From Expert Judgement to AI Insights: The Evolving Landscape of Scoring Methods in Creativity Assessments

Jue Wang[1]

[1] Hong Kong Examinations and Assessment Authority, 68 Gillies Ave South, Hung Hom, Kowloon, Hong Kong
jwang@hkeaa.edu.hk

Abstract. Creativity is increasingly recognized as a core competency in global education systems, yet valid, reliable and fair scoring approaches for both large-scale and classroom assessments remain a longstanding challenge. Traditional human scoring methods, most notably the Consensual Assessment Technique, provide nuanced, context-sensitive judgments but are constrained by rater variability, resource demands, and psychometric issues such as scoring severity, centrality and judgmental biases. Advances in automated scoring have introduced computational techniques ranging from semantic distance metrics to AI-powered models that approximate human ratings with growing accuracy. However, these systems struggle to detect culturally situated novelty and unconventional responses that underpin creative performance. This manuscript traces the evolution of creativity scoring from expert judgment to AI-powered methods and proposes a human-in-the-loop framework that integrates the scalability and consistency of AI scoring with the interpretive depth of human experts. The system incorporates automated prescreening, AI-generated scores on conventional responses, followed by targeted human review of atypical or high-uncertainty responses, and psychometric examination using modern measurement techniques (e.g., many-facet Rasch model) to monitor both human and AI scoring performance. This hybrid approach enhances validity, reliability, fairness, while supporting culturally responsive, transparent and accountable assessment practices, offering a pathway for advancing creativity assessment in both large-scale and classroom settings.

Keywords: Creativity assessment, expert scoring, automated scoring, human-AI collaborative scoring system, many-facet Rasch modeling.

1 Introduction

Creativity as one of the 21st century skills has earned increasing attention. The OECD Learning Compass 2030 provides a global vision for education where student agency and co-agency are emphasized as the foundation for learning in an uncertain and rapidly changing world (OECD, 2019). Creativity is one of the skills that the learning compass highlighted that learners require to thrive and shape their future. Many countries and regions in the Pacific Rim have recognized creativity as a core competency to be assessed and integrated into curriculum design. For example, Singapore's 21st Century

© The Author(s) 2026
C. Y. Lye et al. (eds.), *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS 2025)*, Atlantis Highlights in Social Sciences, Education and Humanities 56,
https://doi.org/10.2991/978-94-6239-834-4_5

Competencies Framework highlights inventive thinking as a core competency (Singapore Ministry of Education, 2023). Hong Kong's Learning to Learn Curriculum emphasizes creative thinking skill as one of the foundations for fostering students' whole-person development and lifelong learning capabilities (Hong Kong Education Bureau, 2001).

With the global administration of PISA 2022 creativity assessment, a standardized approach for scoring is of paramount importance that could ensure the validity, reliability, fairness, and comparability of the creativity scores. Creativity assessment has a long history relying on the judgement of human experts. Expert raters interpreted originality, elaboration, flexibility, and aesthetic quality through professional experience and their interpretation of creativity. This approach brought nuance and contextual sensitivity, but it also carried challenges due to the subjectivity, such as limited agreement or inter-rater reliability and inconsistent perceptions among experts (Long, 2014; Wang et al., 2025).

As digital tools matured, automated scoring methods emerged as a promising alternative. Primi et al. (2025) summarized the existing automated scoring methods for different types of creativity tasks. Endeavors include unsupervised models that extract word embeddings to compute a score based on semantic similarity (e.g., Beaty & Johnson, 2021), supervised approaches through fine-tuned large language models (LLMs) which significantly increased the prediction accuracy with human ratings (e.g., Patterson et al., 2023, 2024), and few-shot and zero-shot prompting strategies (e.g., Organisciak et al. 2023) that do not require substantial model pre-training. Automated scoring methods offered speed and consistency for large-scale assessments and classroom-level feedback. Yet these systems often struggled with the evaluation of originality within cultural-nuance contexts, which is the core of creativity (Primi et al., 2025).

Organisciak et al. (2023) compared different methods and indicated that both the fine-tuned model and few-shot prompting outperformed the traditional semantic distance methods ($r = 0.12 - 0.26$), while the few shot approaches ($r = 0.70$) had a slightly lower accuracy with human ratings than fine-tuned model ($r = 0.76 - 0.81$). Although methods based on prompt engineering showed slightly weaker performance than fine-tuned models, they offer greater flexibility to generalize across different contexts and make it more accessible for educational researchers and practitioners to design their own scoring engines. With that accessibility and flexibility, it creates the pathway for building a human-AI collaborative scoring system that could blend the strengths of expert judgement with the efficiency of automation. While AI models handle the heavy-lifting work, including pre-scoring large volumes of responses to identify relevance or appropriateness, detect common patterns, and flag unconventional yet potentially original responses, human experts then refine, validate, and interpret these outputs, ensuring that the final scores remain aligned with the measurement construct, cultural relevance, and the evolving definitions of creativity. This hybrid model represents more than a technical upgrade. It signals a shift in philosophy that

creativity assessment is no longer a choice between human insight and computational power; instead, it is an integrated ecosystem where AI amplifies human expertise, and human guides AI toward a more meaningful evaluation. As classroom-level and large-scale assessment systems continue to evolve, this human-AI partnership may offer a scalable, reliable, and context-sensitive path forward that values the complexity of creativity while taking the advantage of the efficiency of intelligent technology.

The following sections will first focus on human scoring challenges and strengths through a measurement perspective, followed by a brief overview of the recent developments on automated scoring methods. Next, we describe a human-in-the-loop AI-based scoring system with a conceptualized workflow that can be integrated into either large-scale or classroom-level creativity assessment systems.

2 Measurement Challenges and Strengths of Human Scoring in Creativity Assessments

The formalization of subjective creativity assessment emerged with Amabile's (1982) Consensual Assessment Technique (CAT), which remains the most popular human-scored paradigm in the field. The CAT is grounded in a consensual definition of creativity that a product is creative to the extent that appropriate judges independently agree it is creative. This definition intentionally avoids explicit criteria or a detailed rubric. Under this framework, interrater agreement is treated as a proxy for construct validity (Amabile, 1982), reflecting the assumption that domain experts implicitly share consistent criteria for judging creativity. The flexibility and ecological validity of CAT contributed to its widespread uses across domains, age groups, and cultural contexts, and it has been described as the gold standard for assessing creativity (Carson, 2006; Long & Plucker, 2015; Plucker & Makel, 2010).

Subjective creativity assessment can be situated within the broader framework of rater-mediated assessment, in which creative performance is interpreted through the lens of human judgement. Human scoring is inherently a rater-mediated process where observed scores reflect not only the creativity of the product but also the cognitive, experiential, and interpretive processes of the human raters (Engelhard & Wang, 2024). Contemporary perspectives, aligned with the *Standards for Psychological and Educational Testing* (AERA, APA, NCME, 2014), emphasize that validity evidence must account for the extent to which raters' interpretive processes support the intended meaning of the scores. The lens model depicts a conceptual picture in which raters rely on different cues when forming their cognitive judgments, where task domains, response features, and raters' own characteristics may lead to systematic differences in how the same creative performance is perceived by different experts (Long & Wang, 2022). Under this conceptualization, the measurement process reflects an interaction between the ratee's underlying creativity and the rater's judgment processes, making rater-related sources of variation as consequential as the characteristics of the creative products (Wang & Long, 2024). Taken together, these perspectives position subjective

creativity assessment as a rater-mediated measurement system in which understanding, modeling, and evaluating rater behavior is essential for establishing valid, reliable, and fair interpretations of creativity scores.

The psychometric literature has proposed the need to examine subjective creativity assessments through the rater-mediated assessment frameworks (Wang & Long, 2024). These frameworks conceptualize rating data as the product of multiple interacting facets, typically persons/students, items/tasks, and raters, each of which contributes to the variability in observed creativity scores. Traditional approaches to estimating reliability in creativity assessment, most commonly Cronbach's alpha or correlations among judges, are limited in their ability to disentangle these facets or to account for the ordinal nature of rating scales, which calls for a more comprehensive treatment on the psychometric quality of creativity scores (Long & Wang, 2022). Modern measurement models, particularly the Many-Facet Rasch Model (MFRM; Linacre, 1989), offer a more comprehensive examination of the rating quality. MFRM enables concurrent calibration of person creativity, item/task difficulty, and rater scoring severity on an underlying latent scale, and assesses the fit between data and model to detect aberrant scoring behaviors.

The expanding research and practices of subjective creativity assessments has revealed psychometric issues related to rater effects, including differences in scoring severity, inconsistent uses of rating categories, and differential rater functioning across different tasks or student subgroups (Barbot et al., 2019; Hung et al., 2012; Primi et al., 2019; Wang & Long, 2024; Wang et al., 2025), that introduce construct-irrelevant variance into creativity scores. Even when raters share a common conceptual understanding of creativity (typically novelty and appropriateness), their operationalizations of these components may vary. For example, novelty may be interpreted as deviation from personal experience, deviation from the sample responses, or deviation from canonical solutions, and appropriateness may be understood as task-fit, usefulness, or adherence to implicit disciplinary norms (Long, 2014). These findings demonstrate that rater effects are not merely noise but systematic sources of construct-irrelevant variance that may distort score interpretations.

At the same time, human scoring continues to offer irreplaceable strengths in creativity assessment, particularly in domains where nuanced interpretation, contextual sensitivity, and originality-driven inference are essential. Decades of work on the CAT has demonstrated that expert judges, those with deep familiarity with the domain's norms, conventions, and standards, are uniquely capable of recognizing creativity in ways that align with disciplinary expectations (e.g., Amabile, 1982; Kaufman et al., 2008; Myszkowskil & Storme, 2019). Human judges can detect subtle indicators of originality, stylistic sophistication, and appropriateness that are difficult to formalize into rule-based or algorithmic criteria. This sensitivity is especially critical in open-ended tasks, where creativity may manifest through unconventional structure, unexpected thematic connections, or emergent ideas that do not resemble typical exemplars.

Human experts also excel in interpreting contextual cues. Creativity is inherently situated where the same response may be judged differently depending on cultural norms, task framing, audience expectations, or the evolving conventions of a field (Glăveanu, 2014). Although such contextual sensitivity can introduce bias, as experts may hold different interpretations of what matters in a given situation, it is also precisely what enables humans to recognize nuance, originality, and meaning that automated systems may overlook. Expert judges bring tacit knowledge about these contextual layers, allowing them to evaluate not only the product but also its appropriateness and meaningfulness within a broader sociocultural frame (Hennessey & Amabile, 2010). This dual nature of contextual attunement, both a potential source of inconsistency and a unique human strength, underscores the importance of systematic and collaborative efforts that incorporate diverse perspectives when understanding and judging creativity.

In summary, the theory and practice of human scoring in creativity assessment reveal a dual reality. On one hand, human judgment remains essential for capturing the contextual, interpretive and original features of creativity. On the other hand, human scoring is resource-intensive, variable, and susceptible to systematic biases that challenge the psychometric quality of creativity scores. These tensions have motivated the field's gradual movement toward automated scoring paradigms relying on computational modeling. The next section traces this evolution, examining how automated and scoring methods have emerged in response to the limitations of purely human-scored approaches.

3 Development of Automated Scoring Methods in Creativity Assessments

Researchers have been increasingly developing computational approaches to scoring creativity tasks, leading to a surge of interest in automated creativity measures (e.g., Acar et al. 2024; DiStefano et al. 2025; Dumas et al. 2021; Forthmann et al. 2023; Johnson et al. 2023; Organisciak et al., 2023; Patterson et al. 2023, 2024; Yang et al., 2024). Early work focused on the use of latent semantic analysis (LSA; Landauer et al., 1998) and related models that estimate semantic distance based on word co-occurrence patterns in large corpora (Günther et al., 2019). These methods draw on associative theories of creativity (Kenett, 2019; Mednick, 1962), which propose that creative idea involves linking remote concepts. Supporting this view, existing studies show that highly creative individuals tend to possess more flexible, interconnected semantic networks that facilitate access to distant associations (Christensen et al., 2018; Gray et al., 2019; Kenett et al., 2014, 2018; Kenett & Faust, 2019).

Empirical work demonstrates that semantic-distance metrics can capture meaningful variation in creative performance. Prabhakaran et al. (2014) showed that prompting participants to "think creatively" led to responses with greater semantic distance, and

these scores correlated with established creativity measures, providing early validity evidence. Subsequent studies have replicated and extended these findings across different models and tasks, though with mixed results of validity (Forster & Dunbar, 2009; Forthmann et al., 2018, 2020; Harbison & Haarmann, 2014; Hass, 2017a, 2017b; Heinen & Johnson, 2018). For example, Forthmann et al. (2020) indicated that the semantic scores based on LSA are confounded by the degree of elaboration where a longer response tends to have a higher similarity value. Additionally, semantic distance could be useful for scoring creative response that are required to be both novel and appropriate (Heinen & Johnson, 2018).

Recent advances in natural language processing have highlighted important limitations of the semantic-distance models traditionally used in automated creativity scoring. Although LSA and related count- or prediction-based models have been central to prior work, they have been surpassed in most other machine-learning based applications by deep neural architectures that capture richer contextual relationships between words (Wang et al., 2018, 2019). Transformer-based models such as BERT, GPT, T5, and their derivatives (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2018; Raffel et al., 2020; Vaswani et al., 2017; Yang et al., 2019) offer a more sophisticated representation of language by encoding meaning at the level of sentences and passages rather than treating words as isolated units. Organisciak et al. (2023) further introduced a supervised LLM-based approach to scoring originality. By fine-tuning pretrained LLMs on human-rated examples, the model achieves substantial performance gains ($r = 0.76 - 0.81$) over existing semantic-distance methods ($r = 0.12 - 0.26$). Fine-tuning allows the model to leverage broad linguistic knowledge while learning task-specific mappings between responses and originality judgments, enabling improved scalability and stronger generalization to new tasks. These models benefit from their ability to capture complex contextual and syntactic nuances, and can be adapted via supervised fine-tuning, few-shot, or zero-shot learning paradigms to diverse creative tasks and languages (Primi et al., 2025).

As automated approaches to creativity assessment have evolved from early semantic-distance models to contemporary LLMs, questions of transparency, fairness, and interpretability have become increasingly salient (Organisciak et al., 2023; Primi et al., 2025). While LLM-based systems offer substantial gains in accuracy, contextual sensitivity, and scalability, they also inherit biases from their training data and from the human judgments used as ground truth, raising concerns about equity in potential high-stake applications. Although automated models provide advantages in consistency, their complexity makes bias detection and explanation more challenging than with more straightforward semantic-based methods (Barredo Arrieta et al., 2020; Gunning et al., 2019; Rudin, 2019). Supervised LLM approaches nonetheless represent progress, as they approximate the reliability of multiple human raters and can be iteratively improved through corrective feedback. These developments underscore the need for rigorous oversight, including emerging regulatory frameworks such as the European Union's AI Act, which emphasizes transparency and human intervention in educational decision-making (Veale & Borgesius, 2021).

Meanwhile, human scoring remains the gold standard for creativity assessment due to experts' ability to recognize novelty that departs from established patterns. Automated scoring systems, particularly those trained on historical corpora, tend to reward responses that resemble high-scoring exemplars in the training data, which could inadvertently penalize highly original or unconventional ideas. Human judges, by contrast, can appreciate boundary-pushing responses that challenge norms or introduce new ways of thinking. This capacity to detect emergent creativity is central to the construct.

In short, the trajectory of automated scoring reflects a shift from simple textual indices to sophisticated and context-aware AI models that complement rather than replace human expertise, offering promising avenues for scalable and equitable creativity evaluation. Integrating AI-powered scoring with expert judgment positions a human-in-the-loop AI-assisted scoring approach that leverages the strengths of both systems. The next section proposes such a hybrid framework that can preserve the benefits of human insight alongside the efficiency of automated scoring.

4 Human-in-the-Loop AI-Assisted Scoring System

AI technologies, particularly generative methods, are rapidly reshaping creative work by expanding possibilities and reducing manual effort, yet their growing influence also demands responsible use and thoughtful integration into assessment contexts (Anantrasirichai et al., 2026). In large-scale international assessments, AI-powered scoring pipelines offer clear advantages, including improved multilingual scoring, reduced rater fatigue, and near real-time scoring and monitoring for massive responses (Primi et al., 2025). At the same time, interpretability remains essential. Educators, students, and policymakers must understand why a response is judged as more or less creative, especially when decisions carry meaningful consequences. Human-AI collaboration addresses this need by combining the scalability, consistency, and efficiency of automated scoring with the nuanced judgment, contextual sensitivity, and interpretive depth of expert raters. Human-in-the-loop frameworks position AI as an assistive decision-support tool rather than a replacement, enabling iterative collaboration across the assessment lifecycle, from rubric design and calibration to scoring, feedback, and validation.

Creativity is a domain where a "one-size-fits-all" scoring rubric is rarely sufficient. In PISA 2022, the current scoring approach of creativity tasks begins by filtering out inappropriate or irrelevant responses that are not connected to the task context, then appropriate responses that fit one of the predefined conventional themes received partial credit unless they incorporated an innovative way of elaborating, while responses reflecting entirely original themes earned full credit (OECD, 2024). Both steps align well with the capabilities of LLMs and can be made more efficient through automated scoring functions. Rather than relying on LLMs to independently judge the

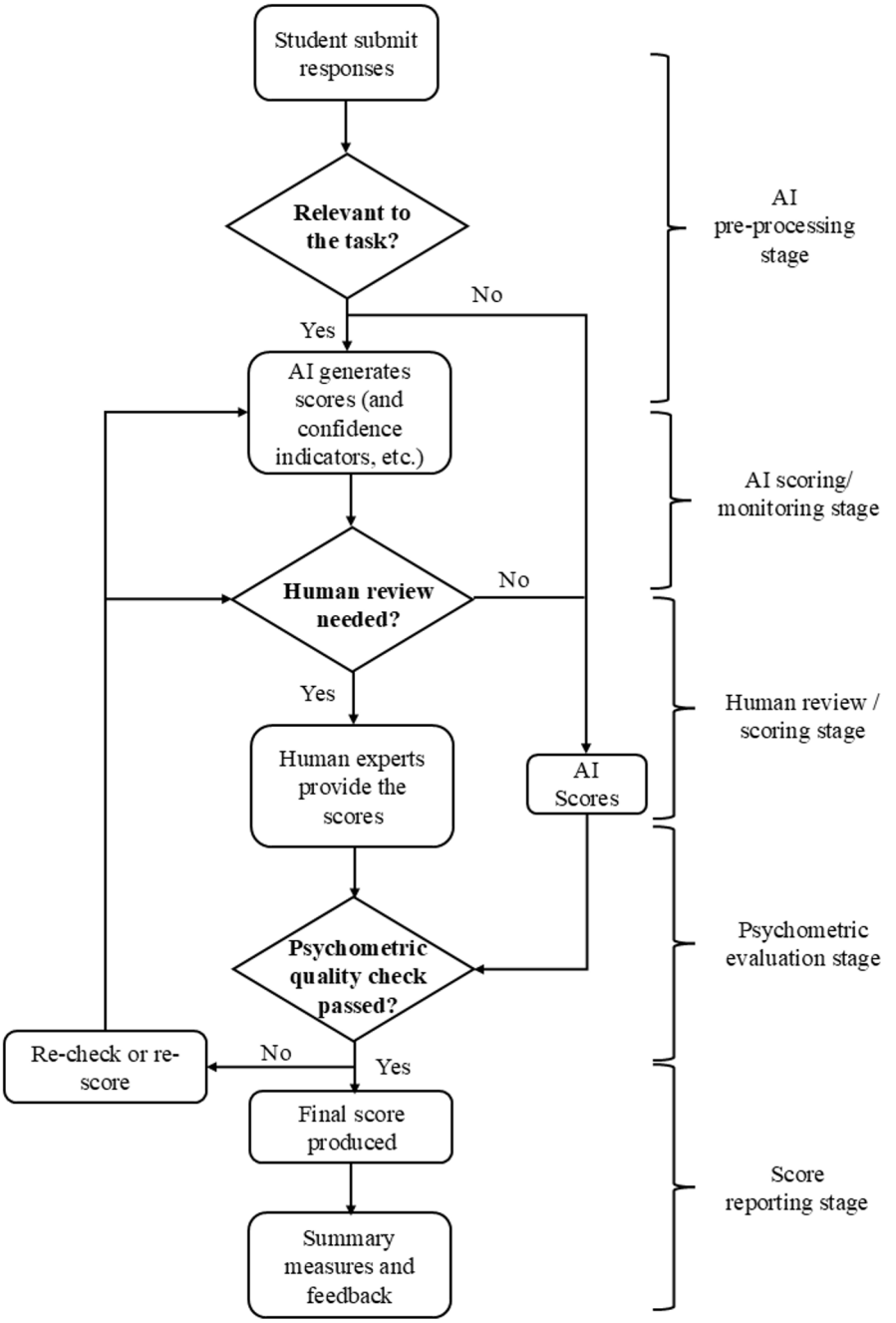
full spectrum of creative quality, a more trustworthy and transparent approach is to let the model handle the identification and labeling of common responses. Human experts can then focus their attention on a more manageable size of highly original responses that require nuanced interpretation. We aim to outline a human-in-the-loop system that automates these stages to improve efficiency while preserving the depth and subtlety of expert judgment.

The human-in-the-loop automated scoring system uses a LLM to handle the initial stages of scoring, including pre-screening and categorizing creative responses. The model can group answers into conventional and unique clusters, allowing those that deviate from typical patterns to be flagged for more nuanced human review. In this workflow, the AI scores the majority of responses automatically, while routing highly original responses or those AI is uncertain about to expert raters for closer examination. Human raters then validate, adjust, or override the AI's judgments to ensure alignment with domain expectations, task requirements, and cultural context. This process can be enhanced by user-friendly interfaces that display AI-generated insights, such as semantic similarity measures, exemplar responses at each score level, confidence estimates, and brief natural-language explanations, helping to strengthen transparency, trust, and ongoing monitoring of model performance.

Figure 1 illustrates the proposed human-in-the-loop scoring system for creativity assessment, which begins when students or participants submit their creative responses to the tasks such as generating creative titles, designs, or solutions to problem-solving prompts. Responses first pass through a pre-processing stage, where the system checks their relevance and appropriateness based on the task context and training examples if any. Any response deemed irrelevant or inappropriate could be assigned a score of zero and may not need to proceed to the subsequent scoring pipeline.

In the core scoring phase, an AI model provides an initial evaluation of each response on designated creativity dimensions such as originality, flexibility, and elaboration. Alongside these scores, the model also generates confidence or uncertainty indicators. The system then applies routing rules to determine which responses require human review. High-confidence and conventional responses will not be flagged for a required human review but will be randomly chosen for quality check. While responses that meet certain conditions are diverted to human raters. These conditions may include low model confidence or unique but relevant content of a creative response. Human experts work within an interface that first presents the raw student responses for scoring and then displays the AI's preliminary insights, enabling them to validate, adjust, or override the model's judgments. This subset of unique or ambiguous responses therefore receives a validated score by human experts for psychometric examination. This set will also be further used as training samples to improve AI performance.

Figure 1. Human-in-the-loop AI-assisted scoring system for creativity assessment



A random sample of responses is also reviewed regularly to support ongoing quality assurance. Drawing on established rater monitoring practices in operational assessment systems, a monitoring procedure can be developed to track and maintain the quality of AI-generated scores. For instance, human experts could work within an interface that displays the original response alongside the AI's scores, confidence estimates, and concise explanations of how those scores were produced. Experts can then confirm the AI's judgments, adjust them when they disagree, or override them entirely while providing justifications that help improve future scoring quality.

Once both AI and human scores are available, the system integrates them through a measurement layer grounded in modern psychometric methods. For example, the use of many-facet Rasch model can treat the AI as an additional rater whose severity and consistency are estimated alongside those of human raters. Automated statistical checks can then identify issues such as centrality, halo effects, or potential bias, prompting additional review or re-scoring when needed. After passing these psychometric quality checks, the final scores are used to generate summary measures and targeted feedback for test developers, educational administrators, learners, and other stakeholders.

The final component of the system is a continuous improvement loop. At this stage, the platform monitors the scoring performance of both AI and human raters, checks for judgmental bias, and iteratively strengthens the scoring quality over time. Human-scored responses, especially those the AI flags as non-conventional or those where its score differs from human judgment, serve as valuable feedback for updating and improving the model performance. Periodic fine-tuning of the prompts for AI, along with adjustments to routing thresholds, helps maintain and enhance model performance as tasks, populations, and curricula evolve. This cyclical process keeps the system accurate, fair, and responsive, while ensuring that human oversight remains central for novel responses and high-stake decisions.

From a measurement perspective, a human-in-the-loop AI-assisted scoring system benefits from modern psychometric models that integrate human and AI ratings within a unified framework. Many-facet Rasch models, for example, can jointly calibrate student creativity, item or task difficulty, human and AI scoring severity along a latent continuum, as well as examine the scoring effects of both human and AI. This approach can further support systematic detection and correction of bias patterns, such as cases where the AI consistently underrates certain response styles or demographic groups relative to expert consensus, thereby strengthening the quality of creativity scores.

From a machine-learning perspective, the system can incorporate active learning and continual learning strategies in which human feedback directly refines the AI model. Human experts can focus on labeling the most unique or high-uncertainty responses identified by the AI, maximizing the impact of each human judgment on model improvement while reducing human workload and associated labor costs. Over iterative cycles, the AI gradually learns to approximate expert judgments on creativity more closely, adapts to new tasks or domains, and reduces reliance on manual scoring, while

still preserving human oversight for unique responses or high-stake decisions. This dynamic feedback loop is particularly important in creativity assessment, where the construct continues to evolve as new forms of expression and problem-solving emerge.

5 Summary

This article traced the evolution of scoring methods within the context of creativity assessment, from traditional human judgment to contemporary AI-powered approaches. Human experts continue to offer irreplaceable strengths with nuanced interpretation, contextual sensitivity, and cultural awareness, yet face persistent challenges related to consistency and bias. Advances in automated scoring, from early semantic-distance to transformer-based LLMs and generative AI, now enable more accurate scoring of multilingual creative responses. Building on this progress, we described a human-in-the-loop scoring framework that integrates AI-assisted functions with expert oversight to enhance validity, interpretability, and accountability. Such a hybrid model offers a scalable and culturally responsive solution for large-scale creativity assessment, supporting equitable evaluation practices while preserving space for unexpected, unconventional, and domain-specific forms of creative expression.

Several considerations must guide the responsible deployment of AI-assisted creativity scoring. Algorithmic bias remains a central concern, as automated systems may inadvertently privilege certain linguistic styles, cultural norms, or response patterns unless datasets are carefully curated, and models transparently documented.

Over-automation could risk narrowing the construct of creativity or overlooking atypical yet meaningful responses, while excessive manual review may undermine the efficiency gains that AI is designed to provide. Creativity is inherently shaped by cultural and contextual factors, making it essential for human-in-the-loop systems to support culturally responsive scoring approaches, localized exemplars, and flexible rubrics that can be adapted to diverse learning environments and disciplinary domains. Ensuring data privacy, maintaining human accountability, and designing workflows that balance efficiency with meaningful oversight are equally critical.

Future work should prioritize transparent and interpretable AI architectures that can highlight salient features of responses and providing clear justifications for scoring decisions. Research is also needed on user-centered interface design that can reduce cognitive load and support higher-order thinking process for human experts' evaluative judgment. Empirical and longitudinal studies are also needed to explore how hybrid scoring systems influence students' creative self-beliefs, risk-taking, and strategy use, and whether AI-generated feedback on dimensions such as originality, flexibility, and elaboration can be leveraged to scaffold creative growth. Additionally, continued exploration of active-learning and continual-learning strategies will further strengthen AI performance as tasks, generations, and cultural expressions evolve.

Disclosure of Interests. Nil

References

1. Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2024). Measuring original thinking in elementary school: Development and validation of a computational psychometric approach. *Journal of Educational Psychology, 116*(6), 953-981. https://doi.org/10.1037/edu0000844
2. Amabile, T. M. (1982). Social psychology of creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology, 43*(5), 997-1013. https://doi.org/10.1037/0022-3514.43.5.997
3. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *The standards for educational and psychological testing.* AERA, APA, and NCME.
4. Anantrasirichai, N., Zhang, F., & Bull, D. (2026). Advances in artificial intelligence: a review for the creative industries. *Artificial Intelligence Review, 59,* 89. https://doi.org/10.1007/s10462-026-11494-w
5. Barbot, B., Hass, R. W., & Reiter-Palmon, R. (2019). Creativity assessment in psychological research: (Re)Setting the standards. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 233-240. https://doi.org/10.1037/aca0000233
6. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58,* 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
7. Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior research methods, 53*(2), 757-780.
8. Carson, S. H. (2006, April). *Creativity and mental illness*. Invitational panel discussion hosted by. Yale's Mind Matters Consortium.
9. Christensen, A. P., Kenett, Y. N., Cotter, K. N., Beaty, R. E., & Silvia, P. J. (2018). Remotely close associations: Openness to experience and semantic memory structure. *European Journal of Personality, 32*(4), 480–492. https://doi.org/10.1002/per.2157
10. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: *Pre-training of deep bidirectional transformers for language understanding.* arXiv. https://arxiv.org/abs/1810.04805
11. DiStefano, P. V., Patterson, J. D., & Beaty, R. E. (2025). Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal, 37*(4), 555-569. https://doi.org/10.1080/10400419.2024.2326343
12. Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts, 15*(4), 645–663. https://doi.org/10.1037/aca0000319
13. Engelhard, G. & Wang, J. (2024). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences (2nd ed.).* Routledge.
14. Forthmann, B., Beaty, R. E., & Johnson, D. R. (2023). Semantic spaces are not created equal—How should we weigh them in the sequel? On composites in automated creativity

scoring. *European Journal of Psychological Assessment, 39*(6), 449–461. https://doi.org/10.1027/1015-5759/a000723

15. Forthmann, B., Paek, S. H., Dumas, D., Barbot, B., & Holling, H. (2020). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology, 90*(3), 683-699. https://doi.org/10.1111/bjep.12325
16. Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2018). Application of latent semantic analysis to divergent thinking is biased by elaboration. *Journal of Creative Behavior, 53*(4), 559–575. https://doi.org/10.1002/jocb.240
17. Forster, E. A., & Dunbar, K. N. (2009). Creativity evaluation through latent semantic analysis. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 602–607). Austin: Cognitive Science Society.
18. Glăveanu, V. P. (2014). *Thinking through creativity and culture: Toward an integrated model.* Transaction Publishers.
19. Gray, K., Anderson, S., Chen, E. E., Kelly, J. M., Christian, M. S., Patrick, J., … Lewis, K. (2019). "Forward flow": A new measure to quantify free thought and predict creativity. *American Psychologist, 74*(5), 539–554. https://doi.org/10.1037/amp0000391
20. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120
21. Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372
22. Harbison, J. I., & Haarmann, H. (2014). Automated scoring of originality using semantic representations. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)* (pp. 2327–2332). Quebec City: Cognitive Science Society.
23. Hass, R. W. (2017a). Semantic search during divergent thinking. *Cognition, 166*, 344–357. https://doi.org/10.1016/j.cognition.2017.05.039
24. Hass, R. W. (2017b). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory and Cognition, 45*(2), 233–244. https://doi.org/10.3758/s13421-016-0659-y
25. Heinen, D. J. P., & Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts, 12*(2), 144–156. https://doi.org/10.1037/aca0000125
26. Hennessey, B. A., & Amabile, T. M. (2010). Creativity. *Annual Review of Psychology, 61*, 569–598. https://doi.org/10.1146/annurev.psych.093008.100416
27. Hong Kong Education Bureau (2001). *Ongoing renewal of the school curriculum.* Retrieved from https://www.edb.gov.hk/en/curriculum-development/renewal/index.html
28. Hung, S.-P., Chen, P.-H., & Chen, H.-C. (2012). Improving creativity performance assessment: A rater effect examination with Many Facet Rasch Model. *Creativity Research Journal, 24*(4), 345-357. https://doi.org/10.1080/10400419.2012.730331
29. Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., ... & Beaty, R. E. (2023). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior research methods*, *55*(7), 3726-3759.
30. Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. John Wiley & Sons.

31. Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences, 27*, 11–16. https://doi.org/10.1016/j.cobeha.2018.08.010
32. Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience, 8*, 407. https://doi.org/10.3389/fnhum.2014.00407
33. Kenett, Y. N. & Faust, M. (2019). A semantic network cartography of the creative mind. *Trends in Cognitive Sciences, 23*(4), 271–274. https://doi.org/10.1016/j.tics.2019.01.007
34. Kenett, Y. N., Levy, O., Kenett, D. Y., Stanley, H. E., Faust, M., & Havlin, S. (2018). Flexibility of thought in high creative individuals represented by percolation analysis. *Proceedings of the National Academy of Sciences of the United States of America, 115*(5), 867–872. https://doi.org/10.1073/pnas.1717362115
35. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284. https://doi.org/10.1080/01638539809545028
36. Linacre, J. M. (1989). *Many Facet Rasch measurement.* Chicago: University of Chicago [Unpublished Ph.D. dissertation].
37. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv. http://arxiv.org/abs/1907.11692.
38. Long, H. (2014). More than appropriateness and novelty: Judges' criteria of assessing creative products in science tasks. *Thinking Skills and Creativity, 13,* 183-194. https://doi.org/10.1016/j.tsc.2014.05.002
39. Long, H., & Plucker, J. A. (2015). Assessing creative thinking: Practical applications. In R. Wegerif, J. Kaufman, & L. Li (Eds.), *The Routledge international handbook of research on teaching thinking* (pp. 315-329). Routledge.
40. Long, H. & Wang, J. (2022). Dissecting reliability and validity evidence of subjective creativity assessment: A literature review. *Educational Psychology Review*, *34*, 1399-1443.
41. Mednick, S. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220–232. https://doi.org/10.1037/h0048850
42. Myszkowskil, N. & Storme, M. (2019). Judge response theory? A call to upgrade our psychometrical account of creativity judgments. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 167-175. https://doi.org/10.1037/aca0000225
43. OECD. (2024). *PISA 2022 Results (Volume III): Creative Minds, Creative Schools.* Paris: OECD Publishing. https://doi.org/10.1787/765ee8c2-en
44. OECD. (2019). *OECD learning compass 2030.* Paris: OECD Publishing. https://www.oecd.org/en/data/tools/oecd-learning-compass-2030.html
45. Organisciak, P., Acar, S., Dumas, D., & 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
46. Patterson, J. D., Merseal, H. M., Johnson, D. R., Agnoli, S., Baas, M., Baker, B. S., Barbot, B., Benedek, M., Borhani, K., Chen, Q., Christensen, J. F., Corazza, G. E., Forthmann, B., Karwowski, M., Kazemian, N., Kreisberg-Nitzav, A., Kenett, Y. N., Link, A., Lubart, T., . . . Beaty, R. E. (2023). Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts, 17*(4), 495–507. https://doi.org/10.1037/aca0000618
47. Patterson, J. D., Barbot, B., Lloyd-Cox, J., & Beaty, R. E. (2024). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior research methods*, *56*(4), 3619-3636. https://doi.org/10.3758/s13428-023-02258-3

48. Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity (pp. 48-73)*. Cambridge University Press. https://doi.org/10.1017/CBO9780511763205.005
49. Prabhakaran, R., Green, A. E., & Gray, J. R. (2014). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods, 46*(3), 641–659. https://doi.org/10.3758/s13428-013-0401-7
50. Primi, R., Beaty, R. E., Benedek, M., Dumas, D., Organisciak, P., Patterson, J. D., Calico, T. & Piacentini, M. (2025). AI-powered scoring for creative thinking: Methods and challenges in PISA assessment. *The Journal of Creative Behavior*, *60*(1), e70082.
51. Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts, 13*(*2*), 176-186. https://doi.org/10.1037/aca0000230
52. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
53. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67. https://jmlr.org/papers/v21/20-074.html
54. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
55. Singapore Ministry of Education (2023). *21st century competencies framework.* Retrieved from https://www.moe.gov.sg/education-in-sg/21st-century-competencies
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. arXiv. http://arxiv.org/abs/1706.03762
57. Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International, 22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402
58. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems, 32*.
59. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353–355). https://doi.org/10.18653/v1/W18-5446
60. Wang, J., & Long, H. (2024). Reexamining subjective creativity assessments in science tasks: An application of the rater-mediated assessment framework and many-facet Rasch model. *Psychology of Aesthetics, Creativity, and the Arts, 18*(4), 536-549. https://doi.org/10.1037/aca0000470
61. Wang, J., Yang, S., & Long, H. (2025). Behind the scores: Unraveling rater judgment in subjective creativity assessments. *Psychology of Aesthetics, Creativity, and the Arts.* Advance online publication. https://doi.org/10.1037/aca0000771
62. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems, 32*.
63. Yang, T., Zhang, Q., Sun, Z., & Hou, Y. (2024). Automatic assessment of divergent thinking in Chinese language with TransDis: A transformer-based language model approach. *Behavior Research Methods*, *56*(6), 5798-5819. https://doi.org/10.3758/s13428-023-02313-z

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

