

Can AI Assess Creativity? The Validity of Machine Learning Models in Scoring Open-Ended Art and Design Tasks

Bei TU
Jilin University of Arts

Abstract

This paper critically examines the validity of Machine Learning (ML) models in assessing human creativity within open-ended art and design tasks. As AI tools increasingly permeate educational assessment, they offer scalability but face significant challenges in capturing the subjective and nuanced nature of creativity. Through an analysis of current ML paradigms—such as Convolutional Neural Networks (CNNs) and computational aesthetics—the study reveals that while AI achieves high accuracy in replicating human scores, it often relies on surface-level pattern recognition rather than a deep understanding of context, intent, or “appropriateness.” The paper highlights critical limitations, including algorithmic bias, the “black box” opacity of deep learning, and the risk of “teaching to the algorithm.” Ultimately, it argues that AI lacks the construct validity to fully replace human evaluators and should instead be positioned as an assistive tool to enhance human judgment and pedagogical feedback.

Keywords Artificial Intelligence; Creativity Assessment; Art and Design Education; Validity; Human-AI Collaboration

1 Introduction: The Intersection of AI, Creativity, and Arts Education

1.1 The Urgency of Cultivating Creativity in Contemporary Education

In the information age, creative thinking is paramount for professional success and has been identified by employers as a rapidly growing and essential skill. While creative thinking was once viewed as an innate talent that could not be acquired, decades of empirical research have dismantled such myths. Today, robust cognitive models describe how the creative process unfolds, and extensive research confirms that creative thinking possesses pedagogical plasticity. It can be developed through deliberate practice and the improvement of environmental conditions. Consequently, the effective assessment of creativity is vital for the educational system, laying the groundwork for the application of Artificial Intelligence (AI) in this domain.

1.2 The Emerging Role of AI in Educational Assessment

Artificial Intelligence, particularly Machine Learning (ML) and Deep Learning (DL), is being widely explored for various educational tasks, including assessment. AI systems aim to construct computer systems capable of intelligent behavior and assisting in solving practical problems. In recent years, driven by robust computational power and advanced machine learning methods, AI applications have surged in fields such as computer vision and natural language processing. Its potential to enhance efficiency and scalability makes it an attractive option for assessing complex constructs like creativity.

1.3 Articulating the Core Question: The Validity of ML Models in Assessing Creativity in Open-Ended Art and Design Tasks

The central research question of this paper asks to what extent current AI, specifically machine learning models, can effectively and reliably assess human creativity in open-ended art and design tasks. Although AI shows promise in this area, the inherent complexity of creativity assessment, particularly in subjective and open-ended domains like art and design, presents significant challenges. This paper aims to deeply explore the capabilities and limitations of current AI technologies and examine the validity of their assessment results. The push for creativity cultivation in education is converging with technological advancements in AI assessment. This intersection brings both opportunities, such as the large-scale promotion of creativity education, and tensions regarding whether machines can truly comprehend human creativity. Creativity involves complex processes and nuanced human traits, whereas AI excels at pattern recognition based on data, naturally leading to the core inquiry regarding its effective assessment of creativity.

Furthermore, if AI intervenes in assessment, the definition of assessment itself may need re-examination. It remains to be seen whether AI's role is merely to replicate human judgment or to provide new perspectives on creative works that human evaluators might overlook. Traditional art assessment relies on the subjective judgment of human experts, and AI models are typically trained on these human judgments. If AI merely duplicates human judgment, its validity depends on the consistency of its mimicry. However, AI might be able to identify patterns or features that are imperceptible to human evaluators. This raises a deeper question of whether AI assessment should aim for imitation by validating against human consensus or for providing new insights that might challenge established creativity norms. This has profound implications for how validity is defined and measured in AI-driven creativity assessment.

2 Defining and Measuring Creativity in Art and Design

2.1 Theoretical Foundations: Process vs. Product, Novelty vs. Appropriateness

Creativity is a multifaceted concept that requires distinguishing between creative thinking as a process and creativity as judged by the novelty and utility or appropriateness of a product. Creative thinking is an iterative process where individuals generate and manipulate ideas, test and refine these ideas through critical analysis and evaluation, and communicate them effectively to solve problems or advance knowledge in novel ways. Creative thinkers possess sensitivity to problems, fluency of thought, flexibility of mind, divergent and convergent thinking capabilities, the ability to redefine familiar objects and concepts, and the ability to effectively communicate new ideas. Broadly, creativity is typically defined as the ability to produce work that is both novel and appropriate.

2.2 Established Methods for Creativity Assessment in Art and Design

Common methods for human assessment of creativity include the Consensual Assessment Technique (CAT), Rubrics, and Standardized Tests. The Consensual Assessment Technique relies on the collective judgment of experts within a domain to evaluate the creativity of a product. Experts score works based on their own definitions of creativity, and reliability can be secured; for instance, studies involving professional artist judges evaluating collages have achieved high reliability. Rubrics are structured scoring tools that can be analytical, assessing specific sub-criteria, or holistic, assessing overall quality. In the arts, rubrics typically attempt to balance objective criteria like technical skill with subjective criteria such as originality and expression. These guides can observe various learning stages from sketching to self-assessment. While less common in open-ended art domains, standardized tests like Guilford's tests or the Torrance Tests of Creative Thinking (TTCT) focus on components of divergent thinking such as fluency, originality, flexibility, and elaboration.

2.3 Challenges Inherent in Human-Centric Assessment: Subjectivity, Scalability, and Consistency

Despite the richness of human assessment, it faces challenges. Subjectivity is inherent, and while CAT aims for consensus, individual biases may still play a role. Assessing artworks typically involves a mix of objective information and subjective opinion, making it difficult for educators to evaluate student learning outcomes. Scalability is a major issue as expert assessment is time-consuming and expensive. Consistency between different raters, or even the same rater over time, can vary. The perception that creativity is hard to measure remains prevalent, partly due to definitional

issues and practical difficulties. The immense time and human resource demands of manual scoring drive the shift toward AI assessment.

The definition of creativity as novelty and appropriateness introduces a fundamental tension for AI. Novelty might be identified computationally by comparing against existing datasets, but appropriateness is highly context-dependent and often tied to human values, cultural norms, or specific task constraints that AI may not fully grasp. AI might recognize patterns and deviations to identify novelty, but appropriateness implies a judgment of fit for a goal, context, or audience, often involving tacit knowledge and cultural understanding. Therefore, AI models, especially those trained on decontextualized data, may struggle to meaningfully assess appropriateness, potentially leading to outputs that are novel but meaningless or irrelevant. This suggests AI may be better at assessing certain aspects of novelty than the holistic concept of appropriateness, affecting its overall validity in creativity assessment.

Moreover, the focus of creativity research has shifted from eminent creators or Big-C creativity to everyday creativity or Little-c creativity and the personal learning of Mini-c creativity within educational contexts. This makes assessment more relevant in education and potentially more suitable for AI, but risks oversimplifying the construct. Historically, research focused on genius, but education now emphasizes Little-c and Mini-c, which involve everyday problem-solving and personal insights. These smaller forms of creativity may involve more observable processes and products with less historical impact, relating more to personal growth. This narrower scope might make it easier to define operational criteria for AI assessment, such as evaluating specific skills in a design task. However, it also risks reducing creativity to easily measurable components, potentially overlooking the holistic, transformative aspects even in Little-c or Mini-c expressions. AI might then assess adherence to criteria rather than true creative breakthroughs.

The reliability of human expert judgment in the Consensual Assessment Technique sets a benchmark. If AI is to be considered valid, it should ideally meet or exceed this level of inter-rater reliability with human experts, while also demonstrating that it measures the same underlying construct rather than merely correlating on surface features. CAT provides a gold standard for human assessment of creative products with quantifiable reliability. For AI assessment to be deemed valid, its scores must correlate highly with these expert human judgments. However, high correlation alone is insufficient. AI might achieve correlation by latching onto surface features that happen to co-occur with human judgments of creativity without understanding the deeper reasons behind those judgments. Thus, validity also requires evidence that the AI is sensitive to the core dimensions of creativity as perceived by experts, such as originality, technical skill, emotional expression, and conceptual depth. This is significantly harder than simple correlational alignment.

3 Machine Learning Paradigms for Assessing Art and Design Creativity

3.1 Overview of Employed Machine Learning Techniques

The types of machine learning models used to assess creativity are diverse. Traditional Machine Learning methods, such as linear or logistic regression, Support Vector Machines (SVM), and decision trees, typically rely on handcrafted features. For example, research has used logistic regression to identify key compositional elements in AI-generated Neoplasticism works, and SVMs to model feature data, EEG data, contextual data, and subjective aesthetic evaluation data. Deep Learning (DL), particularly Convolutional Neural Networks (CNN) for image tasks, and potentially Recurrent Neural Networks (RNN) or Transformers for sequential data or art-related text, allows for automated feature learning from raw data. DL models have rapidly gained popularity due to their general-purpose approach capable of performing the work of other ML methods with high accuracy. Additionally, Natural Language Processing (NLP) and Computer Vision utilize ML and DL. NLP analyzes textual descriptions or reviews, while computer vision analyzes visual artworks. These tools can identify unique ideas, creative language use, and artistic features based on patterns learned from large datasets.

3.2 Computational Aesthetics: Quantifying Artistic Features for Assessment

AI attempts to understand art by extracting and quantifying features. This involves translating abstract aesthetic concepts like style, texture, composition, and color harmony into measurable quantities such as empty space, grayscale histograms, Gabor filters, Gray Level Co-occurrence Matrices (GLCM), and Local Binary Patterns (LBP). Computational aesthetics is defined as computational methods capable of making applicable aesthetic decisions in a manner similar to humans, with the advantage of being able to quantify various abstract aesthetic features.

3.3 Examples and Case Studies of AI Models

Specific instances of AI application in art and design creativity assessment include Visual Arts, where research indicates that CNNs can be used to assess the creativity of paintings, achieving high accuracy compared to human raters. One study developed a CNN model trained on 600 paintings created by professional artists and children, achieving approximately 90% accuracy in assessing painting creativity with speeds far exceeding human raters. Other research applies CNNs to assess responses to graphic creativity tests like the TCT-DP or creates automated drawing assessment platforms to extract visual creativity scores from simple sketches. Research also explores aesthetic evaluation by AI systems regarding the replication of core Neoplasticism principles. In Music, AI is applied in composition, performance, improvisation, accompaniment, and various aspects of the music industry. While not directly assessment, this shows the deep integration of AI with creative domains. In Poetry and Digital Art, AI generates works and develops corresponding evaluation frameworks, such as adaptations of the Turing Test, the FACE model, and ProFTAP.

3.4 Summary Table of Representative ML Models for Art/Design Creativity Assessment

To clarify the current landscape, the following table summarizes representative machine learning models and research.

Table 1: Representative ML Models for Art/Design Creativity Assessment						
Model/Study	AI Technique	Task/Domain		Performance Metrics		Dimensions Assessed
Cropley & Marrone (2022)	CNN	TCT-DP	Student Response Evaluation	Not aimed at	specified; assessment	Graphic Creativity
Patterson et al. (2023)	Automated Drawing Platform	Visual creativity in simple sketches		Extracts score	creativity	Visual Creativity
Painting Creativity Study	CNN	Creativity assessment of 600 paintings		≈90% accuracy; faster than humans		Overall Creativity Score
Neoplasticism Study	Logistic Regression, Midjourney	Aesthetic evaluation of AI Neoplasticism art		Identifying key compositional elements		Geometric forms, balance, color harmony
Computational Aesthetics	SVM, Feature Extraction	Aesthetic assessment of art images		0.76866 accuracy (Subject + Context)		Composition, tone, texture, context

3.5 Further Reflections on Machine Learning Paradigms

The impressive success of models like the CNN, achieving approximately 90% accuracy, invites questions about what exactly the AI is learning. It is unclear whether it learns deep principles of creativity or merely pattern-matches stylistic features prevalent in its training data of professional versus child art. CNNs are powerful at image feature extraction and classification. Training data containing works by professionals and children likely contains distinct stylistic and skill-level differences. The AI might be learning to associate certain visual features, such as color palettes, brushstroke styles, or compositional elements common in "good" art in the dataset, with the higher creativity scores assigned by humans during training. This does not necessarily mean the AI understands creativity in a human sense, such as intent, novelty relative to broader art history, or emotional impact. It may be performing complex style classification or quality approximation based on learned correlations. Thus, high accuracy might mask a superficial understanding, a critical point for assessing validity.

The reliance of computational aesthetics on quantifiable visual properties may fundamentally limit AI's capacity to assess conceptual or process-oriented creativity, which are significant in contemporary art and design. Computational aesthetics aims to quantify features like composition, tone, and texture. These are primarily formal, visual

attributes of the final product. However, much contemporary art and design emphasizes concepts, processes, artist intent, or social commentary, which may not be directly or fully readable from the visual features of the final product alone. If AI primarily analyzes these quantifiable visual properties, it may struggle to assess works where the creative value lies in non-visual or less tangible aspects. This suggests a potential mismatch between current computational aesthetic approaches and the assessment needs of certain types of open-ended creative tasks.

Distinguishing between AI assessing human-created art and AI assessing AI-generated art is important. The latter might be a simpler task if the AI generator and AI evaluator share underlying architectural biases or operate on similar principles. Research focuses on AI assessing human art, while other studies focus on AI assessing AI-generated art. If an AI system generates art based on certain learned algorithmic and aesthetic principles, another AI system might find it easier to evaluate that art based on similar or identical principles, especially if both are trained on overlapping datasets or aesthetic theories. Assessing human creativity involves understanding a much wider and more unpredictable range of intents, techniques, and cultural influences. Therefore, success in assessing AI-generated art may not directly translate to equivalent success or validity in assessing the more diverse and complex realm of human creativity.

4 Examining the Validity of AI-Driven Creativity Assessment

4.1 Conceptualizing Validity in AI Creativity Assessment

In this context, different facets of validity relevant to creativity assessment need defining. Construct Validity asks if the AI measures the theoretical construct of "creativity" as understood in psychology and art education, such as novelty, appropriateness, and process elements. Criterion Validity asks if the AI's scores correlate with other established measures of creativity, such as expert human judgments via CAT, divergent thinking test scores, or future creative achievements. Content Validity asks if the AI assessment covers all relevant content and skills of the specific creative task. Although existing literature may not explicitly define these validity types, the entire research question revolves around validity.

4.2 Empirical Evidence: Accuracy, Reliability, and Correlation with Human Experts

Reviewing studies reporting these metrics is crucial. For instance, CNN models have achieved high accuracy in assessing painting creativity. The key is to critically examine these claims by asking what "ground truth" the accuracy is relative to. Establishing how human scoring was conducted and its reliability is essential. The reliability of human judges via CAT provides a benchmark for comparison.

4.3 AI's Capacity to Assess Established Creativity Dimensions

It is necessary to assess how AI performs on specific aspects of creativity. Originality, or the novelty of ideas, has historically been weak to score prior to generative AI, but generative AI shows potential. Fluency, the quantity of ideas, can be improved by generative AI assessment. Flexibility, the diversity of ideas or categories, is difficult to score automatically because open-ended text responses require semantic classification. Elaboration, the detail and refinement of ideas, is mentioned as an element AI can assess. Studies are critical for understanding AI's capabilities in assessing these dimensions, discussing the challenges and progress in AI scoring of originality, fluency, and flexibility.

4.4 Critical Limitations: Nuance, Context, and Subjectivity

AI models typically struggle with nuance, context, and subjectivity. Nuance involves the subtleties of artistic expression, irony, metaphor, and emotional depth. Context involves the historical, cultural, and personal context of the artwork and artist. Research notes that context influences human aesthetic ratings, highlighting that AI might be blind to context unless explicitly programmed. Subjectivity involves the personal and interpretative aspects inherent in creative appreciation that go beyond quantifiable features. Studies highlight "subjective opinions" in art assessment and note that "the subjectivity of creativity means complete automation may never fully replace human judgment."

High accuracy, as seen in some CNN models, can be misleading if the "ground truth" itself is based on simplified or easily quantifiable aspects of creativity, or if the task is narrow. Validity of AI assessment is tethered to the richness of the human assessment used to benchmark it. Accuracy is measured against ground truth labels. If human raters focused on easier-to-code features when creating these labels for scalability or simplicity, the AI might learn

those features effectively. However, this means the AI is accurately mimicking a potentially simplified version of human assessment rather than necessarily assessing the full, complex construct of creativity. Thus, the validity of AI assessment depends critically on the depth, nuance, and comprehensiveness of the human judgments used to train and validate it. High accuracy scores against superficial human benchmarks do not strongly prove it assesses true creativity.

AI may be better at assessing convergent aspects of creativity, such as technical skill or adherence to stylistic principles like Neoplasticism, than divergent aspects like radical originality or rule-breaking, especially if trained on datasets reflecting established norms. Creative thinking involves divergent and convergent processes. AI models learn from data patterns. If training data reflects established styles or successful examples within those styles, the AI will learn to recognize and reward adherence to those norms. True divergent creativity, which breaks norms or introduces entirely new paradigms, might be penalized or misjudged by AI trained on historical data as it appears as an outlier or error. This implies a bias towards integrative tendencies in AI, potentially stifling the assessment and encouragement of more radical forms of creativity.

The challenge of automatically scoring flexibility due to the need for semantic classification of open-ended responses points to a deeper limitation: AI's current struggle with abstract conceptualization and understanding diverse meanings in truly open domains, which are hallmarks of art and design. Flexibility in divergent thinking involves shifting between different conceptual categories. Assessing this requires identifying those categories from open-ended responses. It is noted this is hard for AI because it requires generating semantic categories that can cover the range of all possible answers. This reflects a broader challenge for AI: moving beyond surface patterns to grasp deeper semantic structures and the ability to form and understand novel conceptual connections, especially in fluid and ambiguous domains like art. If AI struggles with this fundamental aspect of creative cognition, its ability to validly assess the process of creative thinking is severely hampered.

5 Navigating Challenges and Ethical Imperatives

5.1 Algorithmic Bias and Fairness: The Impact of Non-Representative Datasets

This is a major concern. AI models learn from data, and if that data is biased, the AI will perpetuate or amplify those biases. Cultural and stylistic bias is significant; AI models trained primarily on Western art may poorly evaluate or misunderstand art from other cultures. It is noted that AI music models are often trained on dominant genres like Western classical. Arguments suggest art datasets often perpetuate a Western-centric art history canon. Warnings explicitly state that representational bias in sources (often Western-centric) can be reproduced by generative AI, undermining inclusion. Research on Visual Language Models (VLMs) and art history shows differential performance and bias in author/style prediction reflecting dataset characteristics. Demographic bias is also a risk, where datasets may overrepresent certain groups, leading to biased assessment of underrepresented groups.

5.2 Pedagogical Implications: "Teaching to the Algorithm"

If AI assessment becomes prevalent, students and educators might focus on optimizing for what the AI rewards, potentially narrowing the scope of creative expression and discouraging risk-taking or non-traditional approaches. Warnings exist against AI "stifling human creativity by providing overly prescriptive solutions" and that "reliance on AI tools could lead to a diminution of human original thinking."

5.3 Opacity of "Black Box" Models and the Need for Explainable AI

Many advanced AI models, especially deep learning ones, are "Black Boxes," meaning their decision-making processes are opaque. For educational feedback to be meaningful, students need to understand why their work received a certain score. Lack of interpretability hinders this. The need for "transparency" is noted as an ethical consideration. Discussions highlight the "explainability" problem with generative AI, noting it has "no clear source of information for its outputs."

5.4 Frameworks for Responsible and Ethical Integration

Proposing guidelines for ethical use is crucial. This includes developing generative AI literacy, revising assessment designs to integrate critical thinking with AI tools, establishing transparent policies for fairness and accountability,

and ensuring AI complements rather than replaces human judgment. Recommendations for institutions include AI literacy courses and revised assessment methods. Emphasis is placed on a "balanced approach that combines human creativity with AI capabilities," designing AI to "enhance rather than limit human agency," and maintaining teacher/student autonomy.

The digital divide and the resource-intensiveness of deep learning imply that if high-quality AI creativity assessment tools are developed, they might only be accessible to well-resourced institutions, exacerbating existing educational inequalities in art and design. Deep learning relies on extremely powerful computing resources, creating barriers to access. The digital divide is listed as a broader ethical issue with generative AI. Developing and deploying sophisticated AI for creativity assessment likely requires significant computational power, technical expertise, and potentially expensive software. Under-resourced institutions may not have access. Consequently, if such AI tools become markers of "good" art education, students in resource-poor environments could be disadvantaged.

The issue of bias is not just about unfair grading; it can proactively shape future creative output. If AI trained on biased data consistently rewards certain styles, it could homogenize creative expression. AI assessment tools based on biased models may reflect those biases in their scoring. If students are assessed by such AI, they may consciously or unconsciously create works that conform to the AI's rewarded criteria to get better grades. This creates a feedback loop where AI bias influences creative production, potentially narrowing the range of styles and themes perceived as "creative."

Calling for "human oversight" or AI as a "supportive tool" is essential but complex in execution. Defining effective human-AI collaboration in assessment is challenging. Many texts advocate for AI as an assistant requiring human oversight. However, mechanisms for this are tricky. If AI provides a score and a human disagrees, whose judgment prevails? In mass assessment scenarios where AI's appeal is efficiency, how much oversight is realistic? Defining clear protocols for human-AI interaction in assessment, including dispute resolution and accountability frameworks, is a key next step beyond simply stating the need for human involvement.

6 Future Trajectories: Advancing AI for Robust and Ethical Creativity Assessment

6.1 Developmental Prospects: Multimodal AI and Cultural Awareness

Frontier AI research may improve creativity assessment. Multimodal AI (Visual Language Models - VLM) can process information from multiple modalities like image and text. This could allow AI to assess artwork in conjunction with artist statements. Benchmarking of VLMs for cultural understanding is underway. Advanced feature extraction is moving beyond basic visual attributes to more abstract or semantically rich features. Cultural awareness efforts aim to create more diverse datasets and models sensitive to cultural nuances.

6.2 Positioning AI as an Assistive Tool

Reinforcing the idea of AI as a supportive partner is vital. AI can handle initial screening, identify patterns for human review, provide formative feedback on technical aspects, or offer diverse examples for inspiration, leaving nuanced and holistic judgments to human educators. Conclusions suggest that AI has immense potential when used as a supportive tool alongside teacher judgment. It is emphasized that AI should "enhance rather than limit human agency."

6.3 Interdisciplinary Research and Policy Recommendations

There is a call for collaboration between AI researchers, educators, artists, psychologists, and policymakers. Needs include developing diverse, representative, and ethically sourced training datasets; researching explainable AI models for creative assessment; establishing clear ethical guidelines; and conducting longitudinal studies on the long-term impact of AI assessment. Advocacy for interdisciplinary dialogue is strong.

While Multimodal AI offers promise for richer assessment, it introduces new complexities in data alignment, model architecture, and potential for new types of bias arising from modal interactions. Training such models requires large, well-aligned multimodal datasets, which is a challenge for diverse art forms. Bias can stem from spurious correlations learned between certain visual styles and descriptive language. Thus, multimodal AI presents its own research and ethical challenges.

Pushing for "cultural awareness" in AI is critical, but culture is dynamic and contested. Simply adding more data may not be enough; AI may need mechanisms to understand cultural context and relativity rather than just recognizing more symbols. Efforts highlight attempts to improve multicultural understanding in VLMs. However, true cultural awareness involves understanding shared meanings, history, and values. AI might classify a mask as "African" without understanding its ritual significance. This suggests current "cultural awareness" might be more about broader representation in datasets than deep cultural understanding.

The ideal of AI as an "assistive tool" requires a shift in AI development focus from autonomous decision-making to designing for effective human-AI collaboration. This means prioritizing interpretability, uncertainty estimation, and user control. Key features would include explaining why a score was given, quantifying confidence to know when to defer to humans, and allowing educators to adapt the AI to specific pedagogical goals. This design philosophy differs from simply maximizing accuracy on benchmark datasets.

7 Conclusion: Synthesizing Perspectives on AI and Creativity Assessment

7.1 Recapitulation of Current Capabilities and Limitations

AI shows promise in specific, often quantifiable aspects of art and design assessment, such as the accuracy of painting assessment, but faces challenges with the holistic, nuanced, context-dependent, and deeply subjective dimensions of creativity.

7.2 Reiterating the Centrality of Validity and Ethics

Technical capabilities must be balanced with rigorous validation against meaningful criteria, a steadfast focus on ethical implications like bias and fairness, and a commitment to centering human values and goals when integrating AI in education.

7.3 Concluding Thoughts on the Symbiosis of AI and Human Creativity

AI is neither a panacea for creativity assessment nor necessarily a dire threat. Its future role depends on thoughtful development and implementation. Calls are made for re-evaluating the creative process to acknowledge the interdependence of technical innovation and human agency, while other perspectives envision a balanced approach combining human creativity with AI capabilities. The goal should be a synergistic relationship where AI enhances the human capacity to cultivate and understand creativity.

To Cite This Article Bei TU. (2025). Can AI Assess Creativity? The Validity of Machine Learning Models in Scoring Open-Ended Art and Design Tasks. *Artificial Intelligence Education Studies*, 1(4), 22–30. <https://doi.org/10.6914/aiese.010403>

Reference

- [1] Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997–1013.
- [2] 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.
- [3] Beghetto, R. A., & Kaufman, J. C. (2007). Toward a broader conception of creativity: A case for "mini-c" creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 1(2), 73–79.
- [4] Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd ed.). Routledge.
- [5] Cetinic, E., & She, J. (2022). Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(2), 1–22.
- [6] Colton, S., & Wiggins, G. A. (2012). Computational creativity: The final frontier? *Proceedings of the 20th European Conference on Artificial Intelligence*, 21–26.
- [7] Dumas, D., Organisciak, P., & Doherty, M. (2020). 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.

- [8] Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. *Proceedings of the 8th International Conference on Computational Creativity (ICCC)*, 96–103.
- [9] Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
- [10] Guzik, E. E., Byrge, C. Christian., & Gildea, K. M. (2023). The creativity of large language models: An evaluation of GPT-4 on the Torrance Tests of Creative Thinking. *Thinking Skills and Creativity*, 50, 101416.
- [11] Hertzmann, A. (2018). Can computers create art? *Arts*, 7(2), 18.
- [12] Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four c model of creativity. *Review of General Psychology*, 13(1), 1–12.
- [13] Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.
- [14] Marrone, R., Cropley, D. H., & Chong, S. W. (2024). Automated scoring of divergent thinking: A systematic review of machine learning applications. *Computers & Education: Artificial Intelligence*, 6, 100205.
- [15] Organisciak, P., Acar, S., & Dumas, D. (2023). Beyond semantic distance: Automated scoring of divergent thinking with large language models. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication.
- [16] Rafner, J., Biskjaer, M. M., & Sherson, J. (2022). Hybrid intelligence in creativity: A review and framework. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 36, e6.
- [17] Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, 24(1), 92–96.
- [18] Shaik, T., Tao, X., Li, Y., Dann, C., Xie, J., Li, Y., & Galligan, L. (2022). A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access*, 10, 56720–56739.
- [19] Srinivasan, R., & Uchida, Y. (2020). Biased algorithms and the evaluation of creative work. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.
- [20] Zhang, Z., Singh, S., & Gero, J. S. (2021). Deep learning in the assessment of design creativity: A review. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 35(3), 286–305.

Editor Xin HUANG cpcl@cpcl.hk