Huerta, P. y Silva Fuentealba, E. (2026). Desempeño de una herramienta basada en IA para la corrección del TTCT-Verbal en la evaluación del pensamiento creativo. *Revista de Investigación Educativa*, 44. https://doi.org/10.6018/rie.710621

Desempeño de una herramienta basada en IA para la corrección del TTCT-Verbal en la evaluación del pensamiento creativo

Performance of an AI-Based Tool for Scoring the TTCT-Verbal in the Assessment of Creative Thinking

Ricard Huerta[1]* y Eduardo Silva Fuentealba**

*Instituto de Creatividad e Innovaciones Educativas. Universitat de València (España)
**Doctorando en Educación. Universidad Bernardo O'Higgins (Chile)

Resumen

La corrección del Test de Pensamiento Creativo de Torrance – Forma Verbal (TTCT-Verbal) presenta desafíos asociados a la carga temporal, la variabilidad interjueces y la subjetividad en la evaluación de la originalidad. Este estudio examinó de manera preliminar el desempeño de una herramienta de inteligencia artificial basada en un modelo de lenguaje de gran escala para apoyar la corrección de las dimensiones de fluidez, flexibilidad y originalidad, en comparación con un panel de seis evaluadores/as humanos/as. Se utilizaron 47 protocolos de estudiantes de postgrado de la Universitat de València, de los cuales 30 fueron seleccionados para el análisis comparativo entre la IA y el promedio humano. Se calcularon el Coeficiente de Correlación Intraclase ICC(3,1), el error absoluto medio, la raíz del error cuadrático medio, correlaciones de Pearson y gráficos de Bland-Altman. Los resultados mostraron medias próximas entre ambos sistemas y bajo error absoluto. La asociación más alta se observó en originalidad, mientras que fluidez y flexibilidad evidenciaron menor consistencia relativa. Los análisis gráficos permitieron explorar la distribución de las diferencias sin identificar sesgos sistemáticos evidentes, aunque estos resultados deben interpretarse de forma descriptiva. Se concluye que la herramienta muestra un desempeño prometedor, especialmente en originalidad, pero no homogéneo entre dimensiones. Su uso resulta más pertinente como apoyo complementario en sistemas híbridos de evaluación del pensamiento creativo que como sustituto del juicio humano experto.

1 **Correspondencia:** Ricard Huerta, Universitat de València, Ricard.Huerta@uv.es

Palabras clave: creatividad; Inteligencia Artificial; TTCT-Verbal; ingeniería de prompts.

Abstract

Considering the specific challenges related to the assessment of the Torrance Tests of Creative Thinking – Verbal Form (TTCT-Verbal) concerning time demands, inter-rater variability, and subjectivity in the assessment of originality, this study preliminarily examines the performance of an artificial intelligence tool based on a large language model to support the scoring of fluency, flexibility, and originality, contrasting results with those of a panel of six human evaluators. A total of 47 protocols from postgraduate students at the Universitat de València were used, of which 30 were selected for the comparative analysis between the AI system and the human average. Similarly, the Intraclass Correlation Coefficient ICC (3,1), mean absolute error, root mean square error, Pearson correlations, and Bland-Altman plots were calculated. Results indicate close mean scores between both systems and low absolute error, observing the strongest association in originality, while fluency and flexibility showed lower relative consistency. Graphical analyses allowed the distribution of differences to be explored without identifying evident systematic bias, although these results should be interpreted descriptively. As such, it is concluded that the tool shows promising, but dimension-dependent performance, especially with regards to originality. Its use appears more appropriate as a complementary support within hybrid systems for assessing creative thinking than as a replacement for expert human judgment.

Keywords: creativity; Artificial Intelligence; TTCT-Verbal; prompt engineering.

Introducción

En el campo educativo contemporáneo, el interés por medir y desarrollar procesos cognitivos complejos como la creatividad se ha intensificado junto con metodologías centradas en el estudiante y orientadas al desarrollo de competencias (Huerta y Alfonso-Benlliure, 2025). Los enfoques pedagógicos activos favorecen el pensamiento crítico, reflexivo y creativo, al promover experiencias de construcción del conocimiento y resolución de problemas abiertos (Alcántara Santuario, 2023; Cesário y Nisi, 2023; Huerta, 2023). En este contexto, la creatividad se ha consolidado como una competencia clave, incrementando la necesidad de instrumentos válidos y fiables para su evaluación (Jarquín-Ramírez et al., 2025; Sanz-Leal y Orozco Gómez, 2025).

La evaluación de la creatividad verbal mediante instrumentos estandarizados es relevante en psicología educativa, identificación de talento y estudio del pensamiento creativo. El Test de Pensamiento Creativo de Torrance – Forma Verbal (TTCT-Verbal) se ha consolidado como instrumento de referencia internacional, sustentado en el modelo de estructura del intelecto de J. P. Guilford (1950, 1967) y en una amplia validación psicométrica (Torrance, 1966, 1974; Kim, 2006). Este instrumento evalúa fluidez, flexibilidad y originalidad mediante tareas verbales abiertas, y ha mostrado capacidad predictiva de logros creativos a largo plazo (Runco et al., 2010).

Sin embargo, su corrección manual presenta limitaciones: alta carga temporal, necesidad de jueces/zas entrenados/as y variabilidad interjueces/zas, especialmente

en originalidad, por su carácter interpretativo y dependiente de criterios de rareza estadística (Torrance, 2008; Kim, 2006; Acar et al., 2023). Estas dificultades restringen su uso en investigaciones amplias, programas educativos masivos y evaluaciones sistemáticas, aumentando costos y reduciendo accesibilidad (Welter et al., 2016). Por ello, los mecanismos automatizados de corrección han comenzado a recibir creciente atención científica (Alfonso-Benlliure et al., 2025).

Paralelamente, los sistemas de inteligencia artificial basados en modelos de lenguaje de gran escala (Large Language Models, LLMs) están transformando los entornos educativos y las formas de interacción con el conocimiento (Goenechea y Valero-Franco, 2024). Estas tecnologías posibilitan apoyo al aprendizaje, análisis automatizado de información y retroalimentación personalizada, ampliando el potencial de las herramientas digitales en educación superior (Mella-Mella et al., 2026). Asimismo, el uso pedagógico de tecnologías digitales puede facilitar el acceso a recursos, fortalecer la motivación y promover aprendizajes más autónomos y colaborativos (Sanjurjo Pérez et al., 2026; Verdú-Pina et al., 2026). En este marco, los LLMs ofrecen posibilidades para automatizar procesos evaluativos complejos.

Desde una perspectiva técnica, los LLMs pueden procesar tareas semánticas complejas en tiempos reducidos, operar de manera escalable y ofrecer cierta estabilidad bajo condiciones delimitadas de configuración (Acar et al., 2023; Guzik et al., 2023). Estudios recientes han explorado su uso en evaluación de la creatividad y puntuación automatizada de respuestas abiertas, reportando asociaciones relevantes con juicios humanos en originalidad mediante *text-mining*, *prompt engineering* o *fine-tuning* (Acar et al., 2023; Dumas y Runco, 2018; Perchtold-Stefan et al., 2024). No obstante, persisten interrogantes sobre su fiabilidad psicométrica en instrumentos estandarizados como el TTCT-Verbal, especialmente respecto del acuerdo con evaluadores/as humanos/as y la estabilidad en dimensiones cuantitativas —fluidez y flexibilidad— frente a dimensiones más interpretativas, como originalidad (Huerta, 2025).

En este contexto, el presente estudio evalúa el grado de aproximación entre una herramienta basada en un modelo de lenguaje de gran escala, configurada mediante instrucciones delimitadas, y el juicio de evaluadores y evaluadoras humanas expertos en la corrección del TTCT-Verbal. Mediante ICC, métricas de error, correlaciones lineales y gráficos de Bland-Altman, se busca aportar evidencia sobre sus posibilidades y limitaciones como apoyo complementario en sistemas híbridos de evaluación del pensamiento creativo, considerando su potencial para reducir barreras prácticas de corrección y la necesidad de resguardar transparencia, prudencia interpretativa y supervisión experta en el uso de IA en evaluación educativa (Barragán-Giraldo et al., 2024).

Marco teórico

La creatividad y el pensamiento divergente como constructos cognitivos

La creatividad ha sido definida como la capacidad de generar ideas, productos o soluciones novedosos y apropiados dentro de un contexto específico (Runco, 2023). Desde mediados del siglo XX, este constructo ha sido ampliamente desarrollado desde una perspectiva cognitiva, especialmente a partir de los trabajos de J. P. Guilford, quien

distinguió entre pensamiento convergente y pensamiento divergente. El pensamiento convergente se orienta a resolver problemas con una única respuesta correcta mediante procesos lógicos y secuenciales (Guilford, 1967). En contraste, el pensamiento divergente implica la generación fluida, exploratoria y no lineal de múltiples ideas o soluciones ante un mismo estímulo, constituyendo un núcleo operativo del comportamiento creativo (Guilford, 1950, 1967). Guilford propuso cuatro dimensiones principales del pensamiento divergente: fluidez, entendida como cantidad de ideas generadas; flexibilidad, referida a la diversidad de categorías conceptuales utilizadas; originalidad, asociada a la rareza estadística o inusualidad de las respuestas; y elaboración, vinculada al grado de desarrollo o detalle de las ideas. Estas dimensiones han sido ampliamente empleadas como indicadores empíricos de creatividad cognitiva en educación e investigación. Posteriormente, Ellis Paul Torrance (1966, 1974) operacionalizó estas propuestas mediante pruebas estandarizadas para medir el pensamiento divergente, consolidándolo como una medida empírica de la creatividad. Su evaluación es relevante por su asociación con logros académicos, profesionales e innovadores (Runco et al., 2010). Sin embargo, presenta desafíos metodológicos vinculados a la subjetividad del juicio evaluativo, el tiempo requerido para una corrección rigurosa y la variabilidad interjueces que puede comprometer la consistencia psicométrica (Torrance, 2008; Kim, 2006).

El Test de Pensamiento Creativo de Torrance (TTCT-Verbal) como instrumento de referencia

El Test de Pensamiento Creativo de Torrance en su forma verbal (TTCT-Verbal), desarrollado por E. Paul Torrance en la década de 1960, es uno de los instrumentos más difundidos y validados internacionalmente para evaluar el pensamiento creativo verbal. Basado en la teoría de la estructura del intelecto de Guilford, fue concebido inicialmente como parte de los *Minnesota Tests of Creative Thinking*. Desde su primera versión formal, publicada en 1966, ha sido objeto de diversas revisiones. Puede aplicarse desde el primer grado escolar hasta la adultez, cuenta con formas paralelas A y B para reducir el efecto de aprendizaje, ha sido adaptado a más de 35 idiomas y dispone de evidencia de validez transcultural. La forma verbal estándar incluye seis tareas: *Asking*, *Guessing Causes*, *Guessing Consequences*, *Product Improvement*, *Unusual Uses* y *Just Suppose*. Estas se administran en aproximadamente 45 minutos mediante estímulos verbales abiertos que permiten generar múltiples respuestas creativas (Torrance, 2008). El TTCT-Verbal evalúa tres dimensiones centrales: fluidez, referida al número de respuestas relevantes, interpretables y no redundantes; flexibilidad, vinculada a la diversidad de categorías conceptuales empleadas; y originalidad, asociada a la rareza estadística o infrecuencia de las respuestas según normas establecidas. Esta última es la dimensión más compleja, pues exige criterios normativos y juicio experto. Desde una perspectiva psicométrica, el TTCT-Verbal ha mostrado alta fiabilidad test-retest, con coeficientes entre 0.80 y 0.90, y consistencia interna aceptable (Torrance, 2008). Además, estudios longitudinales han evidenciado su validez predictiva al asociar puntuaciones tempranas con logros creativos hasta cinco décadas después (Runco et al., 2010). No obstante, aunque la fiabilidad interjueces puede ser alta con evaluadores entrenados, tiende a disminuir en originalidad cuando no existe capacitación exhaustiva (Kim, 2006).

Limitaciones de la corrección humana del TTCT-Verbal

La corrección manual del TTCT-Verbal implica una elevada carga cognitiva y temporal. Cada protocolo puede requerir entre 20 y 45 minutos de evaluación por parte de cada juez/a, dependiendo del grado de complejidad de las respuestas y de la dimensión evaluada. Esta carga se intensifica en la dimensión de originalidad, que exige juicios cualitativos con alta carga interpretativa (Torrance, 2008; Acar et al., 2023; Organisciak et al., 2023). Aunque diversos estudios reportan coeficientes de fiabilidad interjueces entre 0.80 y 0.95 cuando los evaluadores han sido entrenados adecuadamente, esta fiabilidad puede descender en contextos donde tal formación no es intensiva, especialmente en la dimensión de originalidad, donde los valores pueden oscilar entre 0.70 y 0.85 (Kim, 2006). Estas variaciones evidencian la sensibilidad del proceso evaluativo a factores contextuales y formativos, lo que introduce un margen de incertidumbre en la consistencia de las puntuaciones. En conjunto, estas limitaciones no solo implican un alto coste económico y una considerable inversión de tiempo, sino que también reducen la viabilidad del uso del TTCT en investigaciones a gran escala o en entornos educativos masivos (Welter et al., 2016). En este sentido, la dependencia de evaluadores y evaluadoras altamente capacitadas y la variabilidad inherente al juicio humano plantean un desafío metodológico relevante, vinculado a la necesidad de explorar alternativas que permitan mantener los estándares psicométricos del instrumento, optimizando al mismo tiempo la eficiencia y consistencia del proceso de evaluación.

El potencial de los modelos de lenguaje de gran escala en la evaluación del pensamiento creativo

Los modelos de lenguaje de gran escala (LLMs) han mostrado capacidad para procesar tareas semánticas complejas, lo que ha impulsado su exploración en evaluación psicológica y educativa. Entre sus ventajas potenciales se encuentran la estabilidad operativa bajo configuraciones controladas, la escalabilidad y la reducción del tiempo requerido para corregir respuestas abiertas (Acar et al., 2023). Estos modelos pueden disminuir problemas asociados a la corrección humana, como la fatiga o la variabilidad intraevaluador/a; sin embargo, no eliminan el sesgo ni garantizan mayor objetividad por sí mismos. Su desempeño depende de los datos de entrenamiento, el diseño del *prompt*, las condiciones de uso y los parámetros disponibles en cada entorno tecnológico. Por ello, su aplicación evaluativa requiere procedimientos delimitados e ingeniería de *prompts* rigurosa, orientada por criterios explícitos de puntuación (Guzik et al., 2023).

Diversos estudios han aplicado métodos computacionales a la corrección automática en creatividad, especialmente en originalidad, mediante distancia semántica, minería de texto y LLMs. Estos trabajos reportan correlaciones relevantes con evaluaciones humanas y evidencian su potencial para estimar componentes específicos del pensamiento creativo. No obstante, subrayan la necesidad de distinguir entre sistemas basados en distancia semántica, minería de texto, entrenamiento supervisado y LLMs, dado que operan bajo supuestos metodológicos distintos (Beaty et al., 2022; Acar et al.,

2023; Organisciak et al., 2023). En conjunto, la literatura reciente muestra avances en la evaluación automatizada de la originalidad y en el uso pedagógico de herramientas generativas para estimular fluidez, flexibilidad y originalidad. Sin embargo, también indica que el desempeño de la IA varía según la dimensión analizada, por lo que es necesario examinar cada componente del pensamiento creativo de manera diferenciada y evitar generalizaciones sobre la fiabilidad global de los sistemas automatizados de corrección (Beaty et al., 2022; Acar et al., 2023; Organisciak et al., 2023; Silva-Fuentealba et al., 2024; Silva-Fuentealba et al., 2025).

Objetivos de la investigación

La evaluación del pensamiento creativo mediante el Test de Pensamiento Creativo de Torrance – Forma Verbal (TTCT-Verbal) es relevante en investigación y práctica educativa, pues permite valorar fluidez, flexibilidad y originalidad. Sin embargo, su aplicación se ve limitada por la subjetividad de dimensiones interpretativas, especialmente originalidad, la variabilidad interjueces y el alto costo temporal y económico de la corrección manual por expertos y expertas (Acar et al., 2023; Kim, 2006; Torrance, 2008; Reiter-Palmon et al., 2019). Estas limitaciones justifican explorar herramientas tecnológicas que apoyen la corrección sin sustituir la supervisión experta. En este contexto, el estudio examina el desempeño de una herramienta basada en IA, específicamente un modelo de lenguaje de gran escala configurado mediante instrucciones delimitadas, como apoyo complementario para la corrección del TTCT-Verbal.

Objetivo general

Evaluar el grado de aproximación entre las puntuaciones generadas por una herramienta de inteligencia artificial y las asignadas por evaluadores/as humanos/as expertos/as en la corrección del TTCT-Verbal.

Objetivos específicos

Analizar la consistencia relativa entre las puntuaciones generadas por la IA y el promedio de las evaluaciones humanas en fluidez, flexibilidad y originalidad, mediante el Coeficiente de Correlación Intraclase ICC(3,1) y métricas complementarias de error y asociación (Koo y Li, 2016).

1. Examinar el desempeño diferenciado de la herramienta según cada dimensión evaluada por el TTCT-Verbal, considerando la mayor variabilidad humana reportada en originalidad por su carácter subjetivo y dependiente de criterios de rareza (Kim, 2006; Torrance, 2008).
2. Comparar la eficiencia temporal del sistema automatizado frente a la corrección humana, considerando el tiempo de procesamiento de los protocolos y la necesidad de supervisión experta, dado que la corrección manual del TTCT-Verbal suele demandar entre 20 y 45 minutos por protocolo (Acar et al., 2023; Perchtold-Stefan et al., 2024; Reiter-Palmon et al., 2019).

En conjunto, estos objetivos buscan aportar evidencia preliminar sobre el desempeño de una herramienta de IA en la corrección del TTCT-Verbal, considerando tanto su grado de aproximación al juicio humano como sus límites para reproducir de manera homogénea las distintas dimensiones del pensamiento creativo.

Método

Se llevó a cabo un estudio cuantitativo de tipo instrumental, orientado a estimar el grado de consistencia y aproximación entre las puntuaciones generadas por una herramienta de inteligencia artificial y las puntuaciones asignadas por la evaluación humana experta en la corrección del Test de Pensamiento Creativo de Torrance, Forma Verbal (TTCT-Verbal). El análisis se centró en tres dimensiones del pensamiento creativo evaluadas por el instrumento: fluidez, flexibilidad y originalidad.

Participantes y protocolos

La muestra fue obtenida en el contexto de la Universitat de València, con estudiantes matriculados/as en un programa de postgrado profesionalizante orientado a la formación de futuros/as docentes de secundaria. La aplicación del test se realizó en el marco de una cátedra universitaria centrada en el desarrollo del pensamiento creativo, vinculada a experiencias de educación artística, museos y mediación cultural (Huerta, 2023; Huerta y Alfonso-Benlliure, 2025; Cisternas San Martín et al., 2025).

Se utilizaron 47 protocolos de respuesta del TTCT-Verbal. De este conjunto, 30 protocolos, correspondientes al 64% del universo total, fueron seleccionados aleatoriamente para la comparación entre la herramienta de inteligencia artificial y los/as evaluadores/as humanos/as. Los 47 protocolos fueron procesados por la herramienta automatizada, mientras que los 30 seleccionados fueron corregidos también por el panel de evaluadores/as humanos/as, permitiendo establecer la comparación entre ambos sistemas de puntuación.

Evaluadores humanos

Participaron seis jueces y juezas expertas, con una experiencia promedio de 13.6 años en áreas como educación, filosofía, comunicación social y ciberseguridad e informática. El grupo estuvo conformado por tres hombres y tres mujeres, lo que permitió contar con una participación equilibrada.

Antes de iniciar el proceso de evaluación, todos los y las juezas recibieron una capacitación común en los criterios de corrección del TTCT-Verbal, con el propósito de asegurar una comprensión compartida sobre cómo interpretar y puntuar las respuestas de los y las participantes. Cada evaluador revisó los protocolos de manera independiente, siguiendo rúbricas adaptadas a las dimensiones de fluidez, flexibilidad y originalidad. Para ello, dispusieron de un plazo de diez días, durante los cuales entregaron tanto sus puntuaciones como observaciones cualitativas sobre los protocolos analizados.

Instrumento: TTCT-Verbal

El instrumento utilizado fue el Test de Pensamiento Creativo de Torrance – Forma Verbal (TTCT-Verbal), empleado para evaluar dimensiones del pensamiento creativo mediante tareas abiertas de producción verbal. En este estudio se consideraron tres dimensiones: fluidez, flexibilidad y originalidad. Esta selección se fundamenta en la estructura de puntuación del TTCT-Verbal descrita en la literatura especializada, donde sus componentes principales corresponden a fluidez, originalidad y flexibilidad, mientras que la elaboración forma parte de las subescalas normativas del TTCT-Figural junto con abstracción de títulos y resistencia al cierre prematuro (Kim, 2017).

La fluidez corresponde al número de respuestas relevantes, interpretables y no redundantes producidas por cada participante. La flexibilidad refiere a la diversidad de categorías conceptuales presentes en las respuestas, es decir, a la capacidad para cambiar de perspectiva o utilizar distintos marcos de referencia. La originalidad se relaciona con el grado de rareza, singularidad o infrecuencia de las ideas producidas en comparación con el corpus de respuestas analizado.

En el presente estudio, la originalidad fue evaluada a nivel de idea dentro de cada protocolo. Cada idea recibió 2 puntos cuando fue considerada altamente original, es decir, cuando no aparecía en otros protocolos del corpus total analizado; 1 punto cuando aparecía en algunos protocolos, pero no en la mayoría; y 0 puntos cuando correspondía a una respuesta frecuente o recurrente en el corpus. Posteriormente, los puntajes asignados a las distintas ideas fueron sumados para obtener una puntuación total de originalidad por protocolo. En consecuencia, aunque cada idea individual podía recibir entre 0 y 2 puntos, la puntuación total de originalidad de un protocolo podía superar ese rango.

Sistema automatizado de corrección

La herramienta de inteligencia artificial se implementó utilizando el modelo Gemini 1.5 Pro (Google), configurado mediante un *prompt* de sistema estructurado y ejecutado en un entorno conversacional web. La especificación completa del sistema, incluyendo reglas de puntuación, criterios de análisis y estructura del *prompt*, se detalla en el Anexo A, con el propósito de favorecer la transparencia y replicabilidad del procedimiento.

Las instrucciones estructuradas delimitaron el análisis en tres dimensiones:

- Fluidez (F): descomposición de frases en unidades mínimas de información, siguiendo el criterio de fluidez atómica.
- Flexibilidad (FX): clasificación de las respuestas en cinco bloques conceptuales: físico, tecnológico, utilitario, social y artístico.
- Originalidad (O): asignación de puntajes a nivel de idea, considerando la recurrencia o singularidad de cada respuesta en relación con el corpus total analizado.

El procesamiento de los protocolos mediante la herramienta de inteligencia artificial se realizó en bloques de 10 casos como estrategia operativa para facilitar el ingreso de las respuestas al entorno conversacional y mantener la trazabilidad del análisis. No obstante, la estimación de la originalidad no se realizó de manera independiente dentro de cada bloque, sino considerando el corpus total de protocolos analizados. Es decir,

la recurrencia o singularidad de cada idea fue valorada a partir de la comparación con el conjunto completo de respuestas procesadas.

Procedimiento

Los y las juezas humanas evaluaron los 30 protocolos seleccionados de forma ciega e independiente. El tiempo promedio de corrección humana fue de aproximadamente 22 minutos por cada conjunto de cinco protocolos, lo que equivale a cerca de 4–5 minutos por protocolo.

La herramienta de inteligencia artificial procesó el total de los 47 protocolos en bloques de 10 casos. Posteriormente, se seleccionaron para el análisis comparativo los 30 protocolos que también habían sido revisados por los y las evaluadoras humanas. La salida generada por el sistema fue estructurada por dimensiones y puntuaciones, permitiendo comparar las puntuaciones automatizadas con el promedio de las y los seis evaluadores humanos en fluidez, flexibilidad y originalidad.

Análisis de datos

Para evaluar la consistencia y el grado de aproximación entre las puntuaciones generadas por la herramienta de inteligencia artificial y el promedio de los y las seis evaluadoras humanas, se empleó el Coeficiente de Correlación Intraclase (ICC), siguiendo las recomendaciones de Koo y Li (2016). Este coeficiente es adecuado para variables continuas y permite estimar la consistencia entre sistemas de evaluación. La especificación del sistema automatizado se presenta en el Anexo A, con el fin de favorecer la transparencia y replicabilidad del procedimiento.

Se utilizó el modelo ICC(3,1), correspondiente a un diseño bidireccional de efectos mixtos con evaluadores fijos y orientado a la consistencia. Esta elección responde a que los/as evaluadores/as humanos/as fueron seleccionados por su experiencia profesional y a que el objetivo fue analizar en qué medida la herramienta automatizada reproduce el patrón de puntuaciones del juicio humano experto, más que establecer equivalencia exacta entre cada puntuación humana e IA.

El ICC se calculó comparando las puntuaciones de la IA con el promedio humano en las tres dimensiones del TTCT-Verbal: fluidez, flexibilidad y originalidad. Su interpretación siguió los criterios de Koo y Li (2016): valores inferiores a 0.50 indican baja consistencia relativa; entre 0.50 y 0.75, consistencia moderada; entre 0.75 y 0.90, consistencia buena; y superiores a 0.90, consistencia excelente.

Como análisis complementarios, se calcularon el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE) y correlaciones de Pearson. El MAE y el RMSE permitieron estimar la magnitud de las discrepancias entre la IA y el promedio humano, mientras que Pearson permitió analizar la asociación lineal entre ambos sistemas.

Finalmente, se aplicó el método gráfico de Bland-Altman para examinar la distribución de las diferencias, identificar posibles sesgos sistemáticos y observar el comportamiento del error a lo largo del rango de medición. Dado el tamaño muestral, este análisis se utilizó como complemento descriptivo y visual de las métricas cuantitativas, no como evidencia concluyente de estabilidad psicométrica del sistema automatizado.

Resultados

Se emplearon medidas de acuerdo absoluto (MAE, RMSE, sesgo), consistencia (correlación de Pearson) y análisis gráfico de concordancia (Bland–Altman) para evaluar el desempeño del sistema automatizado en las dimensiones de fluidez (F), flexibilidad (FX) y originalidad (O)

Para interpretar la dimensión de originalidad, se debe precisar que la escala de 0 a 2 puntos fue aplicada a cada idea individual dentro del protocolo, no a la puntuación final del protocolo. La puntuación de originalidad reportada en la Tabla 1 corresponde a la suma de los puntajes asignados a las distintas ideas de cada participante. Por ello, las medias de 5.87 en evaluadores/as humanos/as y 5.90 en IA no deben interpretarse como valores dentro de una escala final de 0 a 2, sino como puntuaciones agregadas por protocolo.

Estadísticos descriptivos

La Tabla 1 resume las medias, desviaciones estándar de las puntuaciones y el sesgo humano-IA.

Tabla 1

Estadísticos descriptivos de las puntuaciones (n = 30)

Dimensión	Media Humanos (DE)	Media IA (DE)	Sesgo (Humano – IA)
Fluidez (F)	6.83 (1.02)	6.60 (0.56)	+0.23
Flexibilidad (FX)	4.00 (0.69)	4.10 (0.55)	–0.10
Originalidad (O)	5.87 (1.91)	5.90 (1.09)	–0.03

Nota. Las medias entre ambos sistemas son prácticamente equivalentes en las tres dimensiones, con diferencias inferiores a 0.25 puntos, lo que sugiere ausencia de sesgo sistemático relevante en las estimaciones de la IA respecto del promedio humano.

Concordancia entre IA y promedio de evaluadores humanos

Se calcularon métricas de acuerdo absoluto y consistencia para evaluar la relación entre las puntuaciones de la IA y el promedio de los y las evaluadoras humanas.

Tabla 2

Métricas de concordancia IA vs. promedio humanos (n = 30)

Dimensión	MAE	RMSE	r de Pearson (p)
Fluidez (F)	0.83	1.10	0.120 (0.527)
Flexibilidad (FX)	0.43	0.75	0.272 (0.146)
Originalidad (O)	1.10	1.33	0.721 (< 0.001)

El error absoluto medio (MAE) y la raíz del error cuadrático medio (RMSE) indican desviaciones bajas entre ambos sistemas, particularmente en la dimensión de flexibilidad (MAE = 0.43), lo que sugiere un alto grado de proximidad en términos absolutos. En fluidez y originalidad, los errores se mantienen en rangos moderados, sin evidenciar discrepancias sistemáticas relevantes.

Las correlaciones de Pearson muestran un comportamiento diferenciado por dimensión. En originalidad se observa una asociación fuerte ($r = 0.721$, $p < 0.001$), lo que indica que la IA captura el patrón general de variación del juicio humano en esta dimensión. En contraste, en fluidez ($r = 0.120$, $p = 0.527$) y flexibilidad ($r = 0.272$, $p = 0.146$) las correlaciones son bajas, lo que sugiere limitaciones en la capacidad del sistema para replicar las diferencias relativas entre sujetos en estas dimensiones.

Para complementar las métricas anteriores, se calculó el ICC(3,1) entre las puntuaciones de la IA y el promedio humano. La Tabla 3 presenta los valores por dimensión, con sus intervalos de confianza al 95% e interpretación.

Tabla 3

Coeficiente de Correlación Intraclase entre IA y promedio humano por dimensión

Dimensión	ICC(3,1)	IC 95%	Interpretación
Fluidez (F)	0.11	[-0.25, 0.45]	Baja consistencia relativa
Flexibilidad (FX)	0.26	[-0.10, 0.57]	Baja consistencia relativa
Originalidad (O)	0.62	[0.34, 0.80]	Consistencia relativa moderada

El análisis mediante ICC(3,1) expuesto en la tabla 3, mostró un desempeño diferenciado por dimensión. Fluidez y flexibilidad presentaron baja consistencia relativa, mientras que originalidad alcanzó un nivel moderado, sugiriendo una mayor aproximación de la IA al patrón de puntuaciones humanas en la dimensión más interpretativa del TTCT-Verbal.

En conjunto, estos resultados indican que, si bien la IA reproduce adecuadamente las medias y presenta bajo error absoluto, su capacidad para reflejar variaciones individuales es desigual según la dimensión evaluada.

En relación con la eficiencia temporal, la corrección humana requirió aproximadamente 22 minutos por cada conjunto de cinco protocolos, equivalente a cerca de 4–5 minutos por protocolo. En contraste, la herramienta de inteligencia artificial procesó los 47 protocolos en un tiempo considerablemente menor, una vez configurado el prompt y organizado el ingreso de los casos por bloques. Esta diferencia sugiere una reducción sustantiva de la carga temporal asociada al proceso de corrección, aunque debe interpretarse con cautela, dado que el tiempo de configuración, revisión y control humano del procedimiento automatizado también forma parte del proceso evaluativo.

Análisis gráfico de acuerdo (Bland-Altman)

Se elaboraron gráficos de Bland-Altman para cada dimensión, representando la diferencia entre las puntuaciones del promedio humano y la IA frente a la media de

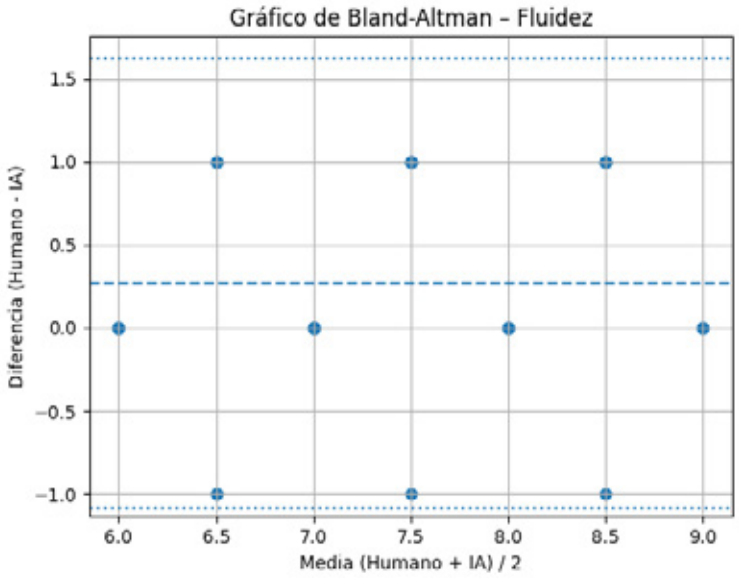
ambos sistemas de puntuación (Bland y Altman, 1986, 1999; Giavarina, 2015). Este análisis se utilizó como complemento descriptivo de las métricas cuantitativas, con el propósito de explorar la distribución de las discrepancias entre ambos métodos de corrección.

En las tres dimensiones, el sesgo medio fue cercano a cero, con valores entre –0.10 y +0.23. Asimismo, entre el 93% y el 97% de las observaciones se ubicó dentro de los límites de acuerdo, calculados como ±1.96 desviaciones estándar de las diferencias. Visualmente, no se observaron patrones marcados de heterocedasticidad, aunque este resultado debe interpretarse con cautela debido al tamaño muestral del estudio.

En la Figura 1 se presenta el análisis de Bland-Altman correspondiente a la dimensión de fluidez. Cada punto representa la diferencia entre la puntuación del promedio humano y la puntuación de la IA en función de la media de ambas mediciones.

Figura 1

Gráfico de Bland-Altman para la dimensión de fluidez.

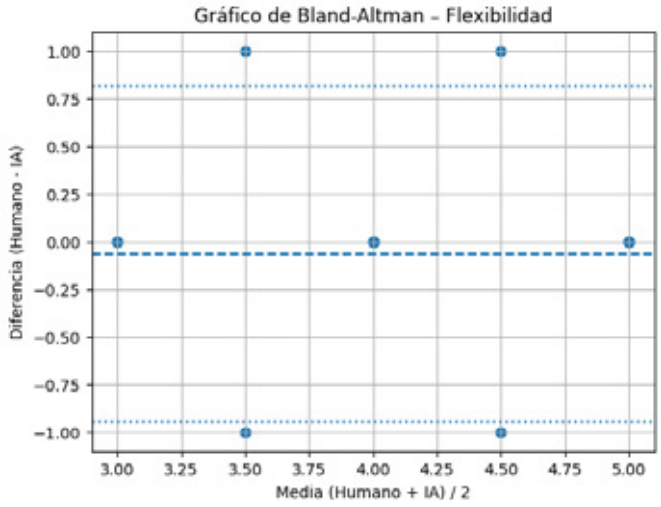


Nota. Elaboración propia con datos del TTCT

En la Figura 2 se muestra el análisis de Bland-Altman correspondiente a la dimensión de flexibilidad, en el que se examina la distribución de las discrepancias entre el promedio humano y la IA respecto de la media de sus puntuaciones.

Figura 2

Gráfico de Bland–Altman para la dimensión de flexibilidad.

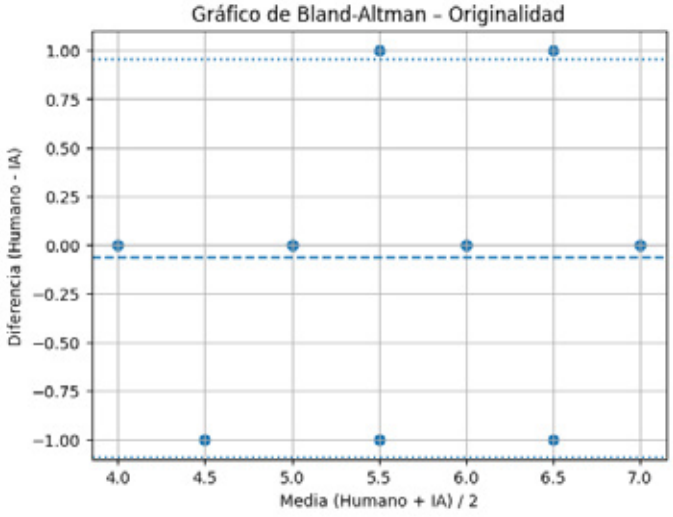


N*ota.* Elaboración propia con datos del TTCT.

En la Figura 3 se expone el análisis de Bland-Altman para la dimensión de originalidad, evidenciando la relación entre las diferencias de puntuación y la media de las evaluaciones humanas y de la inteligencia artificial.

Figura 3

Gráfico de Bland–Altman para la dimensión de originalidad.



Nota. Elaboración propia con datos del TTCT

En conjunto, los gráficos de Bland-Altman sugieren que las diferencias entre la IA y el promedio humano se distribuyen sin sesgos sistemáticos evidentes en las tres dimensiones analizadas. En términos metodológicos, estos hallazgos son consistentes con los criterios de interpretación del método de Bland-Altman, donde el interés radica no en la coincidencia exacta entre mediciones, sino en la magnitud y distribución de las discrepancias dentro de un rango aceptable (Bland y Altman, 1986, 1999; Giavarina, 2015). No obstante, estos resultados deben entenderse como evidencia descriptiva complementaria, y no como una demostración concluyente de estabilidad psicométrica del sistema automatizado a lo largo de todo el rango de puntuaciones.

Estabilidad y consistencia relativa

En síntesis, los resultados muestran que la herramienta de inteligencia artificial presenta bajo error absoluto y ausencia de sesgo sistemático relevante a nivel agregado. No obstante, su capacidad para reproducir las diferencias relativas entre protocolos varía según la dimensión evaluada. La mayor aproximación al patrón de puntuaciones humanas se observó en originalidad, mientras que fluidez y flexibilidad evidenciaron menor consistencia relativa. Por tanto, el desempeño del sistema automatizado debe interpretarse de manera diferenciada por dimensión y no como evidencia de fiabilidad homogénea del procedimiento de corrección.

Discusión

Los resultados evidencian un desempeño diferenciado de la herramienta de inteligencia artificial según la dimensión evaluada y el nivel de análisis considerado. A nivel agregado, las medias obtenidas por la IA fueron próximas al promedio humano y los errores absolutos se mantuvieron en rangos bajos o moderados. Sin embargo, las correlaciones de Pearson y los coeficientes ICC(3,1) muestran que esta proximidad no implica una reproducción homogénea de las diferencias relativas entre protocolos: fluidez y flexibilidad presentaron baja consistencia relativa, mientras que originalidad mostró una aproximación más sólida al patrón de puntuaciones humanas.

Este hallazgo es relevante porque la originalidad constituye una de las dimensiones más interpretativas del TTCT-Verbal. Su corrección exige valorar la rareza, singularidad o infrecuencia de las ideas en relación con un corpus de respuestas, lo que introduce una carga importante de juicio experto. En este sentido, el mejor desempeño relativo de la herramienta en originalidad sugiere que los modelos de lenguaje de gran escala pueden aportar valor en tareas que requieren procesamiento semántico, comparación entre respuestas y detección de patrones de recurrencia, en línea con estudios sobre puntuación automatizada de la originalidad y métodos computacionales aplicados a la creatividad (Acar et al., 2023; Organisciak et al., 2023).

No obstante, los resultados también muestran límites importantes. La baja consistencia relativa en fluidez y flexibilidad indica que la herramienta no reproduce con suficiente precisión las diferencias individuales entre protocolos en todas las dimensiones. Esto puede deberse a que, aunque estas dimensiones parecen más cuantificables, dependen de decisiones de segmentación, categorización y agrupación conceptual

que pueden variar entre evaluación humana y sistemas automatizados. Por tanto, la cercanía de las medias no debe interpretarse como equivalencia plena entre ambos sistemas de corrección.

El análisis de la variabilidad del juicio humano aporta un marco interpretativo relevante. La evaluación del pensamiento creativo involucra componentes interpretativos y contextuales que pueden generar diferencias entre evaluadores, incluso con criterios comunes de corrección. Esta situación ha sido ampliamente señalada en la literatura sobre el TTCT-Verbal, especialmente en originalidad (Torrance, 2008; Kim, 2006). En este estudio, la IA parece alinearse mejor con tendencias generales del promedio humano que con la reproducción exacta del juicio evaluativo en cada protocolo individual.

En consecuencia, los hallazgos no permiten sostener que la herramienta constituya un sistema autónomo y plenamente validado para corregir el TTCT-Verbal. Su valor debe entenderse como apoyo complementario dentro de sistemas híbridos de evaluación, donde puede contribuir a reducir carga temporal, organizar respuestas, ofrecer una primera estimación de puntuaciones y apoyar la revisión experta. Este enfoque es consistente con perspectivas actuales sobre el uso de IA en evaluación educativa, que enfatizan su función como apoyo a la toma de decisiones y no como sustituto del juicio humano (Guzik et al., 2023).

Asimismo, los resultados sugieren que la evaluación automatizada del pensamiento creativo debe analizarse por dimensión y no solo mediante indicadores globales de fiabilidad. El mejor desempeño en originalidad refuerza la necesidad de evitar generalizaciones sobre la fiabilidad global del sistema y de examinar de forma diferenciada el comportamiento de los modelos de lenguaje de gran escala cuando se aplican a instrumentos psicométricos complejos.

Finalmente, deben considerarse algunas limitaciones: la comparación humano-IA se realizó sobre una submuestra de 30 protocolos; el sistema fue evaluado mediante un procedimiento de *prompting* específico en un entorno conversacional web, sin control directo de parámetros como temperatura o semilla; el procesamiento se organizó en bloques, aunque la originalidad se estimó considerando el corpus total; y no se incorporó la dimensión de elaboración del TTCT-Verbal. Estas limitaciones refuerzan la necesidad de replicar el estudio con muestras más amplias, ejecuciones repetidas, modelos comparativos y condiciones de configuración más controladas.

Conclusiones, limitaciones y líneas futuras

El presente estudio aporta evidencia empírica sobre el desempeño de una herramienta de inteligencia artificial basada en un modelo de lenguaje de gran escala para apoyar la corrección del TTCT-Verbal. Los resultados muestran bajo error absoluto y una aproximación relevante al promedio de evaluadores/as humanos/as a nivel agregado, especialmente en originalidad.

No obstante, la herramienta mostró limitaciones para replicar diferencias individuales entre protocolos, particularmente en fluidez y flexibilidad, donde se observaron bajos niveles de consistencia relativa. Por tanto, el sistema automatizado no reproduce de manera homogénea el juicio humano en todas las dimensiones, sino que presenta un desempeño diferenciado según el componente evaluado. En este sentido, los resultados

respaldan el uso de la IA como apoyo complementario dentro de sistemas híbridos de evaluación del pensamiento creativo. Su potencial no radica en sustituir el juicio experto, sino en reducir la carga temporal, apoyar la revisión inicial de protocolos y aportar una referencia adicional en procesos supervisados por evaluación humana.

Entre las principales limitaciones se encuentran el tamaño y la homogeneidad de la muestra, así como la comparación humano-IA basada en una submuestra de 30 protocolos. Además, el desempeño del sistema depende del procedimiento de *prompting*, de las condiciones específicas de uso del modelo y del procesamiento por bloques, aunque la originalidad se estimó considerando el corpus total. También debe considerarse la posible variabilidad entre ejecuciones, dado el uso de un entorno conversacional web sin control directo de parámetros como temperatura o semilla. Finalmente, no se incorporó la dimensión de elaboración del TTCT-Verbal, lo que limita el alcance de la evaluación automatizada explorada.

Futuras investigaciones deberían ampliar la muestra a contextos educativos y culturales diversos, incorporar la dimensión de elaboración, comparar distintos modelos de lenguaje y evaluar la estabilidad del sistema mediante ejecuciones repetidas bajo condiciones controladas. Asimismo, resulta pertinente explorar estrategias de mejora como el refinamiento del prompt, el entrenamiento con datos anotados por expertos/as (*fine-tuning*) y modelos híbridos que integren evaluación automatizada y revisión humana selectiva.

En conjunto, este trabajo contribuye a comprender el potencial de los modelos de lenguaje de gran escala en la evaluación del pensamiento creativo, mostrando que su valor actual radica en complementar y fortalecer procesos de corrección existentes, más que en reemplazarlos.

Financiación

El presente trabajo es resultado del Proyecto I+D "DECHADOS digital. Creatividad inclusiva en museos y patrimonios", con referencia PID2024-155552OB-I00 financiado por el Ministerio de Ciencia, Innovación y Universidades del Gobierno de España MCIN/AEI/10.13039/501100011033 y los fondos FEDER de la Unión Europea. Una manera de hacer Europa.

Referencias

Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, Ch., y Organisciak, P. (2023). Applying automated originality scoring to the verbal form of Torrance Tests of Creative Thinking. *Gifted Child Quarterly, 67*(1), 3–17. https://doi.org/10.1177/00169862211061874

Alcántara Santuario, A. (2023). La inteligencia artificial y sus implicaciones en educación. *Perfiles Educativos*, *45*(Especial), 5–8. https://doi.org/10.22201/iisue.24486167e.2023.Especial.61687

Alfonso-Benlliure, V., Checa, I. y Meléndez, J. C. (2025). Long in the tooth for creativity? Differences in divergent thinking between young and older adults. *Thinking Skills and Creativity, 57, 101847* https://doi.org/10.1016/j.tsc.2025.101847

Barragán-Giraldo, D. F., Pirela-Morillo, J. E., y Riaño Diaz, J. A. (2024). Datificación en contextos educativos. Entre subjetivación y ética. *REICE. Revista Iberoamericana Sobre Calidad, Eficacia y Cambio en Educación*, *22*(2), 119-132. https://doi.org/10.15366/reice2024.22.2.007

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., y Forthmann, B. (2022). Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal, 34*(3), 245–260. https://doi.org/10.1080/10400419.2022.2025720

Bland, J. M., y Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bland, J. M., y Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Cesário, V., y Nisi, V. (2023). Designing mobile museum experiences for teenagers. *Museum Management and Curatorship, 38*(3), 272-292. https://doi.org/10.1080/09647775.2022.2111329

Cisternas San Martín, N., Guzman Muñoz, E., y Rivas Poblete, A. (2025). Inteligencia Artificial como asistente para desarrollar un currículum de educación contínua basado en competencias. *Revista de Investigación Educativa*, (43). https://doi.org/10.6018/rie.619441

Dumas, D., y Runco, M. (2018). Objectively Scoring Divergent Thinking Tests for Originality: A Re-Analysis and Extension. *Creativity Research Journal*, *30*(4), 466–468. https://doi.org/10.1080/10400419.2018.1544601

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica, 25*(2), 141–151. https://doi.org/10.11613/BM.2015.015

Goenechea, C., y Valero-Franco, C. (2024). Educación e Inteligencia Artificial: Un Análisis desde la Perspectiva de los Docentes en Formación. *REICE. Revista Iberoamericana Sobre Calidad, Eficacia y Cambio en Educación*, *22*(2), 33-50. https://doi.org/10.15366/reice2024.22.2.002

Guilford, J. P. (1950). Creativity. *American Psychologist, 5*(9), 444–454. https://doi.org/10.1037/h0063487

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Guzik, E. E., Byrge, C., y Gilde, C. (2023). The originality of machines: AI takes the Torrance Test. *Journal of Creative Behavior, 57*(1), 45–58. https://doi.org/10.1002/jocb.546

Huerta, R. (2023). Artivismo y creatividad en la formación docente: cementerios, arte y literatura. *Arte y Políticas de Identidad, 29*, 65-86. https://doi.org/10.6018/reapi.598721

Huerta, R. (2025). Aprendizaje basado en proyectos (ABP). Visitas a museos y fotografía creativa en la formación del profesorado. *Observar. Revista Electrònica De Didàctica De Les Arts, (19)*, 48–81. https://doi.org/10.1344/observar.2025.19.3

Huerta, R., y Alfonso-Benlliure, V. (2025). Museums as Catalysts for Creativity in Adolescence: A Review. *Heritage, 8*(8), 327. https://doi.org/10.3390/heritage8080327

Jarquín-Ramírez, M.-R., Alonso-Martínez, H., y Díez-Gutiérrez, E.-J. (2025). Chat GPT para un Sistema Educativo Justo, Democrático y Transformador. *REICE. Revista Iberoamericana Sobre Calidad, Eficacia y Cambio en Educación*, *23*(3), 1-14. https://doi.org/10.15366/reice2025.23.3.001

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity Research Journal, 18*(1), 3–14. https://doi.org/10.1207/s15326934crj1801_2

Kim, K. H. (2017). The Torrance Tests of Creative Thinking - Figural or Verbal: Which one should we use? *Creativity. Theories – Research – Applications*, *4*(2), 302–321. https://doi.org/10.1515/ctra-2017-0015

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

Mella-Mella, F. M., Calatayud Salom, M. A., y Lucas Calatayud, Á. J. (2026). Uso de la IA como Estrategia de y para el Aprendizaje en Educación Superior. *REICE. Revista Iberoamericana Sobre Calidad, Eficacia y Cambio en Educación*, *24*(1). https://doi.org/10.15366/reice2026.24.1.007

Organisciak, P., Acar, S., Dumas, D., y Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity, 49*, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Perchtold-Stefan, C. M., Fink, A., Rominger, C., y Papousek, I. (2024). Social exclusion increases antisocial tendencies: Evidence from retaliatory ideation in a malevolent creativity task. *Psychology of Aesthetics, Creativity, and the Arts, 18*(6), 1014–1025. https://doi.org/10.1037/aca0000500

Reiter-Palmon, R., Forthmann, B., y Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *The Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 144-152. https://doi.org/10.1037/aca0000227

Runco, M. A. (2023). Updating the Standard Definition of Creativity to Account for the Artificial Creativity of AI. *Creativity Research Journal, 37*(1), 1 –5. https://doi.org/10.1080/10400419.2023.2257977

Runco, M. A., Millar, G., Acar, S., y Cramond, B. (2010). Torrance Tests of Creative Thinking as predictors of personal and public achievement: A fifty-year follow-up. *Creativity Research Journal, 22*(4), 361–368. https://doi.org/10.1080/10400419.2010.523393

Sanjurjo Pérez, P., Solana Domínguez, I., y Arana Cuenca, A. (2026). Un año sin móvil: Estudio de caso sobre políticas de prohibición y su influencia en el rendimiento académico. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, *24*(1). https://doi.org/10.15366/reice2026.24.1.003

Sanz-Leal, M., y Orozco Gómez, M. L. (2025). Validación en español de una escala de Competencia Global en docentes en formación y en servicio. *Revista de Investigación Educativa*, (43). https://doi.org/10.6018/rie.564141

Silva Fuentealba, E. (2024). ChatGPT como catalizador del pensamiento creativo. [Chat GPT as a Catalyst for Creative Thinking]. *European Public & Social Innovation Review, 9*, 1-19. https://doi.org/10.31637/epsir-2024-410

Silva-Fuentealba, E., Valdés-León, G., y Oyarzún Yáñez, R. (2025). Inteligencia artificial en el aula: potenciando la resolución de problemas a través del pensamiento creativo. *Revista de Comunicación de la SEECI, 58*, 1–19. https://doi.org/10.15198/seeci.2025.58.e927

Torrance, E. P. (1966). *Torrance Tests of Creative Thinking: Norms-technical manual*. Personnel Press.

Torrance, E. P. (1974). *Torrance Tests of Creative Thinking: Norms-technical manual (Research edition)*. Personnel Press.

Torrance, E. P. (2008). *Torrance Tests of Creative Thinking: Technical manual and scoring guide*. Scholastic Testing Service.

Verdú-Pina, M., Serrano, V., Grimalt-Álvaro, C., y Usart, M. (2026). Perfiles docentes según la competencia digital autopercibida y el uso de la tecnología: Un análisis de clústeres. *Revista de Investigación Educativa*, (44). https://doi.org/10.6018/rie.672661

Welter, M. M., Jaarsveld, S., van Leeuwen, C., y Lachmann, T. (2016). Intelligence and Creativity: Over the Threshold Together? *Creativity Research Journal*, *28*(2), 212–218. https://doi.org/10.1080/10400419.2016.1162564

Fecha de recepción: 11 abril, 2026
Fecha de revisión: 15 abril, 2026
Fecha de aceptación: 18 mayo, 2026