



La evaluación de la escritura en español en la educación superior: una conversación con ChatGPT¹

Sergio Álvarez Uribe 

Universidad del Norte
Barranquilla, Colombia

Resumen

En un contexto de uso masivo de ChatGPT, este artículo explora la evaluación de la escritura académica en la educación superior a través de dos preguntas de investigación: ¿cómo evaluar la escritura en la universidad en la era de la inteligencia artificial (IA) y qué tan confiable es usar la IA para evaluar la escritura académica? El estudio aplica un diseño secuencial exploratorio de dos fases. La primera fase consiste en una entrevista a ChatGPT sobre sus capacidades de escritura. La segunda fase analiza la confiabilidad de ChatGPT para evaluar una muestra de 40 respuestas construidas, evaluadas por 5 evaluadores, de una prueba de español académico respaldada por una argumentación de validez. Los resultados apuntan a dos conclusiones provisionales: la evaluación de la escritura académica necesita incrementar prácticas específicas que minimicen los riesgos del uso de la IA; y las evaluaciones que hace la IA necesitan una verificación constante de su confiabilidad debido al sesgo de tendencia central.

Palabras clave: ChatGPT; evaluación de la escritura; confiabilidad; autoría; IA.

1/42

¹ Artículo de investigación
Lenguaje 53(1), e20414433

<https://doi.org/10.25100/lenguaje.v53i1S.14433>

Abstract

The assessment of writing in Spanish in higher education: a conversation with ChatGPT

In a context of massive use of ChatGPT, this article explores the assessment of academic writing in higher education through two research questions: how to assess writing at university in the age of artificial intelligence (AI) and how reliable is it to use AI to assess academic writing? The study applies a two-phase exploratory sequential design. The first phase consists of an interview with ChatGPT about its writing abilities. The second phase analyzes the reliability of ChatGPT to assess a sample of 40 constructed responses, evaluated by 5 raters, of an academic Spanish test supported by a validity argument. The results point to two tentative conclusions: the assessment of academic writing needs to increase specific practices that minimize the risks of using AI; and the assessments made by AI need constant verification of their reliability due to the central tendency bias.

Key words: ChatGPT; writing assessment; reliability; authorship; AI.

Résumé

L'évaluation de l'écriture en espagnol dans l'enseignement supérieur : une conversation avec ChatGPT

2/42 Dans un contexte d'utilisation massive du ChatGPT, cet article explore l'évaluation de l'écriture académique dans l'enseignement supérieur à travers deux questions de recherche : comment évaluer les écrits universitaires à l'ère de l'intelligence artificielle (IA) et quelle est la fiabilité de l'utilisation de l'IA pour évaluer l'écriture académique ? L'étude applique un modèle exploratoire séquentiel en deux phases. La première phase consiste en un entretien avec des ChatGPT au sujet de leurs compétences en matière d'écriture. La seconde phase analyse la fiabilité du ChatGPT pour évaluer un échantillon de 40 réponses construites, évaluées par 5 évaluateurs, d'un test académique d'espagnol soutenu par une argumentation de validité. Les résultats permettent de tirer deux conclusions provisoires : l'évaluation de l'écriture académique doit accroître les pratiques spécifiques qui minimisent les risques liés à l'utilisation de l'IA ; et les évaluations effectuées par l'IA doivent faire l'objet d'une vérification constante de leur fiabilité en raison d'un biais de tendance centrale.

Mots clés : ChatGPT ; évaluation de l'écriture ; fiabilité ; paternité ; IA.

SOBRE EL AUTOR

Sergio Álvarez Uribe

Profesor asistente del Departamento de Español del Instituto de Idiomas de la Universidad del Norte. Doctor en Lingüística aplicada a la enseñanza de lenguas de la Universidad Antonio de Nebrija (calificación sobresaliente *cum laude*). Docente del área de formación básica en eficacia comunicativa. Director del Centro de escritura Eficacia Comunicativa - ECO. Investigador en el área de evaluación de la escritura académica.

Correo electrónico: sergioa@uninorte.edu.co

3/42

CÓMO CITAR ESTE ARTÍCULO

Álvarez Uribe, S. (2025). La evaluación de la escritura en español en la educación superior: una conversación con ChatGPT. *Lenguaje*, 53(1S), e20414433.
<https://doi.org/10.25100/lenguaje.v53i1S.14433>

INTRODUCCIÓN

El número de artículos sobre inteligencia artificial en educación (AIED por su sigla en inglés) ha aumentado de forma constante desde 2001 (Paek y Kim, 2021). El mayor pico de publicaciones sucede en 2023 debido al lanzamiento de ChatGPT en noviembre de 2022 como la primera aplicación de chatbot de inteligencia artificial (IA²) de acceso masivo (Nguyen y Truong, 2025). Este aumento de publicaciones coincide con la identificación de ChatGPT como uno de los tres temas predominantes registrados por Qin y Zhang (2025).

Nguyen y Truong (2025) resumen el estado actual de la investigación de IA en educación con el foco en el aprendizaje personalizado y la evaluación automática que coincide con el tema de evaluación inteligente de la revisión sistemática de Wang *et al.* (2024). Este foco también es consistente con la evolución de la evaluación de la escritura mediante IA hacia la construcción de sistemas sofisticados que implican desafíos éticos, técnicos y de implementación (Astutik *et al.*, 2024).

El análisis del impacto de la IA en la educación superior revela un panorama complejo de transformación que necesita una visión crítica (Ruano-Borbalan, 2025) y demanda exploraciones innovadoras para ser comprendido. Steele (2023) sintetiza la necesidad de investigar el uso masivo de ChatGPT desde el abordaje de tres tipos de problemas en educación: de medición, de exactitud de la información y de devaluación de habilidades. En este artículo, estos problemas se enmarcan en los componentes teórico-metodológicos que se han consolidado en la investigación de validación de la teoría de la evaluación de la lingüística aplicada (Álvarez Uribe, 2023; Chapelle, 2020; Fulcher, 2015): el dominio (qué y cómo se evalúa), la confiabilidad (cómo se mide el dominio), la interpretación (qué significan los resultados de evaluación) y el uso (para qué se evalúa y cuáles son las consecuencias).

A la luz de los componentes de dominio y confiabilidad de la teoría de la evaluación, se plantean dos preguntas de investigación: ¿cómo evaluar la escritura en la universidad en la era de la IA?, y ¿qué tan confiable es usar la IA para evaluar la escritura académica?

² Se usa la abreviatura simplificada de IA para referirse en términos generales a aplicaciones de la inteligencia artificial generativa (IAG) como el chatbot ChatGPT. Estas aplicaciones se basan principalmente en modelos de lenguaje de gran tamaño (LLM por su sigla en inglés) que usan el procesamiento del lenguaje natural para producir respuestas a tareas basadas en lenguaje. En la literatura no especializada, a veces se mezclan los conceptos de IA, IAG y LLM, aunque no son lo mismo. En algunos casos, también se incluyen otros modelos multimodales de aprendizaje automático para el procesamiento de imágenes, video y audio como GPT-4o en la misma denominación general de IA.

REVISIÓN DE LITERATURA

Antes de ChatGPT, la investigación sobre la calificación automatizada de ensayos (AES por su sigla en inglés) y la evaluación automática de la escritura (AWE por su sigla en inglés) reporta varios desafíos. Zupanc y Bosnic (2015) identifican la necesidad de evaluar el contenido semántico y asegurar la validez y confiabilidad de las calificaciones. Hussein *et al.* (2019) señalan el potencial de engaño en la calificación y las limitaciones para evaluar la creatividad. Ramesh y Sanampudi (2022) se enfocan en la dificultad de evaluar la relevancia del contenido y la coherencia. Por su lado, Zhang (2021) da cuenta de la complejidad de dos sistemas de evaluación: IntelliMetric (Rudner *et al.*, 2006) y sus aplicaciones (Elliot, 2013); y el sistema e-rater (Burstein *et al.*, 2004) del Educational Testing Service, conocido en Latinoamérica por el examen TOEFL. Zhang (2021) resalta la capacidad de estos sistemas de identificar rasgos lingüísticos al nivel del léxico, la gramática y la estructura discursiva, y el beneficio potencial en el ahorro de tiempo en la evaluación y en la retroalimentación; pero también indica los retos de infraestructura, financiamiento y mantenimiento. La revisión sistemática de Shi y Aryadoust (2023) señala la falta de correspondencia entre lo que se puede evaluar por un sistema de evaluación automática y las tareas de escritura en un entorno real de uso del lenguaje con propósitos específicos.

5/42

En español, Lillo-Fuentes *et al.* (2023) identifican el bajo desarrollo de la evaluación automatizada de la calidad de la escritura, que coincide con la ausencia de estudios actuales en contextos de uso de IA para la evaluación de la escritura en español. Se encuentran pocos ejemplos respaldados por investigación: la herramienta formativa de ArtText (da Cunha *et al.*, 2017), el programa Gallito-API (Castien *et al.*, 2021), la aplicación Writing Mentor en español (Cahill *et al.*, 2021), la herramienta Estilector (Nazar y Renau, 2023) y el uso especializado de ChatGPT-4 para la revisión de textos académicos en español (Barrezueta-Guzman *et al.*, 2023). La revisión reciente de Martínez-Olmo y González Catalán (2024) incluye algunas investigaciones sobre los retos éticos, de calidad y de autonomía intelectual en la escritura académica en español. En un acercamiento más general, Salas-Pico y Yang (2022) revisan las aplicaciones de IA en la educación superior en Latinoamérica y el aumento de adopción de IA en Brasil, Colombia, México, Ecuador, Perú, Chile y Argentina, y aunque no profundiza en cómo las instituciones pueden abordar las preocupaciones éticas relacionadas con el uso de IA, sí identifica la utilidad de los modelos predictivos para el aprendizaje como una de las áreas de mayor impacto. Adicionalmente, la revisión de Pérez *et al.* (2025) incluye referencias en español, pero falta un análisis más riguroso y sistemático de cómo las herramientas basadas en IA impactan realmente en la calidad de la escritura académica por la naturaleza exploratoria de la mayoría de los estudios revisados.

Las exploraciones del impacto de la IA en la escritura académica en inglés señalan las oportunidades de integración tecnológica para la mejora de la retroalimentación y la corrección en la escritura con efectos positivos en la competencia escrita (Dong, 2023; López-Regalado *et al.*, 2024; Wise *et al.*, 2024). Parker *et al.* (2023) presentan los resultados de evaluación con comentarios de ChatGPT a una muestra tomada del corpus MICUSP (Römer y O'Donnell, 2011). Su conclusión es que ChatGPT supera los sistemas basados en reglas al ofrecer comentarios relacionados con aspectos de la complejidad, la precisión y la fluidez (CAF por su sigla en inglés) a través de criterios de evaluación como la escritura de tesis de textos argumentativos, el propósito general del texto y la adecuación de las evidencias utilizadas como soporte (Parker *et al.*, 2023). Bucol y Sangkawong (2024) comparan las evaluaciones de 10 ensayos de estudiantes de inglés como lengua extranjera (LE) hechas en 5 ocasiones distintas por ChatGPT contra las evaluaciones de 5 docentes. El estudio muestra una correlación moderada de .652 entre la IA y los evaluadores humanos.

Otros estudios hacen énfasis en las implicaciones del uso de IA generativa de los chatbots relacionadas con la creatividad, el pensamiento crítico, la investigación y las prácticas éticas de escritura (Aljuaid, 2024; Imran y Almusharraf, 2023; Malik *et al.*, 2023). El estudio de Herbold *et al.* (2023) analiza 658 calificaciones de 117 evaluadores de 270 ensayos escritos por humanos, ChatGPT-3 y ChatGPT-4 y concluyen que ChatGPT “supera significativamente la calidad de textos producidos por humanos” (Herbold *et al.*, 2023, p. 8).

Un tema recurrente es la confiabilidad de las tareas que pueden ser generadas por IA (Zaheer *et al.*, 2024) y la confiabilidad de las evaluaciones de exámenes escritos que pueden presentar problemas de baja explicabilidad y sesgos que comprometen la validez, la confianza de las personas evaluadas y la de la opinión pública en general (Aloisi, 2023; Chen *et al.*, 2024; Owan *et al.*, 2023). El desafío para Tarp y Nomdedew-Rull (2024) está en garantizar el uso de las herramientas de IA para producir trabajos de mayor calidad de manera más eficiente. Zirar (2023) hace una síntesis de 25 artículos sobre el uso de ChatGPT en la educación y concluye que su integración por parte estudiantes y docentes en el proceso de aprendizaje, sea para generar o evaluar textos escritos, solo es útil si se verifica la precisión, la validez y la confiabilidad de cualquier contenido generado por IA.

Mizumoto y Eguchi (2023) reportan un coeficiente de Kappa de .388 entre las evaluaciones de GPT y las de los 3 niveles de referencia del corpus TOEFL11 hechas por profesionales en evaluación; y un valor de Kappa de .605 cuando se le suman 45 rasgos lingüísticos a la evaluación, provenientes de la herramienta para el análisis automático de la diversidad léxica (TAALED, por su sigla en inglés), la de sofisticación léxica (TAALES), la de complejidad y sofisticación sintáctica (TAASSC), y la de cohesión (TAACO). Esta diferencia les permite afirmar que “la combinación

de GPT + rasgos lingüísticos basados en investigación puede producir el mejor resultado en la predicción de evaluaciones hechas por profesionales en evaluación” (Mizumoto y Eguchi, 2023, p. 9). Su estudio concluye que “la calificación automática de ensayos con GPT solo puede ser una herramienta de apoyo y no puede reemplazar a los evaluadores humanos ni al profesor en el aula” (p. 10). Además, señalan la limitación de los modelos de aprendizaje profundo como GPT por funcionar como una caja negra y recomiendan el desarrollo del concepto de IA explicable de Kumar y Boulanger (2020). Awidi (2024) analiza una muestra de 108 textos de reflexión de estudiantes de ingeniería de primer año evaluados por 39 tutores expertos que se comparan con dos evaluaciones consecutivas de ChatGPT. El estudio encuentra que “no hay una relación estadísticamente significativa entre las evaluaciones de los tutores expertos y las de ChatGPT” (Awidi, 2024, p. 7).

La investigación de Pfau *et al.* (2023) evidencia problemas de consistencia de ChatGPT para identificar errores de bajo nivel relacionados con la ortografía que se categorizan de distintas formas en dos ocasiones de evaluación. La correlación entre las dos ocasiones entre humanos fue de 0.98 y la de ChatGPT fue de 0.80. El estudio evidencia que los humanos identifican errores que ChatGPT no, como oraciones largas (*run-on sentences* en inglés), uso de preposiciones y uso de artículos. Además, se identifica una tendencia de ChatGPT a ubicar errores en categorías muy amplias que no permiten una comprensión precisa de lo que se considera un error de escritura. Por estos hallazgos, los autores recomiendan la verificación humana y el reporte detallado de procedimientos para evaluar la precisión de cualquier IA (Pfau *et al.*, 2023). Lundgren (2024) identifica un patrón conservador de calificación de GPT-4; calcula un valor de Kappa entre la IA y los evaluadores humanos de 0.18 que evidencia una baja confiabilidad y una tendencia de la IA a las características generales en vez de adaptarse a los matices de los criterios solicitados en la instrucción de evaluación.

Por su parte, el experimento de Naismith *et al.* (2023), con una muestra estratificada de 500 respuestas de la prueba estandarizada de inglés de Duolingo (DET por su sigla en inglés), compara una línea base de evaluación construida con indicadores de coherencia (e. g. repetición de nombres entre pares de oraciones) basados en Cohmetrix (Graesser *et al.* 2004) y una evaluación humana basada en una rúbrica holística de coherencia de 6 puntos. La evaluación humana se compara con el modelo GPT-4 en tres condiciones: a) una evaluación sola sin justificación, b) justificación primero más evaluación y c) evaluación primero más justificación. El valor de Kappa de mayor acuerdo es de 0.43 y el de correlación de Spearman de 0.82 en la condición de una evaluación primero más justificación después. Todas las evaluaciones del modelo GPT-4 superan el acuerdo con la evaluación humana en relación con los indicadores lingüísticos de coherencia. Naismith *et al.* (2023) concluyen que las evaluaciones con justificación del modelo GPT-4 prometen un

camino hacia la explicabilidad de los resultados cuantitativos en términos de constructos lingüísticos relevantes en oposición a la naturaleza de caja negra del aprendizaje profundo. Sin embargo, advierten sobre la posibilidad de sesgos difíciles de identificar, porque es un hecho que las justificaciones ofrecidas por GPT-4 “no revelan ningún proceso de pensamiento” sobre las evaluaciones de una prueba dada (Nasmith *et al.*, 2023, p. 399).

Entre las visiones más críticas de la IA, Chomsky (2023) hace un llamado a nombrar las inteligencias artificiales generativas como “lo que son: software de plagio” (p. 32). Por su parte, Lanier (2023) afirma que “la manera más fácil de administrar mal una tecnología es malinterpretarla” (párr. 5). Su llamado va en la dirección de que “la meta debería ser, al menos, establecer una nueva clase creativa en vez de una nueva clase dependiente” (Lanier, 2023, párr. 26). En otra publicación, Lanier (2024) continúa su reflexión sobre la tecnología de la IA: “para que sea útil necesita estar acompañada de otros elementos, como la comprensión de la sociedad en general, buenos hábitos y la aceptación de una responsabilidad compartida de sus consecuencias” (párr. 1); y reafirma que “si no podemos entender cómo funciona una tecnología, nos arriesgamos a sucumbir al pensamiento mágico” (Lanier, 2024, párr. 1).

8/42 Las posturas de Chomsky (2023) y Lanier (2023, 2024) visibilizan, desde el periodismo científico, preocupaciones reales en los inicios del uso masivo de ChatGPT en EEUU. Preocupaciones similares en torno al uso inadecuado de ChatGPT se encontraron en el contexto universitario. Por ejemplo, Townsend (2023) encuestó a 1000 estudiantes universitarios y más del 40 % reportaron el uso de ChatGPT para trabajos de clase. Allan Cu y Hochman (2023) reportaron los resultados de una encuesta anónima en Stanford respondida por 4497 estudiantes en la aplicación Fizz, que requiere el uso de una cuenta oficial de la universidad, en la que el 17 % afirmó haber usado ChatGPT para tareas y exámenes finales, en muchos casos, sin ningún trabajo de edición del texto entregado como respuesta.

De igual manera, la encuesta de Barret y Pack (2023) identificó la ausencia de preparación de las instituciones y los salones de clase para la integración de la IA en el aula. Las preguntas de la encuesta se enfocaron en los procesos de planeación, escritura y revisión. En el apartado de escritura específicamente, llama la atención que el 10.8 % de estudiantes y el 5.9 % de docentes están de acuerdo con que se entreguen textos generados por IA sin declarar su uso. En contraste, otra encuesta reciente (Lee *et al.*, 2024) muestra que a nivel nacional en EEUU los estudiantes no reportan más plagio con ChatGPT que antes del uso masivo de aplicaciones de IA.

Hicks *et al.* (2024) caracterizan las denominadas “alucinaciones” de las IA como el comportamiento indiferente de los LLM respecto a la verdad de sus respuestas. Desde la perspectiva de un estudio empírico, Durmus *et al.* (2024) evalúan la capacidad de persuasión de distintas versiones del modelo de la IA

Claude a través de un diseño experimental en el que 3832 participantes deben valorar qué tanto los persuade una tesis antes y después de ser expuestos a un argumento. Los argumentos varían entre escritos por humanos y escritos por distintos perfiles de IA. El perfil denominado de “Estrategia engañosa”, que le permite al modelo de IA fabricar información, recibe la calificación más alta de persuasión. Los autores resaltan la conexión potencial entre las capacidades persuasivas de los modelos de lenguaje y la difusión de desinformación (Durmus *et al.*, 2024). En el contexto del área de física, Yeadon *et al.* (2024) encontraron que 5 evaluadores no diferenciaron si los ensayos cortos que les presentaron en un experimento eran escritos por humanos o por IA.

En la revisión de revisiones de Bond *et al.* (2024) se encuentra que la evaluación es el uso más común de la IA en educación (AIEd, por su sigla en inglés). Pero se identifican “pocos estudios que den detalles explícitos sobre el impacto del uso de IA en el aprendizaje” (Bond *et al.*, 2024, p. 24). El problema más recurrente, identificado en el 40.9 % de los estudios, es la necesidad de avanzar consideraciones éticas. La revisión de Bond *et al.* (2024) concluye con el llamado a “enfocarse en tres áreas particulares: ética, colaboración y rigor” (p. 34). Por su parte, la revisión sistemática de Lo (2023) sobre el uso de ChatGPT en la educación identifica que “los estudiantes que usan ChatGPT son más propensos a cometer plagio que los que no lo usan” (p. 10) en oposición a la encuesta de Lee *et al.* (2024). Lo (2023) relaciona este aspecto del plagio que facilita ChatGPT con la posibilidad de anular el propósito de “evaluar el aprendizaje de manera imparcial” (p. 10). En esta misma línea, İpek *et al.* (2023) recalcan que la influencia de ChatGPT en la escritura académica aún no se entiende. Su revisión concluye que el uso de IA para plagio “puede causar que los estudiantes eviten sus responsabilidades académicas; y opten, en su lugar, por transferir este deber a un programa de inteligencia artificial, en vez de cumplir con sus responsabilidades de aprendizaje, lo que se traduce en un funcionamiento de la IA más allá de su propósito previsto” (İpek *et al.*, 2023, p. 45).

Algunas investigaciones prueban las capacidades de ChatGPT como sistema de tutoría para examinar el supuesto de la autonomía que favorece en los estudiantes. Castro Nascimiento y Silva Pimentel (2023) presentan una conversación con ChatGPT sobre conocimiento de química y concluyen que es cuestionable por la variabilidad del rango de precisión entre 25 % y 100 % en las respuestas, que pueden inducir a error en el proceso de aprendizaje. En contraste, Levine *et al.* (2024) presentan un uso ideal de la IA en la escritura de 12 estudiantes de secundaria, que representa una manera más productiva y responsable de abordar la integración de nuevas tecnologías en la escritura. En la misma línea, Su *et al.* (2023) presentan un ejemplo de uso de la IA para la tutoría personalizada en escritura. En un punto medio, Eysenbach (2023), en la segunda parte de una conversación con ChatGPT,

evalúa el potencial de la IA para la educación en medicina como prometedora, desde que se supervisen los problemas de precisión y confiabilidad.

Respecto al acceso a tecnologías basadas en IA, Moorhouse (2024) identifica la rápida y masiva adopción de la IA por estudiantes en comparación con la preparación docente necesaria para afrontar los retos de formación especializada de su uso educativo. A propósito de la preparación para afrontar los retos de la IA, la revisión sistemática de Luo (2024) analiza el contenido de las políticas de 20 universidades de reconocimiento mundial en relación con el uso de IA y la evaluación en la educación superior. Se identifica la posibilidad real de que los estudiantes no entreguen trabajos originales para su evaluación como el problema principal. Luo (2024) señala que el problema debe ser enmarcado más allá de esa posibilidad real para empezar a entender qué significa la originalidad en una era en la que la producción de conocimiento no parece posible sin una mediación tecnológica.

Se puede evidenciar que existe una preocupación recurrente sobre consideraciones éticas del uso de IA y que las comparaciones de evaluación de IA se hacen principalmente con humanos como el estándar de referencia. Además, aunque en las investigaciones mencionadas no se identifica un acuerdo generalizado sobre las implicaciones del uso de IA en la escritura, sí se puede afirmar que desde el lanzamiento de ChatGPT, el proceso complejo de resolver una tarea de escritura y la evaluación de respuestas construidas de pruebas escritas se encuentra al alcance de una instrucción. Para enfrentar esta realidad, no hay todavía una guía clara ni una comprensión compartida. La recomendación de Rudolph *et al.* (2023) de alejarse de un comportamiento policivo puede ser una manera realista de incentivar reflexiones y exploraciones profundas y creativas sobre el uso de IA. De todos modos, tal como lo evidencian Dwivedi *et al.* (2023), la integridad académica siempre ha sido un problema en la educación superior, y el uso de IA lo hace más evidente.

CONTEXTO DE LA INVESTIGACIÓN

Al finalizar el primer semestre académico en junio de 2024, se administró y se evaluó la prueba diagnóstica del español académico escrito esperado en la educación superior, abreviada EA2ES, documentada en Álvarez Uribe (2023), en una universidad privada del Caribe colombiano. La administración contó con 648 estudiantes de primer año de universidad que culminaban su segundo curso de escritura académica. Se descartaron 21.7 % de las pruebas por la sospecha reportada por los evaluadores de uso de IA en la escritura de la respuesta construida³. Los

³ La sospecha de IA corresponde a la valoración independiente de cada evaluador de rasgos no establecidos oficialmente, que fueron discutidos en el entrenamiento previo a la evaluación y que

resultados se enviaron a cada estudiante. A los estudiantes correspondientes a las pruebas descartadas (n=141) se les envió un mensaje al correo electrónico oficial de la universidad que decía: “Tu prueba fue descartada del proceso de evaluación. Es probable que se haya identificado el uso de alguna inteligencia artificial. Si tienes alguna duda puedes escribir al Centro de Escritura para reservar una cita y conversar sobre tu prueba”. Ningún estudiante solicitó una aclaración sobre la razón del descarte de su prueba.

Este evento resuena con los desafíos, las preocupaciones, las oportunidades y las dificultades que se han anticipado o identificado en estudios recientes desde el lanzamiento de ChatGPT; y, además, se conecta con los componentes de dominio y confiabilidad de la teoría de la evaluación en lingüística aplicada. A partir de la relevancia de la IA en la educación y del estado de su investigación en relación con la escritura, el propósito de este estudio es explorar el desafío que representa la evaluación de la escritura académica en la educación superior en el contexto del uso masivo de aplicaciones de IA como ChatGPT.

MÉTODO

Esta investigación aplica un enfoque mixto desde un paradigma pragmático (Cresswell y Cresswell, 2018), usado principalmente para la solución de problemas. Se desarrolla un diseño secuencial exploratorio que se divide en dos fases consecutivas, una cualitativa y otra cuantitativa.

11/42

La primera fase explora la pregunta de investigación: ¿cómo evaluar la escritura en la universidad en la era de la IA? La exploración se hace a través de una entrevista semiestructurada a ChatGPT, como técnica de investigación (Bernard, 2017; Charmaz, 2006; Patton, 2015), basada en estudios recientes de conversaciones con ChatGPT (Almeida, 2023; Carretero y Gartner, 2024; Castro Nascimento y Silva Pimentel, 2023; Currie, 2023; Eysenbach, 2023; Karakose *et al.*, 2023; Menon y Shilpa, 2023; Tülübaş *et al.*, 2023). El protocolo de la entrevista consiste en 3 preguntas generales que enmarcan toda la interacción de preguntas y respuestas con la IA: ¿Cómo afecta el uso de ChatGPT el desarrollo de la función epistémica de la escritura? ¿Qué tipo de tarea de escritura no puede hacer ChatGPT? y ¿Cómo se puede determinar la originalidad y la autoría de un texto generado por ChatGPT?

Se usa la opción de seleccionar y copiar toda la conversación de ChatGPT en un archivo nuevo de Microsoft Word (2021) para facilitar la presentación de las respuestas de la entrevista completa (ver Álvarez Uribe, 2025), tal como se usa en investigaciones similares (Carretero y Gartner, 2024; Currie, 2023; Eysenbach, 2023);

están relacionados con estructuras repetitivas de oposición, abuso del gerundio y oraciones bien construidas, pero vacías.

en vez de la técnica de captura de pantalla, más usada en preguntas sobre definiciones con respuestas cortas (por ej. Karakose *et al.*, 2023; Tülübaş *et al.*, 2023). La expresión “Tú dijiste”, que resulta de copiar y pegar las preguntas y respuestas de la conversación con ChatGPT, fue reemplazada por “Investigador” para favorecer la claridad y la familiaridad del formato de entrevista. Se quitó la sangría de las numeraciones y se omitieron las viñetas de las respuestas de ChatGPT para reducir el espacio ocupado por el texto copiado, pero se mantuvieron los resaltados en negrita originales. Además, se modificó el tipo y tamaño de letra para favorecer la lectura de la entrevista completa. La conversación se desarrolló en dos momentos: abril y junio de 2024. Esta separación temporal significa que las respuestas de ChatGPT de las preguntas 1 y 2 del protocolo fueron generadas con el modelo GPT-3.5, y las respuestas de la pregunta 3 del protocolo se generaron con el modelo GPT-4. Ambos modelos corresponden con la versión libre (sin pago) de ChatGPT, usada en las dos fases del estudio.

12/42

Se realiza un análisis temático basado en Dörnyei (2007). Se usa la función experimental de codificación con IA + intención de la versión web de Atlas.ti 25. Se identifican 5 categorías principales asociadas al propósito general de la investigación y resultan 438 códigos. Se deshace la selección de 293 códigos que tienen una sola cita en la entrevista y resultan 145 códigos distribuidos en las 5 categorías identificadas: Utilidad (34 códigos), Aprendizaje (21 códigos), Desafíos (29 códigos), Inconsistencias (18 códigos) y Estrategias (43 códigos). Los códigos se revisan manualmente hasta llegar a 18 temas divididos en 3 categorías y 3 temas recurrentes en todas las categorías. Por último, se hace una selección de interacciones basada en los 3 bloques temáticos del protocolo de la entrevista para establecer la conexión con los temas identificados.

La segunda fase continúa la conversación con ChatGPT con una solicitud de evaluación de una muestra de respuestas construidas (n=40) de la prueba del español académico escrito esperado en la educación superior (EA2ES), proveniente del proceso de validación reportado en Álvarez Uribe (2023). La muestra corresponde a un grupo de evaluaciones elegidas al azar del conjunto de datos completo (N=1272). Se explora la pregunta de investigación ¿qué tan confiable es usar la IA para evaluar la escritura académica?, a través de un análisis de confiabilidad de las evaluaciones de ChatGPT con el modelo GPT-4. Las evaluaciones de la IA se comparan con las evaluaciones existentes de 5 evaluadores humanos y con los 4 niveles de Rasch-Thurstone (RT), establecidos en el proceso de validación de la prueba EA2ES. Los niveles funcionan como el patrón de referencia (*benchmark* o *gold standard*, como se conoce en inglés). Este método de comparación es una alternativa más precisa para la comparación común de sistemas de evaluación automática y de investigaciones en escritura con IA que usualmente se

comparan con evaluadores humanos o con el acuerdo o el promedio de sus evaluaciones.

La fijación de puntos de corte del proceso de validación que conduce a los 4 niveles de Rasch-Thurstone usa la adaptación del modelo de Rasch (1980) hecha por Linacre (1994) que se conoce como el modelo de facetas múltiples para evaluaciones mediadas por el juicio de evaluadores. El modelo construye un marco de medición a partir de las decisiones de evaluación de un grupo de evaluadores y sirve para confirmar el funcionamiento de una prueba como instrumento de evaluación. Después se usa el análisis multidimensional (basado en Biber, 1988) para interpretar el significado de los resultados a la luz de 93 variables lingüísticas que permiten identificar una dimensión textual, denominada Estructura de la oración, que corresponde con los 4 niveles evaluados (Alto=3, Medio-Alto=2, Medio-Bajo=1, Bajo=0). El proceso de validación explica que los niveles de Rasch-Thurstone (RT), basados en la mediana, establecen puntos de corte entre niveles de desempeño en escritura asociados a características lingüísticas que se pueden identificar en las respuestas construidas de la prueba EA2ES (Álvarez Uribe, 2023).

La exploración de la confiabilidad de las evaluaciones con IA del grupo de pruebas seleccionado (n=40) consiste en 3 análisis: un análisis de confiabilidad basado en las correlaciones de Spearman, el alfa de Cronbach (α), el coeficiente de correlación intraclase (ICC) y el coeficiente de Kappa de Cohen (k); un análisis de los indicadores de confiabilidad, ajuste y tendencia central del modelo de Rasch; y un análisis comparativo de la evaluación repetida por la IA de una prueba por cada uno de los 4 niveles de Rasch-Thurstone.

13/42

La prueba EA2ES consiste en una tarea de simulación de escritura académica que solicita completar un texto académico argumentativo ya iniciado al que le falta un argumento completo y la conclusión. El argumento debe tener una razón diferenciada del desarrollo de dos evidencias que se deben basar en dos fragmentos de información provenientes de fuentes académicas que se proporcionan junto con el texto del estímulo que debe ser completado. La instrucción dada a ChatGPT para cada evaluación presenta los criterios y la escala de evaluación de la prueba como aparece a continuación:

Evalúa la prueba de escritura académica que se presenta más abajo, teniendo en cuenta los siguientes criterios:

Escritura de razón: Establece una relación explícita de sustentación con la tesis. Establece una relación explícita con el contenido de los fragmentos de información proporcionados para el desarrollo de las evidencias. Se diferencia de la razón y la tesis del estímulo sin perder el foco ni dispersarse.

Desarrollo de evidencias: Selecciona la información necesaria de los fragmentos de información proporcionados para construir dos evidencias que respaldan el argumento. Organiza y relaciona la información seleccionada de forma lógica. Integra la información seleccionada de forma adecuada. Se diferencia con

claridad qué se toma de los fragmentos de información (sea de forma textual o parafraseada) y qué aporta la persona que escribe para construir significado. Cita según las normas APA ejemplificadas en el estímulo de la prueba. Cada evidencia establece una relación de respaldo con la razón para construir un argumento consistente con la justificación de la tesis del texto académico argumentativo de la prueba. Elabora el contenido de las fuentes citadas para relacionar las dos fuentes con la razón y las dos evidencias entre sí (anticipa, introduce, expande, explica y/o valora la información citada).

Escritura de conclusión: Reitera la tesis con un grado igual o mayor de certeza, resume los argumentos y da un sentido de cierre al texto. El sentido de cierre puede estar dado por la reiteración de la tesis y el resumen de los argumentos o por el uso de estrategias retóricas adicionales que sean pertinentes para cerrar el texto sin incluir información nueva que disperse el foco de la argumentación.

Registro académico: Mantiene el registro académico ejemplificado en el estímulo de la prueba a lo largo de todo el texto. El uso del léxico es preciso respecto al campo construido en el texto. El tenor es neutral y el modo escrito es cuidado, ambos consistentes con el estímulo de la prueba y con la tendencia a la objetividad propia de un texto académico argumentativo. La citación sigue las normas APA ejemplificadas en la prueba.

Ortografía: Sigue las convenciones ortográficas del español a lo largo de todo el texto.

14/42

Usa una escala de evaluación de 0 a 3, como niveles que expresan el cumplimiento del criterio evaluado, en donde Alto = 3, que significa que hay evidencia clara del cumplimiento del criterio; Medio-Alto = 2, que significa que no hay evidencia clara del cumplimiento del criterio, pero que tiende hacia el cumplimiento; Medio-Bajo = 1, que significa que no hay evidencia clara del cumplimiento del criterio, pero que tiende hacia el incumplimiento; y Bajo = 0, que significa que hay evidencia clara del incumplimiento del criterio. Y da una evaluación final con un nivel general de las respuestas con la misma escala de 0 a 3, en donde Alto = 3, que significa que hay evidencia clara de la competencia en español académico; Medio-Alto = 2, que significa que no hay evidencia clara de la competencia en español académico, pero que tiende hacia la competencia; Medio-Bajo = 1, que significa que no hay evidencia clara de la competencia en español académico, pero que tiende hacia la falta de competencia; y Bajo = 0, que significa que hay evidencia clara de la falta de competencia en español académico. Pon todas tus evaluaciones en una tabla con tres columnas: criterio, resultado y comentario.

Esta es la prueba con el texto que debe ser completado y que funciona como estímulo, junto con la respuesta de un estudiante universitario:

[Prueba con la respuesta construida de un estudiante]

Después de los últimos dos puntos se presenta la prueba con la respuesta construida de cada estudiante (ver Figura 1) como parte de la misma instrucción.

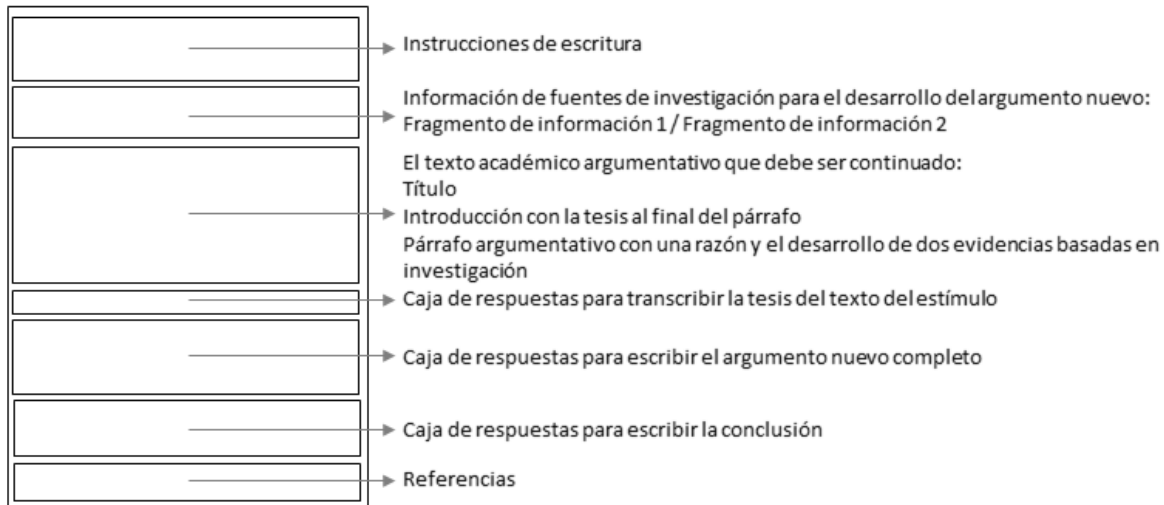


Figura 1. Esquema de la prueba EA2ES

Nota. El esquema corresponde a la Figura 41 de Álvarez Uribe (2023).

El proceso de incluir la respuesta construida se lleva a cabo con cada una de las 40 pruebas seleccionadas para este estudio de manera independiente para evitar las limitaciones de número de palabras de la ventana de contexto de la versión libre de ChatGPT. Todo el contenido de la prueba se puede ver en detalle en Álvarez Uribe (2023) en los capítulos de resultados correspondientes a la definición del dominio de evaluación y a la interpretación de los resultados de evaluación.

15/42

RESULTADOS

Fase 1: la entrevista con ChatGPT

Este apartado presenta un resumen temático de la entrevista a ChatGPT para contestar a la pregunta de investigación ¿cómo evaluar la escritura en la universidad en la era de la IA? A partir de las preguntas y respuestas de la entrevista, se sintetizan los temas de la conversación en una representación gráfica y se comentan apartados representativos de las interacciones con la IA.

El análisis manual permite refinar el análisis automático en 18 temas únicos que se distribuyen en tres categorías principales (ver Figura 2). Se identifican tres temas comunes a las tres categorías que se conectan con los bloques temáticos que organizan toda la entrevista.

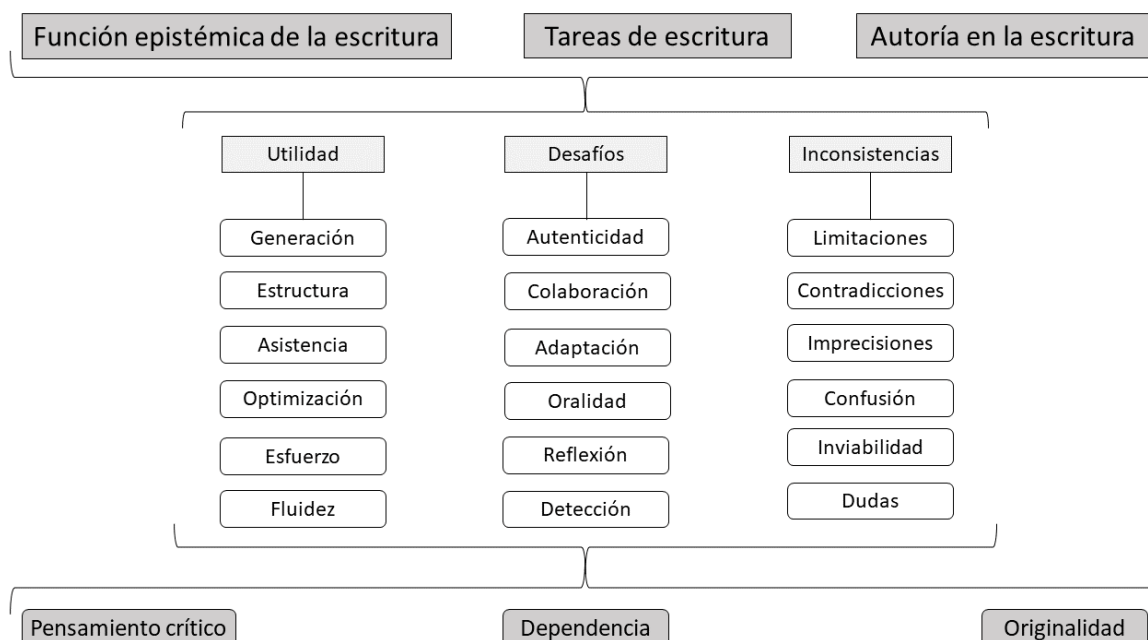


Figura 2. Representación gráfica de los temas de la entrevista a ChatGPT

Nota. Elaboración propia basada en la vista de códigos de Atlas.ti 25 (versión web).

16/42

La categoría Utilidad se centra en las declaraciones de ChatGPT sobre su efectividad y su valor para el proceso de aprendizaje de los estudiantes. Los temas asociados son: Generación (la capacidad de la IA para generar contenido útil para el aprendizaje), Estructura (cómo la IA ayuda a organizar la información de manera efectiva), Asistencia (el apoyo que la IA brinda en tareas académicas), Optimización (mejora en el proceso de aprendizaje gracias a la IA), Esfuerzo (la inversión de tiempo y atención que se puede ahorrar al usar la IA) y Fluidez (la rapidez de las respuestas de la IA y la facilidad con la que los usuarios pueden interactuar con la IA). Ejemplo de respuesta de ChatGPT:

Los estudiantes pueden pedir ayuda para desarrollar ideas o bosquejos para ensayos académicos. ChatGPT puede proporcionar sugerencias y ejemplos para la estructura y el contenido del texto.

Este ejemplo muestra el uso más común asociado a la utilidad de la IA en la escritura.

La categoría Desafíos se refiere a las dificultades y obstáculos que enfrenta el uso de IA en la escritura. Los temas asociados son: Autenticidad (consideraciones sobre la medida en que la IA produce contenido engañoso), Colaboración (definición de límites del trabajo conjunto con la IA para mejorar el aprendizaje), Adaptación (posibilidades de aplicación de la IA en distintos escenarios de escritura), Oralidad (estrategias alternativas de uso de la IA para favorecer el aprendizaje), Reflexión

(controversias sobre el uso de la IA) y Detección (contingencias para identificar y corregir errores en el contenido generado por IA). Ejemplo de respuesta de ChatGPT:

Los docentes que conocen bien el estilo de escritura de un estudiante pueden notar diferencias en su tono, complejidad del lenguaje, estructura o nivel de argumentación.

Este ejemplo destaca el desafío de identificar el trabajo original de los estudiantes para retroalimentar su nivel de desarrollo en escritura y establecer una ruta posible de aprendizaje.

La categoría Inconsistencias se enfoca en los desaciertos que pueden surgir en el uso de la IA. Los temas asociados son: Limitaciones (reservas en el uso de la IA respecto a tareas que pueden impedir el aprendizaje), Contradicciones (incongruencias en la información proporcionada por la IA), Imprecisiones (inexactitud y errores en el contenido generado por IA), Confusión (sensación de desorientación al interactuar con la IA), Inviabilidad (situaciones en las que la información generada por la IA no es aplicable o útil) y Dudas (cuestionamientos sobre la dependencia de la IA para el aprendizaje). Ejemplo de respuesta de ChatGPT:

17/42

A veces, los textos generados por IA contienen información redundante o ligeramente fuera de contexto. Un docente puede revisar si el contenido está alineado con las instrucciones y el propósito de la tarea.

Este ejemplo refleja las inconsistencias que pueden surgir en el uso de IA y el problema de la validez de la información.

Las categorías y los temas se cruzan y se superponen a lo largo de la entrevista. La recurrencia de los temas de Pensamiento crítico, Dependencia y Originalidad en las tres categorías evidencian la discontinuidad y la opacidad entre los riesgos y las oportunidades en la experiencia de uso de la IA en la escritura. El Pensamiento crítico se propone como meta del uso de IA y como estrategia para contrarrestar la Dependencia que puede producir el mismo uso. Por su parte, la Originalidad se presenta como una posibilidad que puede resultar de la interacción con una IA, pero al mismo tiempo como el desafío de cualquier contenido generado por IA en el que no es posible establecer la autoría. Estos temas se conectan con los temas predeterminados de los tres bloques de preguntas del protocolo de la entrevista: la función epistémica de la escritura, las tareas de escritura y la autoría en la escritura, como se puede observar en el ejemplo:

ChatGPT: Los educadores desempeñan un papel clave al orientar a los estudiantes sobre cómo utilizar estas herramientas de manera efectiva y al enfocarse en el desarrollo integral de habilidades, incluido el pensamiento crítico y la capacidad de construir conocimiento de manera autónoma.

Investigador: Pero entonces, ¿cuál es la ventaja de usar ChatGPT en la universidad en tareas de estudiantes si se genera dependencia, se afecta la comprensión profunda y se pierde la originalidad? ¿Qué es lo que complementa ChatGPT?

ChatGPT: El uso de herramientas como ChatGPT en entornos educativos no necesariamente está exento de ventajas. Pueden ser herramientas útiles siempre y cuando se utilicen de manera consciente y equilibrada.

El impacto de ChatGPT en la función epistémica de la escritura

Este bloque de preguntas y respuestas revela la insuficiencia de comprensión de la escritura como tecnología. Por un lado, ChatGPT menciona que la escritura es fundamental para la transmisión y preservación del conocimiento, y por otro lado se sugiere que el uso de herramientas como ChatGPT pueden poner en riesgo la construcción de conocimiento. Esta contradicción deja el cuestionamiento sobre qué tanto el uso de la IA promueve o impide la función epistémica de la escritura:

18/42

Investigador: Entonces, ¿se puede decir que chatGPT pone en riesgo la función epistémica de la escritura cuando los estudiantes prefieren solicitarle que escriba por ellos porque no construyen conocimiento ni desarrollan pensamiento crítico?

ChatGPT: El uso de tecnologías como ChatGPT para generar texto puede plantear ciertas preocupaciones en relación con la construcción del conocimiento y el desarrollo del pensamiento crítico, especialmente si se utiliza de manera indiscriminada o como sustituto completo de la propia reflexión y escritura de los estudiantes.

La falta de originalidad en las tareas de escritura producidas por ChatGPT

Este bloque de preguntas y respuestas exhibe la ambigüedad recurrente de ChatGPT respecto a la evaluación de sus propias capacidades y limitaciones. Primero afirma que no puede generar escritura creativa o argumentativa completamente original, pero luego sugiere ejemplos de tareas que podrían ser realizadas por la IA como se observa en el fragmento siguiente:

ChatGPT: Como IA, no puedo generar escritura creativa o argumentativa completamente original que no se base en patrones o información preexistente, como lo haría un ser humano al aportar ideas nuevas y originales. Además, no puedo tener experiencias personales ni emociones, lo que limita mi capacidad para crear textos con un toque personal o subjetivo. Otra limitación es la incapacidad de comprender completamente el contexto situacional o cultural, lo que puede ser crucial para adaptar una escritura eficaz en ciertas circunstancias. Tampoco puedo participar en procesos de escritura colaborativa en tiempo real, donde intervengan juicios creativos humanos. Aunque puedo realizar una amplia variedad de tareas de escritura, hay algunas que no puedo hacer o que no puedo hacer bien.

...

Investigador: No parece que haya tareas escritas que no se le puedan solicitar a ChatGPT, ¿o sí?

ChatGPT: Es cierto que, en teoría, una tarea escrita podría ser completada por una herramienta como ChatGPT si las instrucciones son lo suficientemente detalladas y específicas. Sin embargo, hay maneras de diseñar tareas que requieran un compromiso más profundo y personal del estudiante, haciendo que la contribución de ChatGPT sea menos significativa.

En resumen, las herramientas como ChatGPT pueden tener ventajas cuando se utilizan de manera consciente y como complemento en lugar de un reemplazo completo de la participación activa de los estudiantes en la construcción de conocimiento. La clave está en encontrar un equilibrio que promueva el aprendizaje autónomo y el desarrollo de habilidades críticas.

Investigador: Pero no entiendo, ¿cómo se logra el aprendizaje autónomo y se desarrollan habilidades críticas si ChatGPT genera el texto de una tarea escrita?

19/42

Desafío en la identificación de la autoría en la escritura de ChatGPT

Este bloque de preguntas y respuestas evidencia el desafío de evaluar la originalidad en el contexto de cualquier tarea de escritura académica y la necesidad de una reflexión sobre la integridad en cualquier ejercicio de escritura en la universidad:

Investigador: ¿Sería posible en este momento identificar una "huella digital" que garantice que un texto es escrito con IA?

ChatGPT: En este momento, no existe una huella digital completamente infalible que garantice que un texto ha sido escrito por IA. Sin embargo, algunos factores y patrones pueden ayudar a detectar si un texto fue generado por una IA, aunque ninguno de ellos es una garantía absoluta.

Fase 2: la confiabilidad de ChatGPT

En este apartado se presentan los resultados del análisis de confiabilidad para contestar a la pregunta de investigación ¿qué tan confiable es usar la IA para evaluar la escritura académica?, a través de la evaluación de una muestra (n=40) de respuestas construidas provenientes del proceso de validación de la prueba EA2ES (Álvarez Uribe, 2023).

Análisis de indicadores de confiabilidad

La exploración inicial de la confiabilidad consiste en el análisis de las correlaciones de Spearman, usadas para medidas ordinales, entre los niveles de Rasch-Thurstone (RT) y la evaluación holística de la competencia global (CG) de ChatGPT (IA), los 5 evaluadores (E1, E2, E3, E4 y E5) y el valor que resulta del acuerdo entre los 5 evaluadores (A5), basado en la moda, tal como se utiliza en la evaluación de pruebas mediadas por el juicio de evaluadores (Dimova *et al.*, 2020) en vez del uso del promedio (ver Tabla 1).

Tabla 1. Correlaciones de Spearman entre IA, humanos y los niveles RT

20/42

	RT	IA	E1	E2	E3	E4	E5	A5
RT	—							
IA	<u>0.618</u> ***	—						
E1	0.728***	0.768***	—					
E2	0.661***	0.441**	0.490**	—				
E3	0.766***	0.306	0.478**	0.367*	—			
E4	0.775***	0.590***	0.682***	0.420**	0.607***	—		
E5	0.734***	0.578***	0.696***	0.627***	0.619***	0.562***	—	
A5	0.843***	<u>0.588</u> ***	0.806***	0.705***	0.733***	0.717***	0.846***	—

Nota. * $p < .05$, ** $p < .01$, *** $p < .001$. RT=Rasch-Thurstone, E=evaluador, A5=acuerdo de 5 evaluadores

Los valores subrayados muestran que las correlaciones entre las evaluaciones holísticas de la competencia global en español académico de la IA con los niveles de Rasch-Thurstone (RT) y con el acuerdo de los 5 evaluadores humanos (A5) son más bajas que las de cada evaluador de manera individual respecto a los mismos niveles de comparación. Aunque junto con las correlaciones de E2 son correlaciones

moderadas, las correlaciones de los demás evaluadores (E1, E3, E4 y E5) son altas, según las recomendaciones de Hinkle *et al.* (2003).

Luego se estima el coeficiente alfa de Cronbach (α) y el coeficiente de correlación intraclase (ICC) entre cada evaluador y la IA contra los niveles de Rasch-Thurstone. El alfa de Cronbach se utiliza como una estimación de la consistencia entre medidas; y el ICC se usa con el modelo 2,1 del módulo seolmatrix (Seol, 2023) del software jamovi (The jamovi project, 2022), para medir el acuerdo absoluto de una muestra posible de evaluadores con efectos aleatorios (ver Tabla 2).

Tabla 2. *Coeficientes de confiabilidad respecto al valor de RT*

Evaluación	α	ICC
E1	0.829	0.671
E2	0.790	0.644
E3	0.817	0.694
E4	0.844	0.676
E5	0.834	0.619
A5	0.910	0.818
IA	0.757	0.603

Todos los valores del coeficiente α están por encima del valor de .70, considerado generalmente como un valor de referencia aceptable para la confiabilidad (Taber, 2018). El valor de la IA es el menor, seguido por el del evaluador 2 (E2). Los demás valores de α están por encima de .80, que se considera un valor meta de confiabilidad (Gliem y Gliem, 2003).

Los valores de ICC muestran una confiabilidad moderada entre .50 y .75 (Koo y Li, 2016), aunque el coeficiente de la IA es el más bajo. La confiabilidad entre los niveles de Rasch-Thurstone (RT) y el acuerdo entre los 5 evaluadores (A5) se encuentra en el rango más alto. Adicionalmente se calcula el coeficiente de acuerdo de Kappa respecto a los niveles de Rasch-Thurstone. Este coeficiente incluye las diferencias de evaluación no capturadas por el acuerdo exacto y se reporta en la mayoría de estudios de evaluación (ver Tabla 3).

Tabla 3. *Coeficiente de Kappa con los niveles de Rasch-Thurstone*

Evaluación	k	Acuerdo	p valor
E1	0.378	Aceptable	<.001
E2	0.395	Aceptable	<.001
E3	0.449	Moderado	<.001
E4	0.352	Aceptable	.001
E5	0.290	Aceptable	.006
A5	0.652	Bueno	<.001
IA	0.287	Aceptable	.011

La IA presenta el valor de Kappa de acuerdo más bajo, aunque se categoriza, según las recomendaciones de Altman (1999), en el rango de acuerdo aceptable, junto con los demás evaluadores. El valor del evaluador 3 (E3) y el valor de acuerdo entre los 5 evaluadores (A5) son los únicos que se encuentran en rangos más altos que el resto de valores. Adicionalmente, se calcula el valor del estadístico de Kappa entre IA y A5. El resultado es un valor de 0.190, que equivale a un acuerdo despreciable con un valor de p no significativo de 0.087.

22/42

Análisis del modelo de Rasch

Se comparan los valores de confiabilidad general del modelo de Rasch entre la evaluación holística y la analítica en 3 situaciones distintas: sin incluir la IA como evaluador en el mismo marco de medición, incluyendo la IA, y reemplazando el último evaluador (E5) por las evaluaciones de la IA (ver Tabla 4).

Tabla 4. *Confiabilidad general del modelo de Rasch*

Evaluación	Sin IA	Con IA	Con IA – E5
Holística (Competencia global)	.85	.87	.84
Analítica (5 criterios)	.95	.97	.95

La confiabilidad general del modelo de Rasch es similar al coeficiente alfa de Cronbach cuando se usa como una medida de la consistencia interna de una prueba; y se puede interpretar de la misma manera: los valores por encima de .70 son aceptables y por encima de .80 son deseables. De este modo, la confiabilidad se puede interpretar como la probabilidad de obtener los mismos resultados con una población equivalente. Se evidencia que las diferencias de confiabilidad entre la evaluación holística y la analítica son apreciables, mientras que las diferencias entre las 3 situaciones son mínimas.

El comportamiento de los evaluadores y la IA se revisa con las estadísticas de ajuste del modelo de Rasch. El ajuste indica qué tan predecible o incierto es el comportamiento de evaluación de los jueces humanos y la IA según se acerquen o se alejen al valor ideal de 1.0. Si los valores se alejan por debajo de 1.0, indican un comportamiento más predecible; y por encima de 1.0 indican un comportamiento más incierto. El ajuste lejano se enfoca en los extremos del marco de medición; y el ajuste cercano en el centro. (ver Tabla 5).

Tabla 5. *Ajuste de la evaluación holística (Competencia global)*

Evaluación	Ajuste lejano	Ajuste cercano
E1	.79	.84
E2	1.22	1.27
E3	1.48	1.51
E4	.78	.81
E5	.63	.66
IA	.90	.86

Los valores de la IA muestran que su comportamiento es el que mejor se ajusta al modelo de Rasch. Linacre (2002) recomienda que los valores de ajuste deben estar entre 0.5 y 1.5 para evaluaciones mediadas por el juicio. La única medida que está fuera de este rango de aceptabilidad es la del ajuste cercano del evaluador 3 (E3). Luego se revisan las estadísticas de ajuste de la evaluación analítica de 5 criterios (ver Tabla 6).

23/42

Tabla 6. *Ajuste de la evaluación analítica (5 criterios)*

Evaluación	Ajuste lejano	Ajuste cercano
E1	.72	.75
E2	1.15	1.18
E3	1.42	1.36
E4	1.13	1.14
E5	.88	.93
IA	.70	.63

Las estadísticas de ajuste de la evaluación analítica muestran un comportamiento relativamente normal. Todos los valores están dentro del rango recomendado entre 0.5 y 1.5. De todos modos, los valores de la IA se identifican con un comportamiento que tiende a ser más predecible, al tener los valores más bajos. Esta tendencia hacia la predictibilidad de las evaluaciones de la IA demanda una confirmación con un análisis de tendencia central.

La tendencia central permite identificar evaluadores más predecibles de lo normal. Se puede inspeccionar con el modelado de cada evaluador como si su escala fuera única para permitir la comparación. La tendencia central se identifica si un evaluador tiene un número alto de evaluaciones en las categorías centrales de la escala (ver Tabla 7).

Tabla 7. *Evaluaciones en las categorías centrales*

Categoría	% E1	% E2	% E3	% E4	% E5	% IA
Medio-Bajo	30	27	34	30	28	42
Medio-Alto	30	39	32	27	23	36
Total	60	66	66	57	51	78

Entre las dos categorías centrales, la más elegida es la de Medio-Bajo. Sin embargo, el evaluador 2 (E2) muestra un patrón opuesto al resto, al presentar el mayor porcentaje en el nivel Medio-Alto. El sesgo de tendencia central de la IA se puede evidenciar con el porcentaje más alto de evaluaciones en la suma de las categorías centrales (resaltado en gris). El valor total de 78 se encuentra 12 puntos porcentuales por encima del porcentaje total de los dos evaluadores con el porcentaje más alto de 66 (E2 y E3). El mapa de variables (conocido como el mapa de Wright) modela cada evaluador y la IA con una escala independiente y muestra gráficamente la magnitud de la tendencia resaltada en gris en la columna correspondiente a la IA (ver Figura 3).

Measr	-Evaluadores	+Estudiantes	-Criterios	E1	E2	E3	E4	E5	IA		
3	+	+	+	+	(3)	+	(3)	+	(3)	+	(3)
		7 11									---
2	+	+	+	+	+	+	+	+	+	+	+
		34 9 40 37		---	---	---					
1	+	+	+	+	+	+	+	+	+	+	+
	E1 E5	2 15 33 38 12 35 39 32		2	2		2	2		2	
	IA	16 8 13 4	DE EC ER			---	---				
0	+	+	+	+	+	+	+	+	+	+	+
	E2 E4	21 23 36 6 30 5 10	RA O	---	---						
-1	+	+	+	+	+	+	+	+	+	+	+
	E3	3 19 24 28 14 29 31 1 20 22 26 17 25 27		1	1	1					
-2	+	+	+	+	+	+	+	+	+	+	+
		18									
-3	+	+	+	+	+	+	+	+	+	+	+
-4	+	+	+	+	(0)	+	(0)	+	(0)	+	(0)

Figura 3. Mapa de variables del modelo de Rasch por evaluador

La mayoría de los estudiantes evaluados quedan ubicados en las categorías centrales de la IA, en comparación con los demás evaluadores. En el nivel 3 (Alto) de la IA se encuentra la respuesta construida correspondiente al texto 7 y en el nivel 0 (Bajo) la del texto 18. Todas las demás respuestas se encuentran en las categorías centrales.

Análisis de evaluaciones repetidas de la IA

Se selecciona una prueba al azar por cada uno de los 4 niveles de Rasch-Thurstone, se evalúan de nuevo con la misma instrucción en ChatGPT y se comparan los valores respecto a los niveles validados de Rasch-Thurstone (ver Tabla 8).

Tabla 8. *Evaluaciones repetidas de la IA*

Texto	Ocasión	ER	DE	EC	RA	O	CG	RT
0007	1	2	2	2	2	2	2	3
	2	2	2	2	2	3	2	
0015	1	2	2	2	2	1	2	2
	2	1	1	1	1	2	1	
0020	1	1	1	1	0	1	1	1
	2	0	0	0	0	1	0	
0024	1	1	0	1	1	0	0	0
	2	1	1	1	1	1	1	

En 5 de 8 ocasiones, resaltadas en gris, hay una discrepancia de un nivel entre la evaluación de la competencia global (CG) de la IA y los niveles de Rasch-Thurstone (RT) provenientes del proceso de validación de la prueba EA2ES. Llama la atención la discrepancia de las dos ocasiones de evaluación en el caso del texto 0007 de nivel Alto (3) y la diferencia de un nivel entre las dos ocasiones en las evaluaciones de la CG en los tres textos restantes, correspondientes a los demás niveles de desempeño. Además, en los criterios de desarrollo de evidencias (DE) y ortografía (O) hay una discrepancia de 1 nivel entre la ocasión 1 y 2 en 3 de las 4 pruebas. A pesar del tamaño pequeño de la muestra (n=4), las evaluaciones repetidas de la IA en dos ocasiones distintas muestran un grado considerable de inconsistencia intraevaluador y falta de acuerdo con los niveles de Rasch-Thurstone.

DISCUSIÓN

Sobre la conversación con ChatGPT

En general, la entrevista se refiere a los tipos de tareas de escritura que se pueden hacer en el contexto de la educación superior, a las capacidades de la IA para producir una tarea de escritura dada y a las consideraciones y recomendaciones éticas de su uso en relación con la originalidad y la autoría. La entrevista completa (ver Álvarez Uribe, 2025) es un resultado autoevidente, tal como se presenta en otros estudios de conversaciones con ChatGPT como el de pensamiento histórico de Carretero y Gartner (2024), el de aprendizaje de medicina nuclear de Currie (2023), el de Castro Nascimento y Silva Pimentel (2023) sobre conocimiento en química y

el de Eysenbach (2023) sobre educación en medicina. Las respuestas revelan argumentos circulares que en muchas ocasiones resultan en contradicciones, imprecisiones o inconsistencias. Si los riesgos de la IA en la escritura dependen de su uso y, al mismo tiempo, el uso de la IA puede crear una “dependencia excesiva”, no es claro cómo se desarrolla un uso adecuado. Las recomendaciones de ChatGPT resultan autorreferenciales.

El primer bloque de preguntas sobre la función epistémica de la escritura evidencia con claridad la tensión existente entre el papel de la escritura en la construcción del conocimiento y cómo la escritura basada en una IA puede ser a la vez un atajo y una barrera para construirlo. La evaluación del español académico en relación con el aspecto de la construcción de conocimiento, que constituye el núcleo del constructo, queda enmarcada en la dificultad de evidenciar esa construcción en tareas asistidas por IA. Álvarez Uribe (2023) afirma, antes del uso masivo de la IA en la escritura, que “la amplitud de la caracterización del español académico resulta en la dificultad de evaluar esta variedad funcional del lenguaje.” (p. 6). El cambio, en la era de la IA, parece ser el aumento de la complejidad de la misma dificultad.

En el segundo bloque de preguntas sobre las tareas de escritura que no puede hacer ChatGPT, la respuesta corta es que ChatGPT puede hacer cualquier tarea de escritura que se le solicite con una instrucción detallada y específica, sin importar qué tan reflexiva, profunda o personal sea la tarea. ChatGPT recomienda la inclusión de elementos creativos en las tareas que se le soliciten a los estudiantes, para luego señalar que uno de los usos comunes de ChatGPT es la práctica de escritura creativa. Las alternativas propuestas que parecen más viables son el seguimiento personalizado a la escritura supervisada de los estudiantes, el uso de la expresión oral y la evaluación en tiempo real.

Aunque la conversación con ChatGPT abre el espacio para considerar opciones alternativas de evaluación de la escritura, no queda claro cómo se puede abordar el aprendizaje de la estructura jerárquica de la argumentación que permite la comunicación académica como una forma eficaz de circulación del conocimiento. La estructura básica de texto que ChatGPT describe como la más común es precisamente esa estructura que se trabaja en la mayoría de cursos de alfabetización académica. ChatGPT afirma que esa estructura “asegura claridad y coherencia en la presentación de ideas” (ver Álvarez Uribe, 2025, p. 11). No es coincidencia que la tarea de evaluación del español académico de la prueba EA2ES, que se evalúa en la segunda fase de esta investigación, se haya diseñado con la misma estructura para dar cuenta de la representatividad y la relevancia respecto a la escritura académica en general.

Las respuestas de ChatGPT enfatizan en que la clave está en “encontrar un equilibrio” entre las ventajas y los riesgos para ocuparse de los objetivos esenciales de la educación “a pesar de la presencia de generadores de texto automáticos” (ver

Álvarez Uribe, 2025, p. 5). Las estrategias que propone ChatGPT para encontrar el equilibrio del uso de IA en la escritura resultan en ejemplos de tareas que promueven la reflexión. Pero como afirma Lo (2023), cuando ChatGPT puede estar involucrado en cualquier proceso de escritura, sea reflexivo o no, “es difícil hacer seguimiento a los problemas reales de aprendizaje de los estudiantes” (p. 10).

El tercer bloque de preguntas de la conversación evidencia el problema central del uso de IA en la escritura: la relación entre originalidad y autoría. Para OpenAI (2024) el usuario es el propietario del contenido, pero la conclusión es otra en la conversación con ChatGPT. La discusión sobre la originalidad y la autoría no se trata de una nueva definición que incluya la mediación tecnológica como propone Luo (2024), sino más bien de un asunto de responsabilidad como propone Lanier (2024).

En este punto quedan claras las limitaciones obvias de la IA: la ausencia de conciencia, emociones, experiencias personales y de comprensión de la cultura que se pueden rastrear en los textos escritos por humanos. Pero sucede que accedemos a gran parte de esas dimensiones humanas a través del lenguaje. La expresión humana de algunos sentimientos con palabras son las palabras, así exista o no un sentimiento real detrás de ellas. Los ejemplos de ChatGPT de la escritura de un aforismo al estilo de Nietzsche y el de la respuesta a una pregunta por un sentimiento (ver Álvarez Uribe, 2025, p. 13-14) evidencian la zona difusa de expresión de emociones y pensamientos en palabras. Tal como concluye la revisión de Lo (2023), queda claro que para evaluar la escritura en la era de la IA hay que considerar otras estrategias que incorporen componentes no digitales como presentaciones orales, pues “estos componentes requieren que los estudiantes demuestren sus habilidades en tiempo real y en persona” (p. 10).

En este sentido, es necesario continuar la reflexión sobre la posibilidad de reformar los formatos de los exámenes, sin perder de vista los avances en la seguridad de las pruebas que se han consolidado como prácticas comunes desde la pandemia de Covid-19 e incluyen el procesamiento de datos de flujo de clics (Hao y Fauss, 2024), técnicas de supervisión y bloqueo de dispositivos electrónicos (Hidayanto *et al.*, 2025). Técnicas que son percibidas por estudiantes como transparentes y como garantías de imparcialidad en la evaluación (Kuleva y Miladinov, 2024).

Además, hay una limitación técnica sobre la autoría que apenas se está empezando a explorar. La idea de la huella digital (*watermarking*, como se conoce en inglés) resuena con la concepción de dignidad de los datos promovida por Lanier (2023; 2024) que defiende “el derecho a no ser manipulado por medios computacionales” (Lanier, 2023, párr. 13). Esta posibilidad de rastreo de la autoría no tiene una ruta de desarrollo clara en este momento, tal como lo evidencia Sala (2024). Su materialización necesita de una acción coordinada de alto alcance a

distintos niveles de la industria de la IA, que debe incluir aspectos de sostenibilidad y de impacto económico y ambiental, tal como se recoge en el reporte anual de inteligencia artificial centrada en humanos de la universidad de Stanford (HAI, 2023), que avanza en una propuesta dirigida a desarrolladores (HAI, 2024) para diferenciar el contenido generado por distintos modelos de lenguaje de gran tamaño.

Aunque hay oportunidades, hay desafíos, preocupaciones y riesgos reales, más allá de las voces de apoyo o rechazo del uso de IA en educación que se evidencian en la literatura revisada para este estudio. El estudio de Levine *et al.* (2024) presenta de manera detallada el uso de IA ideal que permite ver la posibilidad del equilibrio mencionado en la entrevista a ChatGPT, al que se alude sin ejemplos concretos. Es evidente que se necesitan unas condiciones adecuadas y un entrenamiento especializado para que un uso ideal de la IA funcione de manera equilibrada entre los riesgos y las oportunidades.

Sobre la confiabilidad de las evaluaciones hechas por ChatGPT de la fase 2

En general, los resultados del análisis de confiabilidad evidencian que la IA se puede usar como un evaluador adicional para construir un marco de medición confiable, tal como lo recomiendan investigaciones recientes de uso de IA en la evaluación de la escritura (Mizumoto y Eguchi, 2023; Naismith *et al.*, 2023). Pero para lograr una evaluación basada exclusivamente en una IA, se requiere el desarrollo de un sistema sofisticado como el de e-rater del TOEFL (Burstein *et al.*, 2004). Estos sistemas superan las limitaciones de confiabilidad que se encuentran en el uso de las versiones de acceso libre de herramientas como ChatGPT y, al mismo tiempo, requieren de una infraestructura fuera del alcance de docentes individuales. De este modo, se confirma la complejidad de desarrollar sistemas automatizados eficaces para la calificación de ensayos y la evaluación de escritura en español. En este sentido, la promesa del uso de ChatGPT para la evaluación automatizada de la escritura (AWE) no se encuentra aún al alcance de una instrucción. Si se quieren lograr evaluaciones confiables e imparciales, se mantiene un requerimiento de conocimientos especializados en lingüística computacional, el apoyo de equipos interdisciplinarios y una inversión de alto costo.

Por otro lado, el sesgo de tendencia central de la IA identificado en este estudio precisa la sugerencia de Lundgren (2024) sobre la aversión al riesgo de GPT-4 que “parece evitar la asignación de calificaciones bajas o altas” (p. 3). Este sesgo pasa desapercibido en un análisis general de coeficientes de confiabilidad. En este estudio se identifica con las estadísticas de ajuste del modelo de Rasch y se verifica con el modelado del uso de la escala de cada evaluador y de la IA de manera independiente.

Los resultados de este estudio permiten evidenciar que las comparaciones de evaluaciones de IA con evaluadores humanos son limitadas para verificar el comportamiento consistente de una prueba de escritura como instrumento de evaluación. La limitación identificada por Lundgren (2024) apunta en esta dirección: “la ausencia de comparaciones entre evaluadores humanos para el mismo conjunto de ensayos limita el entendimiento de cómo se compara la confiabilidad de GPT-4 frente a la varianza típica de evaluadores humanos” (p. 1).

Los coeficientes de confiabilidad presentados en este estudio usan un referente proveniente de un proceso de validación que se expresa en los niveles de Rasch-Thurstone, respaldados por un análisis de variables lingüísticas. De este modo, los resultados del coeficiente de Kappa de 0.43 y la correlación de Spearman de .82 de Naismith *et al.* (2023), respecto a indicadores de coherencia, son más comparables, aunque no coincidentes, con el valor de Kappa de 0.287 y la correlación de Spearman de 0.618 de este estudio, entre la IA y los niveles de Rasch-Thurstone. La explicación de la diferencia de valores puede estar en la diferencia en el tamaño de la muestra, en las características específicas de ambas tareas y en la construcción de los marcos de medición con los que se realiza la comparación.

30/42 Por otro lado, la correlación de Sparman de 0.588 entre la IA y el acuerdo de los 5 evaluadores (A5) de la prueba EA2ES se acerca al valor de 0.652 del estudio de Bucol y Sangkawong (2024), a pesar de su tamaño de muestra pequeño con 10 ensayos. Además, el coeficiente de Kappa no estadísticamente significativo de 0.190 entre la IA y el acuerdo entre los 5 evaluadores (A5) coincide con las relaciones no significativas de los estudios de Lundgren (2024) y Awidi (2024), aunque difiere levemente de los resultados de Mizumoto y Eguchi (2023) que compara la IA con profesionales en evaluación del corpus TOEFL11. Adicionalmente, el análisis de evaluaciones repetidas hechas por la IA a 4 pruebas puede relacionarse con las diferencias en la identificación de errores en dos ocasiones distintas del estudio de Pfau *et al.* (2023) y con la indicación de la limitación de comparaciones entre evaluadores humanos señalada por Lundgren (2024).

De este modo, este estudio confirma la necesidad de entender mejor la construcción de marcos de medición comunes como los que se pueden construir con el modelo de Rasch, que sirven como un control de calidad o como un patrón de referencia más estable que la comparación exclusiva entre humanos y una IA. La ausencia de procedimientos confiables de fijación de puntos de corte entre niveles de desempeño de pruebas escritas, identificada en Álvarez Uribe (2023) para pruebas de español académico y revisada para el inglés por Aryadoust *et al.* (2021), deja un espacio amplio de exploración de la evaluación con IA más allá del acuerdo entre evaluadores humanos.

En conjunto, las dos fases de esta investigación coinciden con algunas de las conclusiones más representativas de la literatura revisada. Por un lado, los

resultados de Herbold *et al.* (2023) muestran “una fuerte indicación de que el temor que tienen muchos profesionales de la enseñanza es razonable” (p. 9). Rudolph *et al.* (2023) complementa esta conclusión con un comentario sobre la ironía del uso de la IA: “primero se usa una IA para evitar el control de una segunda IA que es evaluada por una tercera IA; y lo único que hacen los humanos es presionar un par de teclas, y nadie aprende nada” (p. 354). Por otro lado, Lanier hace un llamado a cambiar el foco de atención: “Ver la inteligencia artificial como una forma de colaboración social es más factible: nos da acceso a la sala de máquinas, que está hecha de personas” (2023, párr. 30)

El cambio de la educación a través de la IA es una realidad que se hace patente desde la perspectiva de estudiantes con la posibilidad de resolver cualquier tarea de escritura. Desde la perspectiva docente, las posibilidades de enseñanza parecen mezclarse con una lista de nuevas responsabilidades en un contexto de acceso masivo a tecnologías en constante evolución que facilitan a los estudiantes poco motivados la decisión de evitar el proceso de aprendizaje en la misma línea de razonamiento señalada en la revisión de İpek *et al.* (2023). Las implicaciones pedagógicas son evidentes. Steele (2023) defiende que es posible promover el pensamiento crítico si se usa ChatGPT para comprender textos, agregar conocimientos y entender las convenciones de los géneros (Steele, 2023). Pero, estamos en un estadio de exploración del uso de la IA en la educación para poder identificar usos concretos con un impacto real en el aprendizaje.

31/42

En términos de limitaciones de esta investigación, el análisis de la conversación con ChatGPT se dedica a resumir los temas predominantes y a relacionar las interacciones consideradas más relevantes para el propósito de este estudio sin hacer una comparación con otras aplicaciones de acceso masivo de chatbot de IA como Meta AI (Touvron *et al.* 2023), Gemini AI (Anil *et al.*, 2023), Claude AI (Anthropic, 2024) o Copilot AI (Spataro, 2023). Tampoco se validan todas las respuestas de ChatGPT con un análisis contrastado de sus capacidades más allá de algunos ejemplos. De todos modos, se presenta la entrevista completa para su escrutinio en Álvarez Uribe (2025). Por otro lado, en el análisis de confiabilidad no se calcula el coeficiente ponderado de Kappa en su variante cuadrática, que es considerado el estándar en el análisis de modelos predictivos de la ciencia de datos y en el área de evaluación automática de la escritura en general. Sin embargo, en la revisión de literatura hecha para este estudio, no se encontraron suficientes conjuntos de datos con ese coeficiente para comparar.

Por razones de tiempo y espacio, queda pendiente un análisis de los 240 comentarios de justificación de evaluación de ChatGPT (un comentario por cada uno de los 5 criterios más un concepto global por 40 pruebas) y una comparación con los comentarios de pruebas evaluadas reportado en Álvarez Uribe (2023). Este análisis requiere un estudio independiente sobre la confiabilidad de la retroalimentación de

la IA en la línea de estudios como el de Steiss *et al.* (2024). También queda pendiente un análisis textual de las respuestas construidas identificadas en la tabla de respuestas inesperadas del análisis de residuos del modelo de Rasch en el que la IA muestra un comportamiento más impredecible respecto al criterio de Desarrollo de las evidencias y al criterio de Ortografía de la prueba EA2ES. Este resultado coincide, en parte, con el análisis del criterio de la Ortografía en Álvarez Uribe (2023) y se puede relacionar con el análisis de la discrepancia de identificación de errores de bajo nivel de Pfau *et al.* (2023), relacionados con errores de ortografía. Además, deja la pregunta sobre la razón de la inconsistencia de evaluación en el aspecto que se supone más estable y normativo de la lengua.

En resumen, los resultados y las conclusiones de esta investigación se limitan por la naturaleza transitoria propia de un estudio exploratorio, que pretende abrir líneas de investigación sobre el uso de la IA en la escritura en español.

CONCLUSIONES

En el entorno cambiante del desarrollo de la IA, cualquier conclusión puede ser reemplazada rápidamente por los resultados de estudios realizados con nuevas versiones de algún modelo de IA existente o por un modelo nuevo con un funcionamiento distinto. Sin embargo, este estudio exploratorio puede aportar a la acumulación de evidencia de la investigación en progreso sobre el uso de IA en la enseñanza y evaluación de la escritura con dos conclusiones provisionales que se expresan como dos necesidades. Por un lado, la evaluación de la escritura académica necesita incrementar el seguimiento individual de estudiantes, las tareas en tiempo real, las actividades de expresión oral y las estrategias de supervisión de autoría. Por otro lado, la evaluación con IA necesita considerar la posibilidad de su sesgo de tendencia central y el uso de marcos de medición más allá del acuerdo entre evaluadores humanos.

Financiación

El autor no declara haber recibido financiación institucional, gubernamental o privada para este estudio.

Contribución de los autores

Sergio Álvarez Uribe: conceptualización, formulación de objetivos, diseño metodológico, recolección de datos, análisis formal, validación, redacción del borrador original, redacción, revisión y edición.

Conflictos de interés

El autor declara no tener ningún conflicto de interés en relación con la redacción, investigación o publicación de este artículo.

Implicaciones éticas

El autor declara que la escritura y publicación de este artículo no incurre en implicaciones éticas.

Uso de IA

El autor de este artículo declara que no se usó ninguna inteligencia artificial para el proceso de planeación, escritura y revisión. Se usó la IA Elicit (<https://elicit.com/>) para la búsqueda de fuentes incluidas en la introducción; la función experimental de IA en Atlas.ti 25 para la identificación automática de códigos en la entrevista a ChatGPT antes del análisis manual; y la aplicación de IA generativa ChatGPT (<https://openai.com/chatgpt/>) como parte de la investigación que se acredita dentro del artículo sin ninguna ambigüedad.

33/42

REFERENCIAS

- Aljuaid, H. (2024). The Impact of Artificial Intelligence Tools on Academic Writing Instruction in Higher Education: A Systematic Review. *Arab World English Journal (AWEJ) Special Issue on ChatGPT*, 26-55. <https://doi.org/10.24093/awej/ChatGPT.2>
- Allen Cu, M. y Hochman, S. (22 de enero de 2023). *Scores of Stanford students used ChatGPT on final exams, survey suggests*. The Stanford Daily. <https://stanforddaily.com/2023/01/22/scores-of-stanford-students-used-chatgpt-on-final-exams-survey-suggests/>
- Almeida, J. G. (2023). Claves para entender la relación entre la inteligencia artificial y la creatividad humana: una conversación con ChatGPT. <https://doi.org/10.13140/RG.2.2.11522.76480>
- Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithms for assessment and scoring. *European Journal of Education*, 58(1), 98-110. <https://doi.org/10.1111/ejed.12542>
- Altman, D. G. (1999). *Practical statistics for medical research*. Chapman & Hall/CRC Press.
- Álvarez Uribe, S. (2023). Validación de las interpretaciones de una prueba diagnóstica para evaluar el español académico escrito en la educación

- superior (2023). [Tesis doctoral, Universidad Antonio de Nebrija]. Biblioteca Nebrija. <https://biblioteca.nebrija.es/cgi-bin/repositorio?TITN=137362>
- Álvarez Uribe, S. (2025). Una conversación con ChatGPT (entrevista). Zenodo. <https://doi.org/10.5281/zenodo.15619070>
- Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R. ... y Blanco, L. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805
- Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku. Recuperado de https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf
- Aryadoust, V., Ng, L. Y. y Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, 38(1), 6-40. <https://doi.org/10.1177/0265532220927487>
- Astutik, I., Widiati, U., Ratri, D. P., Jonathans, P. M., Nurkamilah, Devanti, Y. M. y Harfal, Z. (2024). Transformative Practices: Integrating Automated Writing Evaluation in Higher Education Writing Classrooms - A Systematic Review. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, 6(3), 423-441. <https://doi.org/10.23917/ijolae.v6i3.23675>
- Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence*, 6, 100226, 1-17. <https://doi.org/10.1016/j.caeai.2024.100226>
- Barrett, A. y Pack, A. (2023). Not quite eye to AI: student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education*, 20(59), 1-24. <https://doi.org/10.1186/s41239-023-00427-0>
- Berrezueta-Guzman, J., Malache-Silva, L. y Krusche, S. (2023). ChatGPT-4 as a Tool for Reviewing Academic Books in Spanish. En: Berrezueta, S. (Ed.). *Proceedings of the 18th Latin American Conference on Learning Technologies (LACLO 2023)*. LACLO 2023. *Lecture Notes in Educational Technology* (pp. 384-397). Springer. https://doi.org/10.1007/978-981-99-7353-8_29
- Bernard, H. R. (2017). *Research methods in anthropology: Qualitative and quantitative approaches*. Rowman & Littlefield.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511621024>
- Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W. y Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and

- rigour. *International Journal of Educational Technology in Higher Education*, 21, 1-41. <https://doi.org/10.1186/s41239-023-00436-z>
- Bucol, J. L. y Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International*, 1-16. <https://doi.org/10.1080/14703297.2024.2363901>
- Burstein, J., Chodorow, M. y Leacock, C. (2004). Automated essay evaluation: The criterion online writing system. *AI Magazine*, 25(3). 27-36. <https://doi.org/10.1609/aimag.v25i3.1774>
- Cahill, A., Bruno, J., Ramey, J., Ayala Meneses, G., Blood, I., Tolentino, F., Lavee, T. y Andreyev, S. (6-11 de junio de 2021). Supporting Spanish Writers using Automated Feedback [Conference]. En A. Sil y X. V. Lin (Eds.). *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations* (pp. 116-124). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-demos.14>
- Carretero, M. y Gartner, E. (2024). Artificial Intelligence and historical thinking: a dialogic exploration of ChatGPT/Inteligencia Artificial y pensamiento histórico: una exploración dialógica del ChatGPT. *Studies in Psychology*, 45(1), 80-102. <https://doi.org/10.1177/02109395241241379>
- Castien, J. I., Cuenca, C. y Zlobina, A. (2021). Aplicación en el nivel universitario del programa informático Gallito-API para la corrección y evaluación de breves ejercicios escritos. En M. A. Martín López y C. S. Rodríguez (Coord.). *Cuestiones transversales en la innovación de la docencia y la investigación de las ciencias sociales y jurídicas: especial referencia al impacto del covid-19, las nuevas tecnologías y metodologías, las perspectivas de género y la diversidad* (pp. 797-820). Dykinson.
- Castro Nascimento, C. M. y Silva Pimentel, A. (2023). Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6), 1649-1655. <https://doi.org/10.1021/acs.jcim.3c00285>
- Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Sage Publications. <https://doi.org/10.4135/9781071878811>
- Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage publications.
- Chen, G. H., Chen, S., Liu, Z., Jiang, F. y Wang, B. (2024). Humans or LLMs as the judge? a study on judgement biases. En Y. Al-Onaizan, M. Bansal y Y. N. Chen (Eds.). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 8301-8327). Association for Computational Linguistics <https://doi.org/10.18653/v1/2024.emnlp-main.474>

- Chomsky, N. (marzo 8 de 2023). *The false promise of ChatGPT*. New York Times. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- Creswell, J. W. y Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Currie, G. (2023). A conversation with ChatGPT. *Journal of Nuclear Medicine Technology*, 51(3), 255-260. <https://doi.org/10.2967/jnmt.123.265864>
- Da Cunha, I., Montané, M. A. e Hysa, L. (2017). The arText prototype: An automatic system for writing specialized texts. En A. Peñas, A. y A. Martins (Eds.). *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017). Software Demonstrations* (pp. 57–60). Association for Computational Linguistics.
- Dimova, S., Yan, X. y Ginther, A. (2020). *Local language testing. Design, implementation and development*. Routledge. <https://doi.org/10.4324/9780429492242>
- Dong, Y. (2023). Revolutionizing Academic English Writing through AI-Powered Pedagogy: Practical Exploration of Teaching Process and Assessment. *Journal of Higher Education Research*, 4(2), 52-57.
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford university press.
- 36/42 Durmus, E., Lovitt, L., Tamkin, A., Ritchie, S., Clark, J., y Ganguli, D. (2024). *Measuring the Persuasiveness of Language Models*. Anthropic. <https://www.anthropic.com/news/measuring-model-persuasiveness>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Elliot, S. (2013). IntelliMetric: From here to validity. En M. D. Shermis y J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Lawrence Erlbaum Associates.
- Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education*, 9(1), e46885. <https://doi.org/10.2196/46885>
- Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge.

- Gliem, J. A. y Gliem, R. R. (2003). *Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales*. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M. y Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & computers*, 36(2), 193-202. <https://doi.org/10.3758/BF03195564>
- HAI (2023). *Artificial Intelligence Index Report 2023*. Human-Centered Artificial Intelligence.
- HAI (2024). *Artificial Intelligence Index Report 2024*. Human-Centered Artificial Intelligence.
- Hao, J. y Fauss, M. (2024). *Test security in remote testing age: Perspectives from process data analytics and AI* [Prepublicación]. arXiv. preprint arXiv:2411.13699
- Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z. y Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific reports*, 13, (1), 18617. <https://doi.org/10.1038/s41598-023-45644-9>
- Hicks, M. T., Humphries, J. y Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, 26(38), 1-26. <https://doi.org/10.1007/s10676-024-09775-5>
- Hidayanto, M.B., Lubis, M. y Jacob, D.W. (2025). Online assessment security through computer lockdown and human proctor methods. En. A. Nagar, D.S. Jat, D. Mishra y A. Joshi (Eds.), *Intelligent sustainable systems. Worlds4 2024. Lecture Notes in Networks and Systems*, (vol 1179, pp. 263-273). Springer. https://doi.org/10.1007/978-981-97-9327-3_22
- Hinkle, D. E., Wiersma, W. y Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (Vol. 663). Houghton Mifflin.
- Hussein, M. A., Hassan, H. y Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5, e208. <http://doi.org/10.7717/peerj-cs.208>
- Imran, M. y Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, 15(4), ep464. <https://doi.org/10.30935/cedtech/13605>
- İpek, Z. H., Gözümlü, A. I. C., Papadakis, S. y Kallogiannakis, M. (2023). Educational applications of the ChatGPT AI System: A systematic review research. *Educational Process: International Journal*, 12(3), 26-55.
- Karakose, T., Demirkol, M., Aslan, N., Köse, H., y Yirci, R. (2023). A conversation with ChatGPT about the impact of the COVID-19 pandemic on education: Comparative review based on human-AI collaboration. *Educational Process: International Journal*, 12(3), 7-25.

- Koo, T. K. y Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kuleva, M. y Miladinov, O. (2024). Exploring the efficacy of online proctoring in online examinations: A comprehensive review. En *Environment. Technologies. Resources. Proceedings of the International Scientific and Practical Conference* (Vol. 2, pp. 192-196). Rezekne Academy of Technologies. <https://doi.org/10.17770/etr2024vol2.8058>
- Kumar, V. y Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in education*, 5, 1-22. <https://doi.org/10.3389/feduc.2020.572367>
- Lanier, J. (20 de abril de 2023). There Is no A.I. *The New Yorker*. <https://www.newyorker.com/science/annals-of-artificial-intelligence/there-is-no-ai>
- Lanier, J. (01 de marzo de 2024). How to picture A.I. *The New Yorker*. <https://www.newyorker.com/science/annals-of-artificial-intelligence/how-to-picture-ai>
- Lee, V. R., Pope, D., Miles, S. y Zárate, R. C. (2024). Cheating in the age of generative AI: A high school survey study of cheating behaviors before and after the release of ChatGPT. *Computers and Education: Artificial Intelligence*, 7, (100253), 1-10. <https://doi.org/10.1016/j.caeai.2024.100253>
- Levine, S., Beck, S. W., Mah, C., Phalen, L. y Pittman, J. (2024). How do students use ChatGPT as a writing support? *Journal of Adolescent & Adult Literacy*, 1-13. <https://doi.org/10.1002/jaal.1373>
- Lillo-Fuentes, F., Venegas, R. y Lobos, I. (2023). Evaluación automatizada y semiautomatizada de la calidad de textos escritos: una revisión sistemática. *Perspectiva Educacional*, 62(2), 5-36. <https://doi.org/10.4151/07189729-Vol.62-Iss.2-Art.1420>
- Linacre, J. M. (1994). *Many-facet Rasch measurement*. Mesa Press.
- Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85-106. <https://acortar.link/RNUcC1>
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- López-Regalado, O., Núñez Rojas, N., López Gil, Ó. R. y Sánchez Rodríguez, J. (2024). Análisis del uso de la inteligencia artificial en la educación universitaria: una revisión sistemática. *Pixel-Bit. Revista de Medios y Educación*, (70), 97-122. <https://dx.doi.org/10.12795/pixelbit.106336>
- Lundgren, M. (2024). *Large language Models in student assessment: Comparing ChatGPT and human graders*. arXiv. <https://doi.org/10.48550/arXiv.2406.16510>

- Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the “originality” of students’ work. *Assessment & Evaluation in Higher Education*, 1-14. <https://doi.org/10.1080/02602938.2024.2309963>
- Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S. y Darwis, A. (2023). Exploring artificial intelligence in academic essay: Higher education students’ perspective. *International Journal of Educational Research Open*, 5, e100296. <https://doi.org/10.1016/j.ijedro.2023.100296>
- Martínez-Olmo, F. y González Catalán, F. (2024). Revisión sistemática de tendencias en la aplicación de la inteligencia artificial al ámbito de la escritura académica en las ciencias sociales. *Digital Education Review*, (45), 37-42. <https://doi.org/10.1344/der.2024.45.37-42>
- Menon, D. y Shilpa, K. (2023). “Chatting with ChatGPT”: Analyzing the factors influencing users' intention to use the Open AI's ChatGPT using the UTAUT model. *Heliyon*, 9(11), e20541. <https://doi.org/10.1016/j.heliyon.2023.e20541>
- Mizumoto, A. y Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Moorhouse, B. L. (2024). Beginning and first-year language teachers’ readiness for the generative AI age. *Computers and Education: Artificial Intelligence*, 6, e100201. <https://doi.org/10.1016/j.caeai.2024.100201>
- Naismith, B., Mulcaire, P. y Burstein, J. (2023, July). Automated evaluation of written discourse coherence using GPT-4. En E. Kochmar, J. Burstein, A. Horbach, R. Laarmann Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan y T. Zesch (Eds.). *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394-403). <https://doi.org/10.18653/v1/2023.bea-1.32>
- Nazar, R. y Renau, I. (2023). Estilector: un sistema de evaluación automática de la escritura académica en castellano. *Perspectiva Educacional. Formación de Profesores*, 62(2), 37-59. <http://dx.doi.org/10.4151/07189729-vol.62-iss.2-art.1427>
- Nguyen, T. N. y Truong, H. T. (2025). Trends and emerging themes in the effects of generative artificial intelligence in education: A systematic review. *Eurasia Journal of Mathematics, Science and Technology Education*, 21(4), em2613. <https://doi.org/10.29333/ejmste/16124>
- OpenAI (2024). Términos de uso. <https://openai.com/policies/row-terms-of-use/>
- Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O. y Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(8), em2307. <https://doi.org/10.29333/ejmste/13428>

- Paek, S. y Kim, N. (2021). Analysis of worldwide research trends on the impact of artificial intelligence in education. *Sustainability*, 13(14), 7941. <https://doi.org/10.3390/su13147941>
- Parker, J. L., Becker, K. y Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education*, 62(12), 721-727. <https://doi.org/10.3928/01484834-20231006-02>
- Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4.^a ed.). Sage publications. <https://acortar.link/wU2rWM>
- Pérez, A., McClain, S. K., Roa, A. F., Rosado-Mendinueta, N., Trigos-Carrillo, L., Robles, H. y Campo, O. (2025). Artificial Intelligence Applications in College Academic Writing and Composition: A Systematic Review. *Íkala, Revista de Lenguaje y Cultura*, 30(1), 1-37. <https://doi.org/10.17533/udea.ikala.355878>
- Pfau, A., Polio, C. y Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, 2(3), 100083. <https://doi.org/10.1016/j.rmal.2023.100083>
- Qin, Q. y Zhang, S. (2025). Visualizing the knowledge mapping of artificial intelligence in education: A systematic review. *Education and Information Technologies* 30, 449–483. <https://doi.org/10.1007/s10639-024-13076-1>
- Ramesh, D. y Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495-2527. <https://doi.org/10.1007/s10462-021-10068-2>
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Ed. Ampliada). University of Chicago Press. (Obra original publicada en 1960).
- Römer, U. y O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 6(2), 159-177. <https://doi.org/10.3366/cor.2011.0011>
- Ruano-Borbalan, J. C. (2025). The transformative impact of artificial intelligence on higher education: A critical reflection on current trends and futures directions. *International Journal of Chinese Education*, 14(1). <https://doi.org/10.1177/2212585X251319364>
- Rudner, L., Garcia, V. y Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *The Journal of Technology, Learning and Assessment*, 4, (4). <https://ejournals.bc.edu/index.php/jtla/article/view/1651>
- Rudolph, J., Tan, S. y Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of applied learning and teaching*, 6(1), 342-363. <https://doi.org/10.37074/jalt.2023.6.1.9>
- Sala, A. (2024). AI watermarking: A watershed for multimedia authenticity. *The UN Agency for Digital Technologies*. <https://www.itu.int/hub/2024/05/ai-watermarking-a-watershed-for-multimedia-authenticity/>

- Salas-Pilco, S. Z. y Yang, Y. (2022). Artificial intelligence applications in Latin American higher education: a systematic review. *International Journal of Educational Technology in Higher Education*, 19(1), 21. <https://doi.org/10.1186/s41239-022-00326-w>
- Seol, H. (2023). *seolmatrix: Correlations suite for jamovi* (Versión 3.7.1) [módulo de jamovi]. <https://github.com/hyunsooseol/seolmatrix>.
- Shi, H. y Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, 28(1), 771-795. <https://doi.org/10.1007/s10639-022-11200-7>
- Spataro, J. (2023). Introducing Microsoft 365 Copilot – your copilot for work. Microsoft Blog. <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/>
- Steele, J. L. (2023). To GPT or not GPT? Empowering our students to learn with AI. Computers and Education: *Artificial Intelligence*, 5, Artículo 100160. <https://doi.org/10.1016/j.caeai.2023.100160>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... y Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Su, Y., Lin, Y., y Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, 57, 100752. <https://doi.org/10.1016/j.asw.2023.100752>
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273-1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Tarp, S. y Nomdedeu-Rull, A. (2024). Who Has the last word? Lessons from Using ChatGPT to develop an AI-based Spanish writing assistant. *Círculo de Lingüística Aplicada a la Comunicación*, 97, 309-321. <https://doi.org/10.5209/clac.91985>
- The jamovi project (2022). Jamovi. (Versión 2.3) [Software]. <https://www.jamovi.org>.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T. Rozière B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. y Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv*. <https://doi.org/10.48550/arXiv.2302.13971>
- Townsend, B. (2023). Exploring the Impact of college rankings on student perceptions. *College Rover*. <https://collegerover.com/campus-library/news/1/exploring-the-impact-of-college-rankings-on-student-perceptions>
- Tülübaşı, T., Demirkol, M., Ozdemir, T. Y., Polat, H., Karakose, T. y Yirci, R. (2023). An interview with ChatGPT on emergency remote teaching: A comparative

- analysis based on human–AI collaboration. *Educational Process: International Journal*, 12(2), 93-110. <https://doi.org/10.22521/edupij.2023.122.6>
- Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T. y Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252, 124167. <https://doi.org/10.1016/j.eswa.2024.124167>
- Wise, B., Emerson, L., Van Luyn, A., Dyson, B., Bjork, C. y Thomas, S. E. (2024). A scholarly dialogue: Writing scholarship, authorship, academic integrity and the challenges of AI. *Higher Education Research & Development*, 43(3), 578–590. <https://doi.org/10.1080/07294360.2023.2280195>
- Yeadon, W., Agra, E., Inyang, O. O., Mackay, P. y Mizouri, A. (2024). Evaluating AI and human authorship quality in academic writing through physics essays. *arXiv*. <https://doi.org/10.48550/arXiv.2403.05458>
- Zaheer, M., Munir, S. y Sherazi, S. N. (2024). Exploring the Issues and challenges of online assessment and evaluation in the era of artificial intelligence. *Journal of Asian Development Studies*, 13(1), 185-197. <https://doi.org/10.62345/jads.2024.13.1.16>
- Zhang, S. (2021). Review of automated writing evaluation systems. *Journal of China Computer-Assisted Language Learning*, 1(1), 170-176. <https://doi.org/10.1515/jccall-2021-2007>
- Zirar, A. (2023). Exploring the impact of language models, such as ChatGPT, on student learning and assessment. *Review of Education*, 11(3), e3433. <https://doi.org/10.1002/rev3.3433>
- Zupanc, K. y Bosnic, Z. (2015). Advances in the field of automated essay evaluation. *Informatica*, 39(4), 383-395. <https://informatica.si/index.php/informatica/article/view/815>