Huerta, P. and Silva Fuentealba, E. (2026). Performance of an AI-based tool for marking the TTCT-Verbal in the assessment of creative thinking. *Revista de Investigación Educativa, 44*. https://doi.org/10.6018/rie.710621

Performance of an AI-based tool for scoring the TTCT-Verbal in the assessment of creative thinking

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

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

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

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

[1] **Correspondence:** Ricard Huerta, Universitat de València, Ricard.Huerta@uv.es

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.

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.

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

Introduction

In the contemporary educational field, interest in measuring and developing complex cognitive processes such as creativity has intensified alongside student-centred methodologies geared towards the development of competencies (Huerta and Alfonso-Benlliure, 2025). Active pedagogical approaches foster critical, reflective and creative thinking by promoting experiences of knowledge construction and open-ended problem-solving (Alcántara Santuario, 2023; Cesário and Nisi, 2023; Huerta, 2023). In this context, creativity has established itself as a key competence, increasing the need for valid and reliable instruments for its assessment (Jarquín-Ramírez et al., 2025; Sanz-Leal and Orozco Gómez, 2025).

The assessment of verbal creativity using standardised instruments is significant in educational psychology, talent identification and the study of creative thinking. The Torrance Test of Creative Thinking – Verbal Form (TTCT-Verbal) has established itself as

an international benchmark instrument, based on J. P. Guilford's model of the structure of the intellect (1950, 1967) and extensive psychometric validation (Torrance, 1966, 1974; Kim, 2006). This instrument assesses fluency, flexibility and originality through open-ended verbal tasks, and has demonstrated predictive power for long-term creative achievement (Runco et al., 2010).

However, manual scoring presents limitations: high time demands, the need for trained raters, and inter-rater variability, particularly regarding originality, due to its interpretative nature and reliance on criteria of statistical rarity (Torrance, 2008; Kim, 2006; Acar et al., 2023). These difficulties restrict their use in large-scale research, mass educational programmes and systematic assessments, increasing costs and reducing accessibility (Welter et al., 2016). Consequently, automated marking mechanisms have begun to receive increasing scientific attention (Alfonso-Benlliure et al., 2025).

At the same time, artificial intelligence systems based on large language models (LLMs) are transforming educational environments and the ways in which we interact with knowledge (Goenechea and Valero-Franco, 2024). These technologies enable learning support, automated information analysis and personalised feedback, expanding the potential of digital tools in higher education (Mella-Mella et al., 2026). Furthermore, the pedagogical use of digital technologies can facilitate access to resources, strengthen motivation, and promote more autonomous and collaborative learning (Sanjurjo Pérez et al., 2026; Verdú-Pina et al., 2026). Within this framework, LLMs offer possibilities for automating complex assessment processes.

From a technical perspective, LLMs can process complex semantic tasks in a short time, operate at scale and offer a degree of stability under defined configuration conditions (Acar et al., 2023; Guzik et al., 2023). Recent studies have explored their use in creativity assessment and the automated scoring of open-ended responses, reporting significant correlations with human judgements of originality through *text mining*, *prompt engineering* or *fine-tuning* (Acar et al., 2023; Dumas and Runco, 2018; Perchtold-Stefan et al., 2024). However, questions remain regarding their psychometric reliability in standardised instruments such as the TTCT-Verbal, particularly regarding agreement with human assessors and stability in quantitative dimensions—fluency and flexibility—compared to more interpretative dimensions, such as originality (Huerta, 2025).

In this context, the present study evaluates the degree of agreement between a tool based on a large-scale language model, configured using defined instructions, and the judgement of human assessors who are experts in marking the TTCT-Verbal. Using ICC, error metrics, linear correlations and Bland-Altman plots, we aim to provide evidence regarding its possibilities and limitations as a complementary support in hybrid systems for assessing creative thinking, considering its potential to reduce practical barriers to marking and the need to safeguard transparency, interpretative caution and expert supervision in the use of AI in educational assessment (Barragán-Giraldo et al., 2024).

Theoretical framework

Creativity and divergent thinking as cognitive constructs

Creativity has been defined as the ability to generate novel and appropriate ideas, products or solutions within a specific context (Runco, 2023). Since the mid-20th century, this construct has been extensively developed from a cognitive perspective, particularly following the work of J. P. Guilford, who distinguished between convergent and divergent thinking. Convergent thinking is geared towards solving problems with a single correct answer through logical and sequential processes (Guilford, 1967). In contrast, divergent thinking involves the fluid, exploratory and non-linear generation of multiple ideas or solutions in response to a single stimulus, constituting an operational core of creative behaviour (Guilford, 1950, 1967). Guilford proposed four main dimensions of divergent thinking: fluency, understood as the quantity of ideas generated; flexibility, referring to the diversity of conceptual categories used; originality, associated with the statistical rarity or unusualness of responses; and elaboration, linked to the degree of development or detail of the ideas. These dimensions have been widely used as empirical indicators of cognitive creativity in education and research. Subsequently, Ellis Paul Torrance (1966, 1974) operationalised these proposals through standardised tests to measure divergent thinking, establishing it as an empirical measure of creativity. Its assessment is relevant due to its association with academic, professional and innovative achievements (Runco et al., 2010). However, it presents methodological challenges linked to the subjectivity of the evaluative judgement, the time required for rigorous marking, and inter-rater variability that may compromise psychometric consistency (Torrance, 2008; Kim, 2006).

The Torrance Test of Creative Thinking (TTCT-Verbal) as a benchmark instrument

The Torrance Test of Creative Thinking – Verbal (TTCT-Verbal), developed by E. Paul Torrance in the 1960s, is one of the most widely used and internationally validated instruments for assessing verbal creative thinking. Based on Guilford's theory of the structure of intellect, it was initially conceived as part of the *Minnesota Tests of Creative Thinking*. Since its first formal version, published in 1966, it has undergone several revisions. It can be administered from Year 1 through to adulthood, features parallel forms A and B to reduce the learning effect, has been adapted into more than 35 languages, and has evidence of cross-cultural validity. The standard verbal form includes six tasks: *Asking*, *Guessing Causes*, *Guessing Consequences*, *Product Improvement*, *Unusual Uses* and *Just Suppose*. These are administered over approximately 45 minutes using open-ended verbal stimuli that allow for the generation of multiple creative responses (Torrance, 2008). The TTCT-Verbal assesses three core dimensions: fluency, referring to

the number of relevant, interpretable and non-redundant responses; flexibility, linked to the diversity of conceptual categories employed; and originality, associated with the statistical rarity or infrequency of responses according to established norms. The latter is the most complex dimension, as it requires normative criteria and expert judgement. From a psychometric perspective, the TTCT- -Verbal has demonstrated high test-retest reliability, with coefficients ranging from 0.80 to 0.90, and acceptable internal consistency (Torrance, 2008). Furthermore, longitudinal studies have demonstrated its predictive validity by linking early scores with creative achievements up to five decades later (Runco et al., 2010). However, although inter-rater reliability may be high with trained assessors, it tends to decline in terms of originality when there is no comprehensive training (Kim, 2006).

Limitations of human scoring of the TTCT-Verbal

Manual scoring of the TTCT-Verbal involves a high cognitive and time burden. Each protocol may require between 20 and 45 minutes of assessment by each rater, depending on the complexity of the responses and the dimension being assessed. This burden is intensified in the originality dimension, which demands qualitative judgements with a high interpretative load (Torrance, 2008; Acar et al., 2023; Organisciak et al., 2023). Although various studies report inter-rater reliability coefficients of between 0.80 and 0.95 when assessors have been adequately trained, this reliability may decline in contexts where such training is not intensive, particularly in the dimension of originality, where values may range from 0.70 to 0.85 (Kim, 2006). These variations highlight the sensitivity of the assessment process to contextual and training factors, which introduces a margin of uncertainty regarding the consistency of scores. Taken together, these limitations not only entail high financial costs and a considerable investment of time, but also reduce the feasibility of using the TTCT in large-scale research or in mass educational settings (Welter et al., 2016). In this regard, the reliance on highly trained assessors and the variability inherent in human judgement pose a significant methodological challenge, linked to the need to explore alternatives that allow the instrument's psychometric standards to be maintained, whilst optimising the efficiency and consistency of the assessment process.

The potential of large-scale language models in the assessment of creative thinking

Large-scale language models (LLMs) have demonstrated the ability to process complex semantic tasks, which has driven their exploration in psychological and educational assessment. Among their potential advantages are operational stability under controlled settings, scalability, and a reduction in the time required to mark open-ended responses (Acar et al., 2023). These models can mitigate issues associated with human marking, such as fatigue or intra-rater variability; however, they do not eliminate

bias nor do they guarantee greater objectivity in themselves. Their performance depends on the training data, the *prompt* design, the conditions of use, and the parameters available in each technological environment. Therefore, their use in assessment requires well-defined procedures and rigorous *prompt* engineering, guided by explicit scoring criteria (Guzik et al., 2023).

Various studies have applied computational methods to the automatic assessment of creativity, particularly in terms of originality, using semantic distance, text mining and LLMs. These studies report significant correlations with human evaluations and demonstrate their potential for estimating specific components of creative thinking. However, they highlight the need to distinguish between systems based on semantic distance, text mining, supervised learning and LLMs, as they operate under different methodological assumptions (Beaty et al., 2022; Acar et al., 2023; Organisciak et al., 2023). Overall, recent literature shows progress in the automated assessment of originality and in the pedagogical use of generative tools to stimulate fluency, flexibility and originality. However, it also indicates that AI performance varies depending on the dimension analysed, making it necessary to examine each component of creative thinking separately and avoid generalisations about the overall reliability of automated correction systems (Beaty et al., 2022; Acar et al., 2023; Organisciak et al., 2023; Silva-Fuentealba et al., 2024; Silva-Fuentealba et al., 2025).

Research objectives

The assessment of creative thinking using the Torrance Test of Creative Thinking – Verbal Form (TTCT-Verbal) is relevant in educational research and practice, as it allows for the assessment of fluency, flexibility and originality. However, its application is limited by the subjectivity of interpretative dimensions, particularly originality, inter-rater variability, and the high time and financial costs of manual marking by experts (Acar et al., 2023; Kim, 2006; Torrance, 2008; Reiter-Palmon et al., 2019). These limitations justify exploring technological tools that support marking without replacing expert supervision. In this context, the study examines the performance of an AI-based tool, specifically a large-scale language model configured using delimited instructions, as a complementary aid for marking the TTCT-Verbal.

Overall objective

To assess the degree of agreement between the scores generated by an artificial intelligence tool and those assigned by expert human markers in the marking of the TTCT-Verbal.

Specific objectives

To analyse the relative consistency between the scores generated by the AI and the average of human evaluations in terms of fluency, flexibility and originality, using the Intraclass Correlation Coefficient ICC(3,1) and complementary error and association metrics (Koo and Li, 2016).

1. To examine the tool's performance across each dimension assessed by the TTCT-Verbal, taking into account the greater human variability reported in originality due to its subjective nature and dependence on criteria of rarity (Kim, 2006; Torrance, 2008).
2. To compare the time efficiency of the automated system against human marking, taking into account the processing time of the protocols and the need for expert supervision, given that manual marking of the TTCT-Verbal typically takes between 20 and 45 minutes per protocol (Acar et al., 2023; Perchtold-Stefan et al., 2024; Reiter-Palmon et al., 2019).

Taken together, these objectives aim to provide preliminary evidence on the performance of an AI tool in correcting the TTCT-Verbal, considering both its degree of alignment with human judgement and its limitations in consistently reproducing the various dimensions of creative thinking.

Method

An instrumental quantitative study was conducted to estimate the degree of consistency and alignment between the scores generated by an artificial intelligence tool and those assigned by expert human assessors in the scoring of the Torrance Test of Creative Thinking, Verbal Form (TTCT-Verbal). The analysis focused on three dimensions of creative thinking assessed by the instrument: fluency, flexibility and originality.

Participants and protocols

The sample was recruited from the University of Valencia, comprising students enrolled in a professional postgraduate programme aimed at training future secondary school teachers. The test was administered within the framework of a university course focused on the development of creative thinking, linked to experiences in arts education, museums and cultural mediation (Huerta, 2023; Huerta and Alfonso-Benlliure, 2025; Cisternas San Martín et al., 2025).

Forty-seven TTCT-Verbal response protocols were used. From this set, 30 protocols, corresponding to 64% of the total sample, were randomly selected for comparison

between the artificial intelligence tool and the human assessors. The 47 protocols were processed by the automated tool, whilst the 30 selected protocols were also assessed by the panel of human evaluators, enabling a comparison to be made between the two scoring systems.

Human assessors

Six expert judges took part, with an average of 13.6 years' experience in fields such as education, philosophy, media studies, and cybersecurity and computer science. The group comprised three men and three women, ensuring a balanced representation.

Before beginning the assessment process, all judges received standard training on the TTCT-Verbal marking criteria, with the aim of ensuring a shared understanding of how to interpret and score participants' answers. Each evaluator reviewed the scripts independently, following rubrics adapted to the dimensions of fluency, flexibility and originality. They were given a ten-day period to do so, during which they submitted both their scores and qualitative observations on the scripts analysed.

Instrument: TTCT-Verbal

The instrument used was the Torrance Test of Creative Thinking – Verbal Form (TTCT-Verbal), employed to assess dimensions of creative thinking through open-ended verbal production tasks. In this study, three dimensions were considered: fluency, flexibility and originality. This selection is based on the scoring structure of the TTCT-Verbal described in the specialist literature, where its main components correspond to fluency, originality and flexibility, whilst elaboration forms part of the normative subscales of the TTCT-Figural alongside title abstraction and resistance to premature closure (Kim, 2017).

Fluency corresponds to the number of relevant, interpretable and non-redundant responses produced by each participant. Flexibility refers to the diversity of conceptual categories present in the responses, that is, the ability to change perspective or use different frames of reference. Originality relates to the degree of rarity, uniqueness or infrequency of the ideas produced in comparison with the corpus of responses analysed.

In the present study, originality was assessed at the level of the idea within each protocol. Each idea was awarded 2 points when it was considered highly original, that is, when it did not appear in other protocols within the total corpus analysed; 1 point when it appeared in some protocols but not in the majority; and 0 points when it corresponded to a frequent or recurring response in the corpus. Subsequently, the scores assigned to the different ideas were added together to obtain a total originality score per protocol. Consequently, although each individual idea could receive between 0 and 2 points, the total originality score of a protocol could exceed that range.

Automated correction system

The artificial intelligence tool was implemented using the Gemini model 1.5 Pro (Google), configured via a structured system *prompt* and run in a web-based conversational environment. The full system specification, including scoring rules, analysis criteria and *prompt* structure, is detailed in Appendix A, with the aim of promoting the transparency and replicability of the procedure.

The structured instructions defined the analysis in three dimensions:

- Fluency (F): breaking down sentences into minimal units of information, following the criterion of atomic fluency.
- Flexibility (FX): classification of responses into five conceptual blocks: physical, technological, utilitarian, social and artistic.
- Originality (O): assignment of scores at the idea level, considering the recurrence or uniqueness of each response in relation to the total corpus analysed.

The processing of the transcripts using the artificial intelligence tool was carried out in blocks of 10 cases as an operational strategy to facilitate the entry of responses into the conversational environment and maintain the traceability of the analysis. However, the estimation of originality was not carried out independently within each block, but rather by considering the total corpus of analysed protocols. In other words, the recurrence or uniqueness of each idea was assessed by comparing it with the complete set of processed responses.

Procedure

The human judges evaluated the 30 selected protocols in a blind and independent manner. The average time taken for human correction was approximately 22 minutes per set of five protocols, equivalent to around 4–5 minutes per protocol.

The artificial intelligence tool processed all 47 protocols in blocks of 10 cases. Subsequently, the 30 protocols that had also been reviewed by the human evaluators were selected for comparative analysis. The output generated by the system was structured by dimensions and scores, allowing the automated scores to be compared with the average of the six human evaluators in terms of fluency, flexibility and originality.

Data analysis

To assess the consistency and degree of agreement between the scores generated by the artificial intelligence tool and the average of the six human evaluators, the Intraclass Correlation Coefficient (ICC) was used, following the recommendations of Koo and Li (2016). This coefficient is suitable for continuous variables and allows the consistency

between evaluation systems to be estimated. The spec s of the automated system are presented in Appendix A, in order to promote the transparency and replicability of the procedure.

The ICC(3,1) model was used, corresponding to a bidirectional mixed-effects design with fixed evaluators and focused on consistency. This choice was made because the human assessors were selected for their professional experience and because the aim was to analyse the extent to which the automated tool replicates the scoring pattern of expert human judgement, rather than to establish exact equivalence between each human and AI score.

The ICC was calculated by comparing the AI scores with the human mean across the three dimensions of the TTCT-Verbal: fluency, flexibility and originality. Its interpretation followed the criteria of Koo and Li (2016): values below 0.50 indicate low relative consistency; between 0.50 and 0.75, moderate consistency; between 0.75 and 0.90, good consistency; and above 0.90, excellent consistency.

As supplementary analyses, the mean absolute error (MAE), root mean square error (RMSE) and Pearson correlations were calculated. The MAE and RMSE enabled the magnitude of the discrepancies between the AI and the human average to be estimated, whilst Pearson allowed the linear association between the two systems to be analysed.

Finally, the Bland-Altman plot was used to examine the distribution of differences, identify potential systematic biases and observe the behaviour of the error across the measurement range. Given the sample size, this analysis was used as a descriptive and visual complement to the quantitative metrics, rather than as conclusive evidence of the psychometric stability of the automated system.

Results

Measures of absolute agreement (MAE, RMSE, bias), consistency (Pearson's correlation) and graphical agreement analysis (Bland–Altman) were used to evaluate the performance of the automated system across the dimensions of fluency (F), flexibility (FX) and originality (O)

To interpret the originality dimension, it should be noted that the 0–2 point scale was applied to each individual idea within the protocol, not to the protocol's final score. The originality score reported in Table 1 corresponds to the sum of the scores assigned to the various ideas of each participant. Therefore, the means of 5.87 for human evaluators and 5.90 for AI should not be interpreted as values on a final scale of 0 to 2, but as aggregate scores per protocol.

Descriptive statistics

Table 1 summarises the means, standard deviations of the scores and the human-AI bias.

Table 1

Descriptive statistics of the scores (n = 30)

Dimension	Human mean (SD)	AI Mean (SD)	Bias (Human – AI)
Fluency (F)	6.83 (1.02)	6.60 (0.56)	+0.23
Flexibility (FX)	4.00 (0.69)	4.10 (0.55)	–0.10
Originality (O)	5.87 (1.91)	5.90 (1.09)	–0.03

Note. The means between the two systems are practically equivalent across all three dimensions, with differences of less than 0.25 points, suggesting the absence of any significant systematic bias in the AI's estimates relative to the human average.

Agreement between AI and the average of human evaluators

Metrics of absolute agreement and consistency were calculated to assess the relationship between the AI scores and the average of the human evaluators.

Table 2

Agreement metrics: AI vs. human average (n = 30)

Dimension	MAE	RMSE	Pearson's r (p)
Fluency (F)	0.83	1.10	0.120 (0.527)
Flexibility (FX)	0.43	0.75	0.272 (0.146)
Originality (O)	1.10	1.33	0.721 (< 0.001)

The mean absolute error (MAE) and root mean square error (RMSE) indicate low deviations between the two systems, particularly in the flexibility dimension (MAE = 0.43), suggesting a high degree of proximity in absolute terms. In fluency and originality, the errors remain within moderate ranges, with no evidence of significant systematic discrepancies.

Pearson's correlations show varying patterns across dimensions. In originality, a strong association is observed ($r = 0.721$, $p < 0.001$), indicating that the AI captures the general pattern of variation in human judgement in this dimension. In contrast, for fluency ($r = 0.120$, $p = 0.527$) and flexibility ($r = 0.272$, $p = 0.146$), the correlations are low, suggesting limitations in the system's ability to replicate the relative differences between subjects in these dimensions.

To complement the above metrics, the ICC(3,1) was calculated between the AI scores and the human mean. Table 3 presents the values by dimension, with their 95% confidence intervals and interpretation.

Table 3

Intraclass Correlation Coefficient between AI and human average by dimension

Dimension	ICC(3,1)	95% CI	Interpretation
Fluency (F)	0.11	[-0.25, 0.45]	Low relative consistency
Flexibility (FX)	0.26	[-0.10, 0.57]	Low relative consistency
Originality (O)	0.62	[0.34, 0.80]	Moderate relative consistency

The ICC(3,1) analysis presented in Table 3 showed varying performance across dimensions. Fluency and flexibility exhibited low relative consistency, whilst originality reached a moderate level, suggesting that the AI came closer to the pattern of human scores in the most interpretative dimension of the TTCT-Verbal.

Taken together, these results suggest that, whilst AI accurately reproduces the means and exhibits low absolute error, its ability to reflect individual variations varies depending on the dimension assessed.

In terms of time efficiency, human correction took approximately 22 minutes per set of five protocols, equivalent to around 4–5 minutes per protocol. In contrast, the AI tool processed the 47 protocols in considerably less time, once the prompt had been configured and the cases organised into blocks. This difference suggests a substantial reduction in the time burden associated with the correction process, although it should be interpreted with caution, given that the time spent on configuration, review and human oversight of the automated procedure also forms part of the evaluation process.

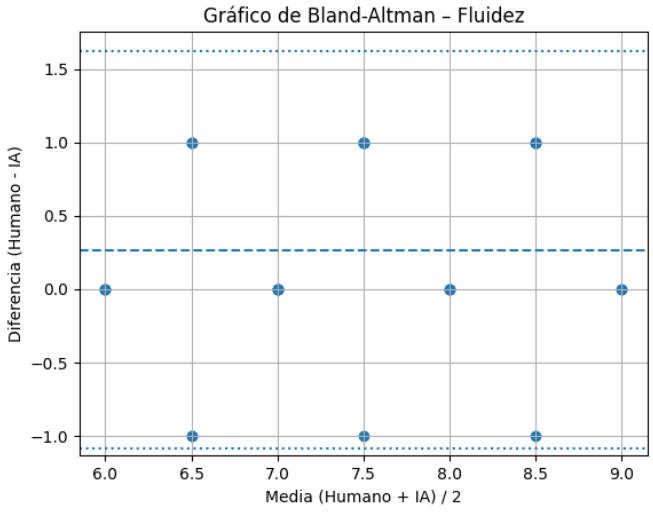
Graphical analysis of agreement (Bland-Altman)

Bland-Altman plots were produced for each dimension, representing the difference between the human and AI mean scores plotted against the mean of both scoring systems (Bland and Altman, 1986, 1999; Giavarina, 2015). This analysis was used as a descriptive complement to the quantitative metrics, with the aim of exploring the distribution of discrepancies between the two scoring methods.

Across the three dimensions, the mean bias was close to zero, with values ranging from –0.10 to +0.23. Furthermore, between 93% and 97% of the observations fell within the limits of agreement, calculated as ±1.96 standard deviations of the differences. Visually, no marked patterns of heteroscedasticity were observed, although this result should be interpreted with caution due to the sample size of the study.

Figure 1 presents the Bland-Altman plot for the fluency dimension. Each point represents the difference between the human mean score and the AI score as a function of the mean of both measurements.

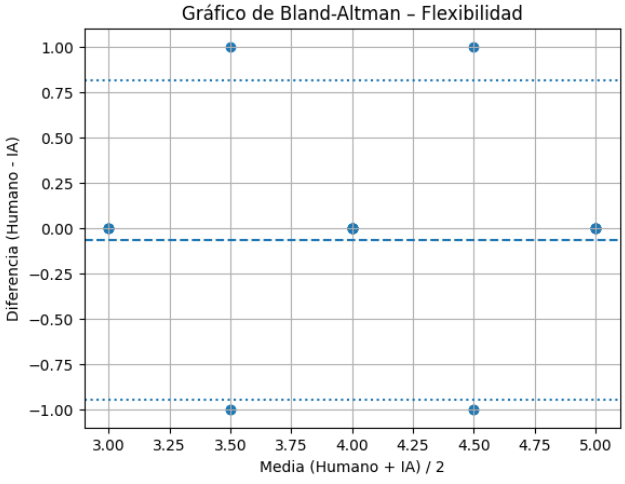
Figure 1. Bland-Altman plot for the fluidity dimension.



Note. Prepared by the authors using TTCT data

Figure 2 shows the Bland-Altman plot for the flexibility dimension, which examines the distribution of discrepancies between the human average and the AI average relative to the mean of their scores.

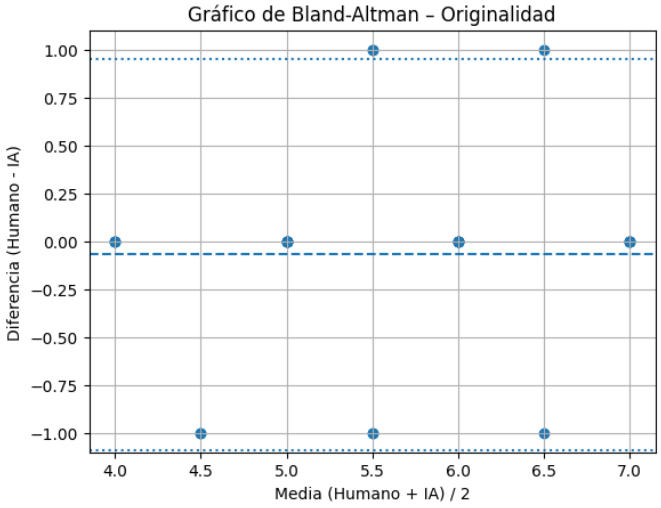
Figure 2. Bland–Altman plot for the flexibility dimension.



Note: Prepared by the authors using data from the TTCT.

Figure 3 presents the Bland-Altman plot for the originality dimension, illustrating the relationship between score differences and the mean of human and AI evaluations.

Figure 3. Bland–Altman plot for the originality dimension.



Note: Prepared by the authors using data from the TTCT

Overall, the Bland-Altman plots suggest that the differences between AI and the human average are distributed without any obvious systematic bias across the three dimensions analysed. In methodological terms, these findings are consistent with the interpretation criteria of the Bland–Altman method, where the focus lies not on exact agreement between measurements, but on the magnitude and distribution of discrepancies within an acceptable range (Bland and Altman, 1986, 1999; Giavarina, 2015). However, these results should be understood as complementary descriptive evidence, and not as conclusive proof of the psychometric stability of the automated system across the entire range of scores.

Stability and relative consistency

In summary, the results show that the artificial intelligence tool exhibits low absolute error and an absence of relevant systematic bias at the aggregate level. However, its ability to reproduce the relative differences between protocols varies depending on the dimension assessed. The closest approximation to the pattern of human scores was observed in originality, whilst fluency and flexibility showed lower relative consistency. Therefore, the performance of the automated system should be interpreted differently for each dimension and not as evidence of homogeneous reliability of the marking procedure.

Discussion

The results demonstrate that the performance of the artificial intelligence tool varies depending on the dimension assessed and the level of analysis considered. At an aggregate level, the means obtained by the AI were close to the human average, and absolute errors remained in low or moderate ranges. However, Pearson's correlations and the ICC(3,1) coefficients show that this proximity does not imply a consistent reproduction of the relative differences between protocols: fluency and flexibility showed low relative consistency, whilst originality demonstrated a more robust alignment with the pattern of human scores.

This finding is relevant because originality constitutes one of the most interpretative dimensions of the TTCT-Verbal. Its assessment requires evaluating the rarity, uniqueness or infrequency of ideas in relation to a corpus of responses, which introduces a significant element of expert judgement. In this regard, the tool's relatively better performance on originality suggests that large-scale language models can add value in tasks requiring semantic processing, comparison of responses and detection of recurring patterns, in line with studies on automated originality scoring and computational methods applied to creativity (Acar et al., 2023; Organisciak et al., 2023).

However, the results also reveal significant limitations. The relatively low consistency in fluency and flexibility indicates that the tool does not reproduce individual differences between protocols with sufficient accuracy across all dimensions. This may be because, although these dimensions appear more quantifiable, they depend on decisions regarding segmentation, categorisation and conceptual grouping that may vary between human assessment and automated systems. Therefore, the closeness of the means should not be interpreted as full equivalence between the two marking systems.

The analysis of the variability in human judgement provides a relevant interpretative framework. The assessment of creative thinking involves interpretative and contextual components that can lead to differences between assessors, even with common scoring criteria. This situation has been widely noted in the literature on the TTCT-Verbal, particularly regarding originality (Torrance, 2008; Kim, 2006). In this study, the AI appears to align better with general trends in the human mean than with the exact reproduction of evaluative judgement in each individual protocol.

Consequently, the findings do not support the claim that the tool constitutes an autonomous and fully validated system for marking the TTCT-Verbal. Its value should be understood as complementary support within hybrid assessment systems, where it can help reduce the time burden, organise responses, provide an initial estimate of scores, and support expert review. This approach is consistent with current perspectives on the use of AI in educational assessment, which emphasise its role as a support for decision-making rather than as a substitute for human judgement (Guzik et al., 2023).

Furthermore, the results suggest that automated assessment of creative thinking should be analysed by dimension and not solely through global reliability indicators. The superior performance in originality reinforces the need to avoid generalisations about the system's overall reliability and to examine the behaviour of large-scale language models in a differentiated manner when applied to complex psychometric instruments.

Finally, a number of limitations should be noted: the human-AI comparison was conducted on a subsample of 30 protocols; the system was evaluated using a specific *prompting* procedure in a web-based conversational environment, without direct control over parameters such as temperature or seed; processing was organised in blocks, although originality was estimated by considering the entire corpus; and the elaboration dimension of the TTCT-Verbal was not incorporated. These limitations reinforce the need to replicate the study with larger samples, repeated runs, comparative models and more controlled experimental conditions.

Conclusions, limitations and future directions

This study provides empirical evidence on the performance of an artificial intelligence tool based on a large-scale language model to support the marking of the TTCT-Verbal. The results show low absolute error and a relevant approximation to the average of human markers at the aggregate level, particularly in originality.

However, the tool showed limitations in replicating individual differences across protocols, particularly in fluency and flexibility, where low levels of relative consistency were observed. Therefore, the automated system does not uniformly replicate human judgement across all dimensions, but rather exhibits varying performance depending on the component being assessed. In this regard, the results support the use of AI as a complementary aid within hybrid systems for assessing creative thinking. Its potential lies not in replacing expert judgement, but in reducing the time burden, supporting the initial review of protocols, and providing an additional reference in processes supervised by human evaluation.

Key limitations include the sample size and homogeneity, as well as the human-AI comparison based on a subsample of 30 protocols. Furthermore, the system's performance depends on the *prompting* procedure, the specific conditions of model use and block processing, although originality was estimated by considering the entire corpus. The potential variability between runs must also be considered, given the use of a web-based conversational environment without direct control over parameters such as temperature or seed. Finally, the elaboration dimension of the TTCT-Verbal was not incorporated, which limits the scope of the automated ' ' evaluation explored.

Future research should expand the sample to diverse educational and cultural contexts, incorporate the elaboration dimension, compare different language models, and assess the system's stability through repeated runs under controlled conditions. It is also

pertinent to explore improvement strategies such as prompt refinement, training with expert-annotated data (*fine-tuning*), and hybrid models that integrate automated assessment and selective human review.

Overall, this work contributes to understanding the potential of large-scale language models in the assessment of creative thinking, showing that their current value lies in complementing and strengthening existing marking processes, rather than replacing them.

Funding

This work is the result of the R&D project "DECHADOS digital. Inclusive creativity in museums and heritage", reference PID2024-155552OB-I00, funded by the Spanish Ministry of Science, Innovation and Universities (MCIN/AEI/10.13039/501100011033) and the European Union's ERDF funds. A way of building Europe.

References

Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, Ch., and 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). Artificial intelligence and its implications for education. *Perfiles Educativos*, *45*(Special), 5–8. https://doi.org/10.22201/iisue.24486167e.2023.Especial.61687

Alfonso-Benlliure, V., Checa, I. and 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., and Riaño Diaz, J. A. (2024). Datafication in educational contexts. Between subjectivation and ethics. *REICE. Ibero-American Journal on Quality, Effectiveness and Change in Education*, *22*(2), 119–132. https://doi.org/10.15366/reice2024.22.2.007

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., and 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., and 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., & 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., and 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., and Rivas Poblete, A. (2025). Artificial Intelligence as an assistant in developing a competency-based continuing education curriculum. *Journal of Educational Research*, (43). https://doi.org/10.6018/rie.619441

Dumas, D., and 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., and Valero-Franco, C. (2024). Education and Artificial Intelligence: An Analysis from the Perspective of Trainee Teachers. *REICE. Ibero-American Journal on Quality, Effectiveness and Change in Education*, *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., and 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). Artivism and creativity in teacher training: cemeteries, art and literature. *Art and Identity Policies, 29*, 65–86. https://doi.org/10.6018/reapi.598721

Huerta, R. (2025). Project-based learning (PBL). Museum visits and creative photography in teacher training. *Observar. Revista Electrònica De Didàctica De Les Arts, (19)*, 48–81. https://doi.org/10.1344/observar.2025.19.3

Huerta, R., and 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., and Díez-Gutiérrez, E.-J. (2025). ChatGPT for a Fair, Democratic and Transformative Education System. *REICE. Ibero-American Journal on Quality, Effectiveness and Change in Education*, *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., & Li, M. Y. (2016). A guideline for 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., and Lucas Calatayud, Á. J. (2026). The use of AI as a strategy for and in learning in higher education. *REICE. Ibero-American Journal on Quality, Effectiveness and Change in Education*, *24*(1). https://doi.org/10.15366/reice2026.24.1.007

Organisciak, P., Acar, S., Dumas, D., and 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., and 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., and 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., and 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., and Arana Cuenca, A. (2026). A year without a mobile phone: A case study on ban policies and their influence on academic performance. *REICE. Ibero-American Journal on Quality, Effectiveness and Change in Education, 24*(1). https://doi.org/10.15366/reice2026.24.1.003

Sanz-Leal, M., and Orozco Gómez, M. L. (2025). Validation in Spanish of a Global Competence Scale for trainee and practising teachers. *Journal of Educational Research*, (43). https://doi.org/10.6018/rie.564141

Silva Fuentealba, E. (2024). ChatGPT 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., and Oyarzún Yáñez, R. (2025). Artificial intelligence in the classroom: enhancing problem-solving through creative thinking. *SEECI Communication Journal, 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., and Usart, M. (2026). Teacher profiles according to self-perceived digital competence and technology use: A cluster analysis. *Journal of Educational Research*, (44). https://doi.org/10.6018/rie.672661

Welter, M. M., Jaarsveld, S., van Leeuwen, C., and 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

Date received: 11 April 2026
Review date: 15 April 2026
Date of acceptance: 18 May 2026