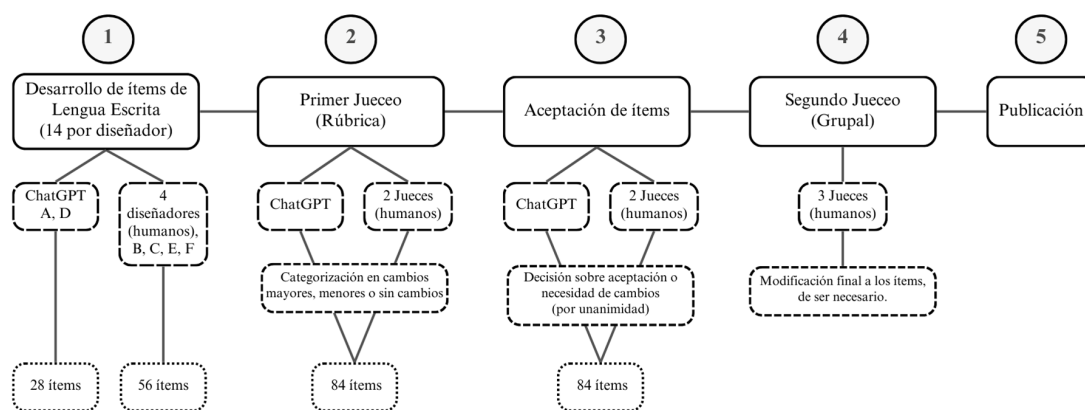


Figura 1. Proceso de evaluación y revisión de los ítems.

Fuente: Elaboración propia.

Para el desarrollo de ítems se hizo uso de la IAGen ChatGPT (chat.openai.com) en su versión de pago nombrada 4.0. Se abrieron conversaciones por cada ítem creado, ya que todavía no estaba disponible la posibilidad de crear un GPT con los criterios específicos para la elaboración de ítems. En este sentido, cada ítem fue solicitado con las características del manual de Lengua Escrita, 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 y a consultar

Por otro lado, la evaluación de los ítems fue realizada por dos jueces humanos y por ChatGPT, siguiendo el modelo de evaluación de contenido descrito por Haladyna (2004) y Lynn (1986). Los jueces evaluaron cada ítem en términos de necesidad de cambios, calidad de distractores y aceptabilidad general del ítem. Ante esto, es importante resaltar que el método utilizado para la evaluación de los ítems fue a doble ciego; a ChatGPT tampoco se le indicó si quien elaboró el ítem era un ser humano o una IAGen. Este proceso de evaluación se alinea con las recomendaciones de Nitko y Brookhart (2011) sobre la importancia de una revisión integral en la construcción de ítems de evaluación.

Los jueces evaluaron los ítems con una rúbrica (véase Tabla 1), por lo que en esta etapa no interactuaron entre sí. Al final debían señalar si el ítem debiese ser aceptado, aceptado con modificaciones o, bien, podían descartarlo. Asimismo, a ChatGPT se le solicitó evaluar cada uno de los ítems con la rúbrica disponible; es importante destacar que era necesario tener una conversación diferente por cada ítem evaluado, puesto que su capacidad de recordar la rúbrica era corta; es posible que en la actualidad haya mejorado, por lo que se tiene que seguir probando esta herramienta.

Tabla 1. Dimensiones de la rúbrica.

Dimensiones	Elementos Clave
Claridad y Pertinencia del Contenido	- Información necesaria y clara - Tema comprensible para el público objetivo
Neutralidad y Accesibilidad	- Libre de sesgos - Inclusión de imágenes/gráficas claras
Concisión y Formato	- Longitud adecuada del texto - Cumplimiento de las especificaciones de formato
Alineación Curricular	- Congruencia con especificaciones - Adecuación al nivel cognitivo y al público objetivo
Claridad Disciplinar y Enfoque	- Brevedad y claridad situacional - Presentación directa y positiva - Claridad en la redacción
Redacción y Ortografía	- Uso de vocabulario y ortografía adecuados - Ausencia de sesgos y temas delicados - Uniformidad y plausibilidad de las opciones
Estructura de Respuestas	- Ausencia de pistas indebidas - Consistencia gramatical
Formato y Presentación	- Uso correcto de elementos de formato

Fuente: Elaboración propia.

Los resultados de las evaluaciones reflejaron una gama de decisiones, desde la aceptación de ítems

sin cambios hasta la sugerencia de modificaciones en menor o mayor grado; o bien, el rechazo del ítem. Estas decisiones se basaron en criterios establecidos por expertos en evaluación educativa, como la relevancia, claridad y justicia de los ítems (Downing, 2003).

Además, algunos ítems fueron seleccionados para un segundo jueceo grupal, reflejando la metodología de revisión colaborativa sugerida por Stiggins (2001), lo cual permite una evaluación más profunda y detallada en casos donde los ítems presentan desafíos particulares o requieren ajustes más significativos. En este segundo jueceo se volvió a hacer uso de la rúbrica, pero escuchando las observaciones de los participantes, además se añadió a un tercer juez para tener una mejor variedad y perspectiva sobre el jueceo.

Una vez finalizado este segundo jueceo en versión grupal, se procedió al proceso de publicación de los ítems en su versión impresa para su aplicación en el mes de noviembre del año 2023, como parte del proceso de selección de ingreso a la universidad. En esta aplicación se contó con la participación de 2,263 sustentantes, siendo 50.06 % mujeres y 49.93 % hombres. Los resultados de aplicación del modelo Rasch, el cual es idóneo para medir actitudes, habilidades o personalidad (López, 1998), se presentan en el Apéndice A, donde todos los ítems resultaron con índices favorables conforme a los principios de la Teoría de Respuesta al Ítem (TRI).

Por otro lado, para el análisis del comportamiento de los jueces se realizaron los siguientes pasos:

- A) Para observar las diferencias entre el jueceo a ítems diseñados por humanos y ChatGPT
 1. Preparación de datos, de forma categórica.
 2. Prueba Chi-cuadrado en SPSS, siguiendo las recomendaciones de (Field, 2013).
 3. Preparación de datos, de forma numérica para proceder a un ANOVA o bien la prueba de Kruskal-Wallis (Field, 2013; Howell, 2012).
 4. Prueba de normalidad en SPSS; el resultado fue no normal, por lo que se procedió a la prueba de Kruskal-Wallis con las variables: Creadores (Humano, ChatGPT), Juez A, Juez B, ChatGPT; además se aplicó la prueba U de Mann-Whitney (Field, 2013) realizando la prueba por cada una de las variables para revisar si había alguna variación. Todos ellos con el *software* SPSS.
 5. Análisis de resultados con estadísticos básicos comparativos.
- B) Para observar la concordancia entre jueces
 1. Prueba de Kappa de Cohen (McHugh, 2012), entre jueces: A y B, A y ChatGPT, B y ChatGPT, a través del *software* SPSS.
 2. Prueba alfa de Krippendorff (Hayes y Krippendorff, 2007): A, B y ChatGPT, a través del *software* R (Versión 2023.12.0+369), con paquetería tidyverse e IRR.
 3. Análisis de resultados con estadísticos básicos comparativos.

4 Resultados

4.1 Jueceo de ítems diseñados por Humanos y ChatGPT

Como se mencionó anteriormente, en este estudio se contemplaron dos etapas. En este segmento se compara el resultado del jueceo según los ítems diseñados por humanos y ChatGPT. Por un lado, se le solicitó a ChatGPT (Open AI, 2023) que elaborara 24 ítems, de forma separada, pero siguiendo los criterios marcados, y por otro, se le solicitó a cuatro diseñadores elaborar un total de 56 ítems, 14 por cada uno de ellos. Con el fin de evaluarlos se sometió a una revisión a doble ciego con dos jueces humanos y un tercer juez que fue el propio ChatGPT, pero con los parámetros de la rúbrica.

Como se describió en la metodología, se procedió a organizar la base de datos y realizar la prueba chi-cuadrado. En la Tabla 2 se pueden observar los resultados de este primer análisis, no se observa una diferencia significativa entre cómo evaluaron los jueces los ítems desarrollados por humanos o por ChatGPT (Juez A, $p = .758$; Juez B ($.264$); ChatGPT, $p = 1.0$). No obstante, este último, ChatGPT, muestra una respuesta rotunda ($p=1$), por lo que habrá que tener cuidado sobre un comportamiento repetitivo más que crítico.

Para complementar y teniendo los mismos resultados, se realizó la prueba Kruskal-Wallis (Field, 2013; Howell, 2012), debido a que los resultados de la prueba de normalidad fueron: no normal. Los resultados de la prueba Kruskal-Wallis se observa en la Tabla 3, con los cuales se puede confirmar que no hay diferencias significativas en las evaluaciones (Juez_A, Juez_B, ChatGPT) entre los ítems

Tabla 2. Resultados chi-cuadrado entre ítems diseñados por humanos y ChatGPT según la visión de los jueces.

Creador	Chi-cuadrado de Pearson	Grados de Libertad (gl)	Significación Asintótica (Bilateral)	Notas
Juez_A	1.179	3	.758	3 casillas (37.5 %) tuvieron un recuento esperado menor que 5. El recuento mínimo esperado fue .67.
Juez_B	1.247	1	.264	2 casillas (50.0 %) tuvieron un recuento esperado menor que 5. El recuento mínimo esperado fue 2.33. Solo para una tabla 2x2.
ChatGPT	.000	2	1.000	4 casillas (66.7 %) tuvieron un recuento esperado menor que 5. El recuento mínimo esperado fue 1.00.

Fuente: Elaboración propia.

creados por humanos y los generados por ChatGPT, recordando que un valor p menor que el nivel de significancia elegido (0.05) indica que hay diferencias estadísticamente significativas entre los grupos. Por ejemplo, los resultados de Juez_A y ChatGPT, donde los valores p son 0.787 y 1.000 respectivamente, lo que sugiere que no hay diferencias significativas en las evaluaciones entre los diferentes creadores de ítems; estos resultados son similares a la realizada con chi-cuadrado Mann Withney.

Tabla 3. Resultados de la prueba Kruskal Wallis entre las evaluaciones de los jueces sobre los ítems realizados entre humanos y ChatGPT.

Variable	H de Kruskal-Wallis	Grados de Libertad (gl)	Significación Asintótica (p-valor)	Prueba de la Mediana - Chi-cuadrado	Prueba de la Mediana - Sig. Asintótica
Juez_A	0.073	1	0.787	0.386	0.534
Juez_B	1.232	1	0.267	1.247	0.264
ChatGPT	0.000	1	1.000	0.000	1.000

Fuente: Elaboración propia.

Por otro lado, se analizaron los resultados por comparación de medias, donde se puede decir, con reservas, que los resultados del jueceo, únicamente realizado por humanos (Juez A y B), revelan diferencias sutiles en la aceptación de ítems entre los creados por ChatGPT y los diseñadores humanos. En la Tabla 4, los ítems de ChatGPT mostraron una tasa de aceptación sin cambios ligeramente superior (67.85 %) en comparación con los diseñadores humanos (65.17 %). Esto sugiere que, en términos de cumplir con los criterios establecidos inicialmente, según los contenidos sobre Lengua Escrita, los ítems generados por ChatGPT (A y D) se alinearon ligeramente mejor con las expectativas de los jueces. Sin embargo, es notable que los ítems humanos (B, C, E y F) tuvieron una tasa menor de cambios menores requeridos, pero una tasa más alta de cambios mayores necesarios. Esto podría indicar que mientras los ítems de ChatGPT generalmente se acercaban más a las expectativas iniciales, los ítems humanos, cuando requerían modificaciones, necesitaban ajustes más sustanciales.

En la Tabla 5, que incluye la evaluación del Juez C (ChatGPT 4.0), se observa un aumento en la tasa de aceptación sin cambios para ambos grupos, siendo ligeramente más pronunciado para los ítems de ChatGPT (A y B). Este aumento podría reflejar una alineación en la forma de evaluar entre el ChatGPT como diseñador y como juez. Sin embargo, la tasa de aceptación sin cambios también aumentó para los ítems humanos cuando ChatGPT actuó como juez, lo que sugiere una evaluación

Tabla 4. Aceptación y comparación entre ítems realizados por humanos vs ChatGPT (Juez A y B).

Diseñadores (14 ítems por c/u)	Aceptados Sin Cambios (Promedio)	Aceptados Con Cambios Menores (Promedio)	Aceptados Con Cambios Mayores (Promedio)	Rechazados (Promedio)
ChatGPT (A y D)	67.8 5%	12.5 %	17.85 %	1.7 %
Humanos (B, C, E, F)	65.17 %	6.25 %	27.67 %	0.89 %

Fuente: Elaboración propia.

consistente y objetiva por parte de la IA. La disminución en la necesidad de cambios menores y mayores para ambos grupos implica que el Juez C (ChatGPT) tuvo una tendencia general a requerir menos modificaciones en los ítems.

Tabla 5. Aceptación y comparación entre ítems realizados por humanos vs ChatGPT (Juez A, B y C).

Grupo de Diseñadores	Aceptados Sin Cambios (Promedio)	Aceptados Con Cambios Menores (Promedio)	Aceptados Con Cambios Mayores (Promedio)	Rechazados (Promedio)
ChatGPT (A y D)	76.19 %	9.52 %	13.9 %	1.19 %
Humanos (B, C, E, F)	74.40 %	5.35 %	19.64 %	0.59 %

Fuente: Elaboración propia.

No obstante, cuando se observan los resultados por diseñador, sin jueces específicos y a partir de medias, se encuentra que no hay grandes diferencias entre los resultados de cada uno de los diseñadores. Según la Tabla 6, que refleja el promedio de decisiones tomadas por los tres jueces (Juez A, Juez B y Juez C - ChatGPT 4), el 75 % de los ítems de todos los diseñadores fueron aceptados sin cambios, lo que indica una alta calidad general y una alineación efectiva con los estándares de evaluación. Este alto porcentaje de aceptación sin cambios sugiere que la mayoría de los ítems fueron considerados adecuados y pertinentes desde su presentación inicial. Sin embargo, hay variabilidad entre los diseñadores, con el Diseñador B alcanzando la tasa más alta de aceptación sin cambios (83.33 %) y el Diseñador F la más baja (66.67 %). Esta variación puede reflejar diferencias en los enfoques de diseño de ítems o en la interpretación de los criterios de evaluación.

Tabla 6. Control de aceptación de ítems (promedio).

Diseñador	Aceptado Sin Cambios (%)	Aceptado Con Cambios Menores (%)	Aceptado Con Cambios Mayores (%)	Rechazado (%)
A	78.57 %	14.29 %	7.14 %	0.00 %
B	83.33 %	9.52 %	7.14 %	0.00 %
C	73.81 %	4.76 %	19.05 %	2.38 %
D	73.81 %	4.76 %	19.05 %	2.38 %
E	73.81 %	4.76 %	21.43 %	0.00 %
F	66.67 %	2.38 %	30.95 %	0.00 %
Media	75 %	6.75 %	17.46 %	0.79 %

Fuente: Elaboración propia.

En cuanto a los ítems que requirieron cambios menores, la media se sitúa en un 6.75 %. Este porcentaje relativamente bajo indica que solo una fracción menor de los ítems necesitaba ajustes leves para cumplir con los criterios de evaluación. Nuevamente, existe una variación entre los diseñadores, siendo el Diseñador B el que más a menudo requirió estos ajustes. Los ítems que necesitaron cambios

mayores presentaron una media del 17.46 %, lo que sugiere que, aunque en menor medida que los aceptados sin cambios, una proporción considerable de ítems necesitó modificaciones más sustanciales. En cuanto a la tasa de rechazo, esta fue bastante baja, con una media del 0.79 %, solo los Diseñadores C y D experimentaron rechazos, aunque en una proporción mínima, lo que indica que la gran mayoría de los ítems fueron considerados válidos y adecuados en cierta medida.

4.2 Consistencia y resultados entre jueces

Sin duda, a pesar de que hubo una rúbrica como guía, como se observa en la Tabla 7, hubo diferencias significativas entre las evaluaciones del Juez A, B y C. El Juez A mostró un enfoque más crítico en la evaluación, con solo un 39.29 % de ítems aceptados sin cambios. Esta tasa más baja sugiere un estándar riguroso o criterios más estrictos en la evaluación. Además, un 16.67 % de los ítems necesitó cambios menores, y una proporción significativa, 41.67 %, requirió cambios mayores. También se observó un pequeño porcentaje de rechazo (2.38 %), lo que indica que algunos ítems no cumplían con los estándares requeridos. Por otro lado, el Juez B adoptó un enfoque más permisivo o alineado con los diseños de ítems, con una alta tasa de aceptación sin cambios del 92.86 %. Esta evaluación indulgente se refleja en la ausencia total de cambios menores y solo un 7.14 % de ítems que necesitaron cambios mayores. Además, no se registraron rechazos, lo que sugiere una percepción generalmente favorable de los ítems presentados.

Tabla 7. Control de aceptación de ítems de los jueces.

Juez	Aceptado Sin Cambios	Aceptado Con Cambios Menores	Aceptado Con Cambios Mayores	Rechazado
Juez A	33 39.29 %	14 16.67 %	35 41.67 %	2 2.38 %
Juez B	78 92.86 %	0	6 7.14 %	0
Juez C (ChatGPT 4)	78 92.86 %	3 3.57 %	3 3.57 %	0
Media	75 %	6.74 %	17.46 %	0.79 %

Fuente: Elaboración propia.

No obstante, uno de los aspectos más relevantes fue observar la consistencia entre los 84 ítems y los 3 jueces, el resultado fue una concordancia baja ($\alpha = 0.228$, véase Tabla 8), que según Hayes y Krippendorff (2007), el resultado varía de 0 a 1, donde valores cercanos a 1 indican alta confiabilidad o acuerdo entre los jueces, y valores cercanos a 0 indican lo contrario. En este sentido, como se ha analizado anteriormente, parece existir una mayor concordancia entre el Juez B y ChatGPT, que entre el Juez A y cualquiera de los otros dos jueces.

Tabla 8. Resultados del alfa de Krippendorff entre los jueces.

Sujetos	3
Evaluadores	84
Alfa	0.228

Fuente: Elaboración propia.

Debido a los resultados del alfa de Krippendorff para tres jueces, se realizó la prueba Kappa de Cohen entre los jueces A y B, y entre cada juez y las evaluaciones de ChatGPT. Kappa de Cohen, según McHugh (2012), es una medida de acuerdo entre dos jueces que tiene en cuenta el acuerdo que podría ocurrir por azar. Un valor de Kappa de 1 indica un acuerdo perfecto, mientras que un valor de 0 indica que cualquier acuerdo es exactamente el que se esperaría por azar, y los valores negativos indican desacuerdo. En la Tabla 9 se pueden observar los resultados entre jueces, donde:

- Juez A vs. Juez B: Hay un mínimo acuerdo entre estos dos jueces, que no es estadísticamente significativo, es decir, sus evaluaciones no concuerdan más allá de lo que se esperaría por azar.

- Juez A vs. ChatGPT: Hay un mínimo, pero estadísticamente significativo acuerdo entre las evaluaciones del Juez A y ChatGPT. Aunque el acuerdo es mínimo, existe cierta concordancia más allá del azar.
- Juez B vs. ChatGPT: Este par muestra un moderado acuerdo que es muy significativo estadísticamente. Indica que las evaluaciones del Juez B y ChatGPT concuerdan en cierta medida más allá de lo que se esperaría por azar. Lo que también podría significar un comportamiento específico a la de un humano con menos rigurosidad que, por ejemplo, un Juez A, más crítico.

Tabla 9. Resultados de la prueba Kappa de Cohen entre jueces.

Comparación	Kappa	Significación (p-valor)	Interpretación
Juez A vs. Juez B	0.075	p = 0.092	Mínimo acuerdo, no significativo
Juez A vs. ChatGPT	0.089	p = 0.017	Mínimo acuerdo, significativo
Juez B vs. ChatGPT	0.429	p < 0.001	Moderado acuerdo, muy significativo

Fuente: Elaboración propia.

4.3 Último jueceo

En el proceso de segundo jueceo, los resultados variaron significativamente entre los distintos diseñadores, pero homogeneizó el resultado final del jueceo; considerando que aquí participaron tres jueces además del A y B y se omitió la participación de ChatGPT. En la Tabla 6 se puede ver que para el Diseñador A, 9 de sus 14 ítems (64.29 %) pasaron a esta etapa, con un ítem (7.14 %) requiriendo cambios menores y otro ítem (7.14 %) cambios mayores, mientras que la mayoría, 7 ítems (50 %), no necesitaron ningún cambio. De manera similar, el Diseñador C tuvo 8 ítems (57.14 %) en la segunda evaluación, con un ítem (7.14 %) necesitando tanto cambios menores como mayores, y 7 ítems (50 %) sin cambios. El Diseñador E también mostró un patrón parecido, con 9 ítems (64.29 %) en el segundo jueceo, donde 1 ítem (7.14 %) necesitó cambios mayores y 7 ítems (50 %) no requirieron modificaciones.

Tabla 10. Resumen del control de ítems que pasaron al segundo jueceo (grupal).

	Aceptados en primer jueceo	Total a segundo jueceo (grupal)	Aceptados Con Cambios Menores (Promedio)	Aceptados Con Cambios Mayores (Promedio)	Sin cambio (Después del jueceo)
Diseñador A	5 35.71 %	9 64.29 %	1 7.14 %	1 7.14 %	7 50 %
Diseñador B	7 50 %	7 50 %	1 7.14 %	0	7 50 %
Diseñador C	6 42.85 %	8 57.14 %	1 7.14 %	1 7.14 %	7 50 %
Diseñador D	4 28.57 %	10 71.43 %	1 7.14 %	0	8 57.14 %
Diseñador E	5 35.71 %	9 64.29 %	0	1 7.14 %	7 50 %
Diseñador F	3 21.42 %	11 78.57 %	1 7.14 %	0	10 71.43 %
Total	30	54	5	3	46

Fuente: Elaboración propia.

Por otro lado, el Diseñador B tuvo una menor proporción de ítems que pasaron a esta etapa, con solo 7 ítems (50 %), y de estos, 1 ítem (7.14 %) necesitó cambios menores y 7 ítems (50 %) fueron aceptados sin cambios. El Diseñador D mostró la mayor proporción de ítems en el segundo jueceo, con 10 ítems (71.43 %), y una alta tasa de aceptación sin cambios, ya que 8 ítems (57.14 %) no requirieron ajustes y solo 1 ítem (7.14 %) necesitó cambios menores. Sobresaliendo en este proceso, el Diseñador F presentó la mayor tasa de aceptación sin cambios, con 11 ítems (78.57 %) pasando al segundo jueceo y 10 de ellos (71.43 %) aceptados tal como estaban inicialmente.

Finalmente, como se observa en la Tabla 10, la cantidad de ítems que realmente sufrieron cambios fue un total de 8 de los 84, lo que equivale a 9.52 % del total. La segunda etapa de jueceo resultó muy necesaria por la disparidad del jueceo. En este sentido, se puede concluir que el Juez A tuvo una actitud muy crítica, lo que podría hacernos pensar en la subjetividad del ser humano. Asimismo, en la Tabla 11, que representa el ajuste final de los ítems después del jueceo grupal y antes de su publicación y aplicación en un examen, se centra en los resultados obtenidos tanto por los ítems diseñados por ChatGPT (A y D) como por los diseñadores humanos (B, C, E, F).

Tabla 11. Ajuste final, después de jueceo grupal, de los ítems.

Grupo de Diseñadores	Aceptados Sin Cambios (Promedio)	Aceptados Con Cambios Menores (Promedio)	Aceptados Con Cambios Mayores (Promedio)	Rechazados (Promedio)
ChatGPT (A y D)	85.71 %	7.14 %	3.57 %	3.57 %
Humanos (B, C, E, F)	89.28 %	5.35 %	3.57 %	1.78 %

Fuente: Elaboración propia.

Los ítems diseñados por ChatGPT mostraron una alta tasa de aceptación sin cambios, con un 85.71 % de los ítems considerados adecuados para su uso sin necesidad de modificaciones adicionales. Esto indica que una amplia mayoría de los ítems generados por ChatGPT alinearon efectivamente con los estándares de evaluación desde su concepción inicial. Además, un 7.14 % de estos ítems requirieron cambios menores, lo que sugiere que solo se necesitaron ajustes leves en una proporción relativamente pequeña de casos. En cuanto a los cambios mayores, solo un 3.57 % de los ítems necesitó este tipo de ajustes, y la misma proporción (3.57 %) fue rechazada, lo que refleja una tasa baja de rechazo y una calidad general alta.

Por otro lado, los ítems diseñados por los diseñadores humanos obtuvieron una tasa ligeramente superior de aceptación sin cambios, alcanzando un 89.28 %. Este resultado sugiere que los ítems humanos estuvieron, en promedio, un poco más alineados con los criterios de evaluación que los ítems de ChatGPT. Sin embargo, la diferencia no es muy marcada, evidenciando una calidad comparable entre ambos grupos. La tasa de ítems que requirieron cambios menores fue del 5.35 %, ligeramente inferior a la de ChatGPT, mientras que la proporción de ítems que necesitaron cambios mayores fue idéntica a la de ChatGPT, 3.57 %. La tasa de rechazo para los ítems humanos fue del 1.78 %, ligeramente inferior a la de ChatGPT, lo que indica un margen muy estrecho en términos de calidad y aceptación general.

5 Discusión y Conclusiones

El uso e investigación del uso de la IAGen se encuentra en un momento crucial de la historia educativa. Este estudio no solo se alinea con la evolución tecnológica en la educación, sino que también aborda la intersección de la IA y las metodologías de evaluación, un tema que ya ha capturado la atención de académicos y educadores por igual (Sadiku *et al.*, 2021; Hosseini; Rasmussen y Resnik, 2023). Esto a pesar de la discusión pública sobre el papel de ChatGPT en la educación, marcada por voces críticas como Chomsky, Roberts y Watumull (2023), que contrasta con perspectivas más optimistas como la de Yell (2023), quien subraya el potencial de la IAGen para enriquecer el aprendizaje.

Uno de estos ejemplos es la capacidad de ChatGPT para generar ítems de examen válidos y relevantes, como se demostró en la investigación de Nasution (2023), refleja un avance hacia la automatización en la creación de contenido educativo, respaldando los argumentos de eficacia y eficiencia en la utilización de tecnologías avanzadas en la educación (Feuerriegel *et al.*, 2024; Dimitriadou y Lanitis, 2023; Tlili *et al.*, 2023). Este estudio trata de abonar a esta visión optimista de Yell (2023) y Nasution (2023) sobre el uso e inclusión de la IAGen en procesos educativos.

Los resultados obtenidos sugieren que, bajo un juicio de expertos cuidadosamente diseñado, los ítems generados por ChatGPT alcanzan un nivel de aceptación comparable a los creados por humanos. Este hallazgo es consistente con las observaciones de Rauber *et al.* (2024), quienes también destacaron

la utilidad de la tecnología de aprendizaje automático en la evaluación educativa. Sin embargo, la variabilidad en la aceptación de ítems entre los diseñadores humanos y ChatGPT resalta la importancia de la supervisión humana y la necesidad de ajustes específicos para alinear los ítems generados por IA con los estándares educativos (Nasution, 2023; Ruiz Mendoza, 2023).

Las contribuciones de este estudio se centran en demostrar el potencial de la IAGen para asistir en la creación de contenido educativo validado, al tiempo que se subraya la necesidad de un marco de juicio de expertos robusto para evaluar la calidad de este contenido. A pesar de los resultados prometedores sobre la no diferencia de comportamiento entre jueces humanos y ChatGPT, existen limitaciones inherentes al estudio, como la dependencia de la especificidad de los *prompts* y la variabilidad en la capacidad de juicio de los expertos, lo que sugiere la necesidad de una investigación futura para optimizar los procesos de generación y evaluación de ítems con IAGen, sobre todo debido a los resultados de la confiabilidad inter-jueces, que a pesar de ser positivos para el ChatGPT, hubo discordancia entre los jueces humanos (Galicia Alarcón *et al.*, 2017).

Finalmente, este estudio no solo refuerza la viabilidad de utilizar ChatGPT y otras tecnologías de IAGen en la educación, sino que también destaca la importancia crítica del juicio humano experto en la validación de contenido generado por IA. Al vincular estrechamente los hallazgos con los objetivos planteados, este trabajo contribuye significativamente a la discusión sobre el equilibrio entre la innovación tecnológica y la necesidad de mantener altos estándares de calidad y relevancia en la educación. La investigación futura debería centrarse en perfeccionar la sinergia entre la inteligencia artificial y el juicio humano para maximizar los beneficios de ambas en el desarrollo educativo.

Algunos estudios posteriores:

- Se dará seguimiento a la evaluación de los ítems, comparando los desarrollados por humanos y por ChatGPT; se puede adelantar que los ítems se muestran consistentes, pero será publicado posteriormente (Apéndice A).
- Se dará seguimiento y se seguirán desarrollando este tipo de ítems con IAGen para observar sus limitaciones y posibles aportes importantes; recordando que hay constantes actualizaciones de ChatGPT y similares.

Referencias

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, American Psychological Association y NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. *Standards for Educational and Psychological Testing*. [S. l.]: American Educational Research Association, 2014.

ANDERSON, L.W. y KRATHWOHL, D. (ed.). *A Taxonomy for Learning, Teaching and Assessing: a Revision of Bloom's Taxonomy of Educational Objectives*. [S. l.]: Longman, 2001.

BLOOM, B. S. *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. New York: David McKay Co Inc, 1956.

CHAPELLE, C. A. *Argument-based validation in testing and assessment*. [S. l.]: SAGE Publications, 2021.

CHOMSKY, N.; ROBERTS, I. y WATUMULL, J. Noam Chomsky: The False Promise of ChatGPT. *The New York Times*, marzo 2023. Disponible en: <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>.

DENZIN, N. K. *The Research Act: A Theoretical Introduction to Sociological Methods*. [S. l.]: McGraw-Hill, 1978.

DIMITRIADOU, E. y LANITIS, A. A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learning Environments*, v. 10, n. 12, 2023. DOI: 10.1186/s40561-023-00231-3.

DOWNING, S. M. Validity: On the meaningful interpretation of assessment data. *Medical Education*, v. 37, n. 9, p. 830-837, 2003. DOI: 10.1046/j.1365-2923.2003.01594.x.

- FEUERRIEGEL, S. *et al.* Generative AI. *Bus Inf Syst Eng*, v. 66, p. 111-126, 2024. DOI: 10.1007/s12599-023-00834-7.
- FIELD, A. *Discovering statistics using IBM SPSS statistics*. 4th. [S. l.]: Sage, 2013.
- GALICIA ALARCÓN, Liliana Aidé *et al.* Validez de contenido por juicio de expertos: propuesta de una herramienta virtual. *Apertura*, v. 9, n. 2, p. 42-53, 2017. DOI: 10.32870/Ap.v9n2.993.
- HALADYNA, T. M. *Developing and Validating Multiple-choice Test Items*. [S. l.]: Lawrence Erlbaum Associates, 2004.
- HALADYNA, T. M.; DOWNING, S. M. y RODRÍGUEZ, M. C. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, v. 15, n. 3, p. 309-333, 2002. DOI: 10.1207/S15324818AME1503_5.
- HAYES, A. F. y KRIPPENDORFF, K. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, v. 1, n. 1, p. 77-89, 2007.
- HOSSEINI, M.; RASMUSSEN, L. M. y RESNIK, D. B. Using AI to write scholarly publications. *Accountability in Research*, p. 1-9, 2023. DOI: 10.1080/08989621.2023.2168535.
- HOWELL, D. C. *Statistical methods for psychology*. Wadsworth, NY: Cengage Learning, 2012.
- KANE, M. T. Current Concerns in Validity Theory. *Journal of Educational Measurement*, v. 38, n. 4, p. 319-342, 2001. DOI: 10.1111/j.1745-3984.2001.tb01130.x.
- KANE, M. T. Validating the interpretations and Uses of Test Scores. *Journal of Educational Measurement*, v. 50, n. 1, p. 1-73, 2013. DOI: 10.1111/jedem.12000.
- LÓPEZ, A. T. *Análisis de Rasch para todos. Una guía Simplificada para evaluadores educativos*. [S. l.]: Instituto de Evaluación e Ingeniería Avanzada, 1998. ISBN 9709225103.
- LYNN, M. R. Determination and Quantification of Content Validity. *Nursing Research*, v. 35, n. 6, p. 382-385, 1986.
- MCHUGH, M. L. Interrater reliability: the kappa statistic. *Biochemia Medica*, v. 22, n. 3, p. 276-282, 2012.
- MESSICK, S. Validity. *In: Educational Measurement*. Edición: R. L. Linn. 3rd. [S. l.]: American Council on Education/Macmillan, 1989. p. 13-103.
- NASUTION, N. E. A. Using artificial intelligence to create biology multiple choice questions for higher education. *Agricultural and Environmental Education*, v. 2, n. 1, em002, 2023. DOI: 10.29333/agrenvedu/13071.
- NITKO, A. J. y BROOKHART, S. M. *Educational Assessment of Students*. Boston, MA: Pearson, 2011.
- OPEN AI. *ChatGPT (versión del 14 de marzo) [Modelo de Lenguaje Grande]*. 2023.
- POPHAM, W. J. *Educational Evaluation*. Boston, MA: Allyn y Bacon, 1990.
- RAUBER, M. F. *et al.* Reliability and validity of an automated model for assessing the learning of machine learning in middle and high school: Experiences from the "ML for All!" course. *Informatics in Education*, v. 00, n. 00, 2024. DOI: 10.15388/infedu.2024.10.
- RUIZ MENDOZA, K. K. El uso del ChatGPT 4.0 para la elaboración de exámenes: crear el prompt adecuado. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, v. 4, n. 2, p. 6142-6157, 2023. DOI: 10.56712/latam.v4i2.1040.
- SADIKU, M. N. O. *et al.* Artificial Intelligence in Education. *International Journal of Scientific Advances*, v. 2, n. 1, 2021.

STIGGINS, R. J. *Student-involved classroom assessment*. [S. l.]: Prentice Hall, 2001.

TLILI, A. et al. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, v. 10, n. 15, 2023. DOI: 10.1186/s40561-023-00237-x.

YELL, M. M. Social studies, ChatGPT, and lateral reading. *Social Education*, v. 87, n. 3, p. 138-141, 2023.

Contribuciones de los autores

Karla Karina Ruiz Mendoza: Investigación, Análisis formal, Redacción – borrador original,; **Luis Horacio Pedroza Zúñiga:** Investigación, Supervisión, Validación, Redacción – Revisión y edición; **Alma Yadhira López García:** Metodología, Supervisión, Validación.

A Resultados de los ítems elaborados por ChatGPT

El estudio de las métricas de cada aplicación se basó en la aplicación del modelo Rasch; que es parte de la Teoría de Respuesta al ítem. Este modelo probabilístico no determinista predice la posibilidad de que una persona seleccione la respuesta adecuada de un ítem, dependiendo de la discrepancia entre el estímulo aplicado y el nivel del atributo del individuo (López, 1998).

Tabla 12. 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

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: Elaboración propia.

B Datos de las evaluaciones por creador y Juez

Tabla 13. Datos de las evaluaciones por creador y Juez

Creador	Evaluación Juez A	Evaluación Juez B	Evaluación ChatGPT
ChatGPT	1	0	0
ChatGPT	2	2	2
ChatGPT	1	0	0
ChatGPT	1	0	0
ChatGPT	1	0	0
ChatGPT	1	0	0
ChatGPT	0	0	0
ChatGPT	1	0	0
ChatGPT	2	0	0
ChatGPT	0	0	0
ChatGPT	0	0	0
ChatGPT	0	0	0
ChatGPT	0	0	0
ChatGPT	0	0	0

ChatGPT	0	0	0
ChatGPT	2	0	0
ChatGPT	0	0	0
ChatGPT	2	0	0
ChatGPT	0	0	0
ChatGPT	2	0	0
ChatGPT	3	0	1
ChatGPT	0	0	0
ChatGPT	0	0	0
ChatGPT	2	0	0
ChatGPT	2	0	0
ChatGPT	2	0	0
ChatGPT	2	0	0
ChatGPT	2	0	0
Humano	1	0	0
Humano	0	0	0
Humano	1	0	0
Humano	1	0	0
Humano	0	0	0
Humano	1	0	0
Humano	0	0	0
Humano	2	0	0
Humano	0	0	0
Humano	0	0	0
Humano	0	0	0
Humano	2	0	0
Humano	0	0	0
Humano	0	0	0
Humano	1	2	2
Humano	0	0	0
Humano	0	0	0
Humano	0	0	0
Humano	2	2	0
Humano	2	2	0
Humano	0	2	0
Humano	0	0	0
Humano	2	0	0
Humano	2	0	0
Humano	0	0	0
Humano	3	0	0
Humano	1	0	1
Humano	2	0	0
Humano	0	0	0
Humano	0	0	0
Humano	2	0	0
Humano	1	0	0
Humano	1	0	1
Humano	0	0	0
Humano	0	0	0
Humano	2	0	0
Humano	2	0	0
Humano	2	0	0
Humano	2	0	0
Humano	2	0	0

Humano	2	0	0
Humano	2	0	0
Humano	2	0	0
Humano	0	0	0
Humano	2	2	2
Humano	2	2	0
Humano	0	0	0
Humano	2	0	0
Humano	2	0	0
Humano	2	0	0
Humano	0	0	0
Humano	2	0	0
Humano	2	0	0
Humano	2	0	0
Humano	2	0	0

Fuente: Elaboración propia.

Sin cambio = 0; Con cambios menores = 1; Con cambios mayores = 2; Rechazado = 3.