Design Science

"Who" designs better? A competition among human, artificial intelligence and human–AI collaboration

Khanh Hoa T. Vo

Eskenazi School of Art, Architecture + Design, Indiana University Bloomington, Bloomington, IN, USA

Abstract

This research examines whether a machine, specifically artificial intelligence (AI), can be creative by comparing design solutions for a practical competition – a light fixture for a pediatric waiting room – among AI, collaboration efforts and a human designer. Amazon Mechanical Turk and Prolific workers observed the design solutions throughout the design process, from sketches (*S*) to three-dimensional renderings (*3D*) to fully developed models in virtual waiting rooms (*VR*). Using the well-established Creative Product Semantic Scale (CPSS), the workers rated each design solution in three distinctive stages – *S*, *3D* and *VR* – on three criteria – *novelty* (freshness or newness), *resolution* (relevance and logic) and *style* (craftsmanship and desirability). Despite some demographic discrepancies, the workers expressed general senses of happiness and calmness, resonating with the competition's requirements. Statistical results of CPSS ratings revealed that while AI excelled in *style* for *3D*, the human designer outperformed in *novelty* for both *S* and *VR*. Collaboration efforts surprisingly finished last. Such findings challenge current assumptions of AI's creative ability in design research and highlight the need to be agile in the age of disruptive technologies. This research also offers guidance for product and interior designers and educators on thoughtfully integrating AI into the design process.

Keywords: Artificial intelligence, Creativity, Human–AI collaboration, Design thinking, Virtual reality, Crowdsourced creativity assessment

Received 07 November 2024
Revised 14 September 2025
Accepted 15 September 2025

Corresponding author
Khanh Hoa T. Vo
hoavo@iu.edu

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use and/or adaptation of the article.

Des. Sci., vol. 11, e37
journals.cambridge.org/dsj
DOI: 10.1017/dsj.2025.10029

1. Introduction

Can a machine be creative? According to Shi *et al.* (2023), this question garners attention of design researchers from multiple disciplines, as shown in 93 publications across different journals between 2007 and 2022. Of these, 34.4% focused on graphic design, 9.7% on product design and 3.2% on interior design. Goldschmidt (2019) even deems artificial intelligence (AI) a pivotal change in design research, especially regarding creativity, with attempts to assess this technology's impact started in Boden's (1991) *The Creative Mind* and continuing to date. But what exactly is AI? We often describe this technology with fear, believing that it will eventually replace humankind. However, AI research focuses on training the machine to adhere to principles of rationality – making optimal decisions based on available information (Russell & Norvig 2003; Jackman *et al.* 2017). Moreover, AI is an umbrella term for various algorithms (e.g., convolutional neural networks

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

and generative adversarial networks) with diverse applications (e.g., text or image generations), each differing in its abilities – all compared to human cognitions (Anantrasirichai & Bull 2022). Whether a machine – or AI – can be creative depends on the ability level that AI can exercise.

Ameen *et al.* (2022) rank AI abilities from *narrow*, to *general* and *super* – with the first two being entirely rational (e.g., learning and reasoning) and the last one being emotionally capable (e.g., empathizing). In other words, narrow AI can perform specific tasks that rely on routine cognitive processes in well-defined contexts. General AI can adapt and solve tasks across different contexts with human-like cognitive flexibility. Super AI can hypothetically solve tasks in any conceivable context beyond human cognitive abilities. Given that creativity remains elusive due to its fluid and ill-defined nature (Ahmed & Fuge 2018), general and super AIs stand the best chance to exercise this ability thanks to their flexibility and superiority in task performance. If that is the case, two scenarios could unfold: AI might compete with human designers, or both could thrive through collaboration.

Despite advancements in AI abilities – from narrow applications like chatbots (Siri and Alexa) to general systems like Generative Pre-trained Transformer (GPT-4) – super AI remains unrealized due to data availability and quality, energy consumption and societal implications like privacy and human-job replacement (Stiefel & Coggan 2023; Lloyd 2024; Raman *et al.* 2025). Since a super AI that rivals humankind has yet to emerge, general AI still adheres to the classic Turing test (1950) – the extent of its abilities depends on whether a human evaluator can discern its response from that of a human. Hence, general AI's creative ability is contingent upon the proximity of its outcomes to an established standard of creativity.

This study provides an empirical analysis of the proximity between solutions generated by general AI, a human designer and a human–AI collaboration within a real-world design competition. The following subsections outline the theoretical framework for defining creativity and current perspectives on AI's creative ability. Findings from this study answer the question of whether a machine exercises creative ability, providing both theoretical and practical insights for product and interior design – two disciplines in which this subject matter remains underexplored (Shi *et al.* 2023). Product and interior design feature "a holistic problem-solving process" that defines functionality, esthetics and experience of a consumer artifact and a spatial entity, respectively (Stöhr, Koldewey & Dumitrescu 2023; Timmermans & Van der Rijst 2023). This study centers on a real-world design competition in which a light fixture – designed to create a comfortable and engaging spatial experience in a pediatric waiting room – serves as a relevant experimental context for both disciplines.

1.1. What is creativity?

There are three movements that have shaped creativity research: personal traits (1950s–1960s), cognitive processes (1970s–1980s), and sociocultural contexts (1980s–1990s) (Sawyer 2012). The first two movements reflected an individualist view of creativity, noting personal qualities and tactics of being creative, as understood and expressed by individuals via tangible outcomes (Merrotsy 2013). An example of creativity as personal traits is an artist who embraces experimenting with unconventional materials, demonstrating openness and willingness to take

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

risks. An instance of creativity as a cognitive process is a designer engaging in the brainstorming process to explore diverse strategies for enhancing a product's usability. Stein (1987) exemplifies the core of these two movements by depicting creativity as everyday activities that are new to the individual, like a child inventing a unique way to organize their toys. In contrast, the third movement adopted a sociocultural view of creativity, defining it as the social act of assessing whether products are novel and valuable by peer groups, as illustrated by Stein's *Big C* (1987). For instance, Picasso's creativity gave rise to the revolutionary movement *Cubism*, transforming the art world and inspiring generations of artists to create under shared esthetics.

Rhodes (1961), Amabile's *componential framework* (1983) and Csikszentmihalyi's *systems model* (1988) unify the above movements, viewing creativity as a multifaceted social phenomenon (Runco & Jaeger 2012). Rhodes (1961) denoted creativity as *product* (socially novel and appropriate), *person* (creative traits), *process* (creative thinking) and *press* (social pressures). Amabile (1983) categorized creativity into domain-relevant skills, creativity-relevant skills and task motivation, with creativity being assessed by peer consensus within a specific domain. Csikszentmihalyi (1988) also described creativity as the intersection of the individual, the field (peers) and the domain (practices), as shown in the domain's acceptance or resistance to novel changes brought forth by the individual to the field. Overall, creativity is an ecosystem in which individuals with specific skills and motivations operate within distinct domains, refining and validating the novelty and appropriateness of their creations through the collective judgment of their peers.

1.2. How to measure creativity?

Creativity measures have evolved through multiple approaches: indirect assessment, global judgment and criterion-based assessment (Horn & Salvendy 2006; O'Quin & Besemer 2011). Indirect assessment involves peer nominations and self-reported achievements, global judgment relies on consensus among judges and criterion-based assessment evaluates specific attributes of creativity. Of these, criterion-based assessment aligns with the comprehensive view of creativity described above – an ecosystem where peers validate each other's creations via specific attributes (Raghunath *et al.* 2023). A measure for this approach is the Creative Product Semantic Scale (CPSS), which evaluates creativity on three criteria: *novelty*, *resolution* and *style* (Besemer 2006). Novelty reflects originality and surprise; resolution covers logic and utility; style addresses craftsmanship and elegance. Interestingly, Norman (2005) and Horn & Salvendy (2009) noted that *resolution* in creative outcomes also manifest via emotional responses, inducing by the coherence between appearance and function. Kayode, Ojo & Sheba (2008) and Ghasemi *et al.* (2024) contextualize *style* as desirability – user-preferred esthetics – and integrity – function-driven esthetics – in product and interior design, respectively.

In addition to CPSS, several legacy creativity measures have been applied to design disciplines, including Guilford's (1967) Alternative Uses Task (AUT), Torrance's (1974) Test of Creative Thinking (TTCT) and Amabile's (1982) Consensual Assessment Technique (CAT). Among these, CAT and CPSS offer the most comprehensive assessments, capturing creativity as a multi-faceted construct that considers both novelty and relevance. In contrast, AUT and TTCT emphasize

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

novelty but fail to account for whether such novelty is meaningful or applicable within a domain-specific context. However, CAT also has a major limitation – the lack of standardized criteria for creativity, which vary across tasks and judges, reducing its generalizability (O'Quin & Besemer 2011). For instance, a study using CAT reported low internal reliability ($\alpha = .18$) due to inconsistent creativity criteria among judges (Lamb, Brown & Clarke 2016), raising concerns about its reliability in empirical research.

CPSS comprises 55 semantic pairs rated on a 7-point scale, across three creativity criteria – novelty (15 pairs), resolution (15 pairs) and style (25 pairs). Multiple studies confirm the high internal reliability ($.69 \leq \alpha \leq .97$) of this measurement (Besemer 1998; Besemer & O'Quin 1999; Horn & Salvendy 2006). White & Smith (2001), Thang *et al.* (2008) and Wei *et al.* (2015) recommended using an abbreviated version of CPSS with fewer semantic pairs, providing sufficient internal reliability ($\alpha > .72$), to reduce potential rater fatigue. However, this study employed the original 55-pair version to provide a thorough and nuanced evaluation of creativity across human, human–AI collaboration and AI-generated solutions. A list of semantic pairs used in this study is available in the Appendix.

1.3. Who should measure creativity? In which context?

In creativity research, the norm is for expert judges to evaluate creations for novelty and appropriateness, ensuring consistent and efficient assessments (Görzen & Kundisch 2019). However, using non-expert judges is increasingly common, yielding reliable results for simple and intermediate creations (Fong *et al.* 2016; Yuan *et al.* 2016; Miceli & Raimondo 2020). For instance, expert judges accurately and efficiently rated the novelty and appropriateness for technical design enhancements, whereas non-expert judges exceled in rating such criteria for consumer design products (Fraser *et al.* 2017; Görzen & Kundisch 2019). Using non-expert judges from crowd-sourced platforms, such as Amazon Mechanical Turk (MTurk) and Prolific, has become a cost-effective and reliable method for assessing design creativity. Many studies have indicated that large-scale crowd-sourced ratings can achieve high reliability and tend to converge with expert evaluations in accuracy (Kudrowitz & Wallace 2013; Ahmed & Fuge 2018; Goucher-Lambert & Cagan 2019). This study uses crowd-sourced evaluations for the creativity of a light fixture, a task of moderate complexity – neither too simple nor too elaborate.

Interestingly, although legacy creativity measures such as AUT, TTCT, CAT and CPSS remain widely accepted, the context of creativity assessment is rapidly evolving. Recent scholarship (Barbot, Kaufman & Myszkowski 2023; Myszkowski 2024) advocates embedding creativity measures within immersive contexts like virtual reality (VR), where evaluators can emotionally and psychologically engage with design solutions as vividly as in real life, enabling more nuanced and accurate assessments of creativity. Barbot *et al.* (2023) further demonstrated that judges with varying expertise levels ($n = 9$) could reliably assess the creativity of visual artworks in VR using a single-item, 7-point rating scale, achieving sufficient internal reliability ($\alpha = .86$). This study also integrates VR into the final evaluation stage, allowing crowd-sourced judges to immerse themselves within the environment as they observe and rate the creativity of various light fixture designs.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

1.4. AI, human and creativity

As product and interior design share a “holistic problem-solving process” (Stöhr *et al.* 2023), the comparison of AI, human and collaboration efforts within these disciplines needs to be situated in an operational framework. This study uses the Double Diamond (DD) – developed by the Design Council (2022) – as the context. The DD is an foundational framework in design research (Flus & Hurst 2021; Martins Pacheco *et al.* 2024) and has been broadly used in AI and creativity research (Bouschery, Blazevic & Piller 2023; Grilli & Pedota 2024). In the book, *Design Creativity*, Jun, Hignett & Clarkson (2024) characterize the DD framework as a two-phase process of *divergent* and *convergent* thinking – exploring possibilities and refining an optimal solution (Guilford 1950; Torrance 1974). The said dual structure makes the DD framework relevant for the comparison of AI, human and collaborative efforts in this study, as it operationalizes the very process of creativity, as well as the disciplines of product and interior design.

Despite its seminal role, the DD framework has limitations. First, it suggests a linear, step-by-step design process, moving from divergent to convergent thinking. In fact, designers often loop back, repeatedly exploring and refining solutions. This toggling between the two diamonds indicates a more fluid and iterative design process, where designers achieve incremental developments by continuously redefining the problem and iterating on solutions (Goel & Pirolli 1992; Jackman *et al.* 2017; Kim & Park 2021). Second, the DD framework operates at a high level of abstraction, failing to accommodate rapid, agile cycles where problem-framing, ideating and solution-testing occur simultaneously rather than sequentially (Zhu & Luo 2022). Nevertheless, the DD framework provides an operational understanding of general AI’s output generation – a process that parallels the first phase of divergent thinking.

According to Boden (1998) and Ameen *et al.* (2022), general AI blends learned solutions into new ones and retrieves domain-specific information to explore possibilities in distantly related areas, following a sequential algorithm to match requests (inputs) with relevant data. Yet AI’s rapid ideation still relies on having humans with expertise to train the models and judge the outputs. As Tørresen (2021) notes, general AI in interior architecture design randomizes limited inputs to generate diverse outputs, thus requiring multiple runs and alternatives. Babakhani (2023) even asserted that the machine lacks the intuitiveness to decide the appropriateness of its outputs. Thus, interior architecture designers play a crucial role in selecting the best-suited design solutions for specific problems (Huang *et al.* 2021; Morrison *et al.* 2023; Abuzuraiq & Pasquier 2024). Runco *et al.* (2024) further found that, on divergent-thinking tasks, humans excelled at “blue sky” ideas, whereas general AI tools bested at pragmatic concepts. Similarly, Wang *et al.* (2024) developed a general AI tool for generative solutions in residential design. Most expert testers ($n = 8$) rated the AI-generated outputs as “Good” to “Very Good” (70% to 90%). However, when interior designers ($n = 6$) collaborated with Wang *et al.*’s (2024) AI tool in residential design projects, they raised concerns about the similarity between generative solutions and a lack of spatial rationality. Results from Runco *et al.* (2024) and Wang *et al.* (2024) both highlight the human edge amidst general AI’s leverage of rapid ideation. Overall, general AI may lack the agility of the human brain, yet the machine manages to emulate the sequential

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

steps of the DD framework, capable in the divergent phase but needs human oversight in the convergent one.

The current design research literature elaborates this observation on AI, humans and creativity. According to Sosa (2019), while divergent thinking increases the quantity of solutions, convergent thinking determines the quality via surprising, synergistic connections among solutions. Expert designers are important, as they intentionally structure early possibilities within domain- and problem-specific constraints, then continually refine relevant options until an optimal one emerges. Thus, AI's rapid generation of numerous solutions does not guarantee creativity. Grilli & Pedota (2024) reinforce this point, noting that AI is bound to its trained knowledge and lacks the flexibility to make connections between distantly related or adjacent domains. Only humans possess the ability to think beyond domain boundaries, connecting less obvious knowledge areas that, though separate, converge within a specific design problem. In other words, while AI can execute algorithms that mimic the divergent thinking phase of the DD framework, it lacks the inherently human capacity for convergent thinking, which limits its ability to achieve true creativity independently.

1.5. Human–AI collaboration and creativity

Another portion of the literature increasingly emphasizes human–AI collaboration – a synergy that pairs human designers' agility and nuanced judgment of creativity with AI's rapid idea generation and pattern recognition. This partnership holds promises for pushing creative boundaries, enabling novel design outcomes beyond what humans alone could achieve. For instance, Bouschery *et al.* (2023) argued that AI expands human capacity for divergent thinking – as illustrated in the DD framework – with transformer-based models like GPT-3 expand access to vast, domain-specific knowledge, a task often time- and labor-intensive in large-scale product design problems. Likewise, Boudier *et al.* (2023) highlight that product designers often experience fixation – bias toward certain solutions due to reliance on familiar knowledge or reasoning. Given proper training, AI can generate diverse possibilities for designers to not only evaluate the novelty and appropriateness of the solutions but also reshape their interpretations of the design problems. Domain-specific expertise then helps refine solutions or draw potential connections between them, enhancing creativity. In other words, AI's rapid generation of potential solutions can counter human designers' fixation on established ruminations, leading to unfamiliar creative explorations.

Rahman, Bayrak & Sha (2024) provide empirical support for human–AI collaboration via a reinforcement learning (RL) model that transfers domain knowledge to predict designers' decisions with up to 78% accuracy, matching the performance of high-achieving designers on specific design problems. This RL model can aid designers in mapping potential decisions for similar problems, making it particularly valuable in cases of design fixation. Likewise, Zhou, Zhang & Yu (2023) found that product design students who used AI-generated possibilities as a springboard for designing luggage produced a higher quantity of solutions ($p = .004$), achieved greater originality, novelty and practicality (all $p < .001$), according to expert ratings. Chen *et al.* (2025) compared creative performance between two groups of novice product designers ($n = 20$) – one using general AI tools (for generative texts and images such as ChatGPT-3.5 and Midjourney) and

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

one without – tasked with designing a baby chair and a set of musical building bricks. Five expert judges rated design solutions of those using general AI tools higher across criteria such as novelty, feasibility, usability and functional diversity. Their findings suggested that human–AI collaboration heightens conceptual explorations and design quality.

Regarding interior design, Gallega & Sumi (2024) developed a general AI tool to support material selection for furniture pieces. Both professionals ($n = 6$) and design students ($n = 5$) collaborated with the tool and reported an average of 72.82 (over 100) for Creativity Support Index, indicating a high perceived effectiveness of general AI in facilitating creative performance. Chandrasekera, Hosseini & Perera (2024) compared two groups ($N = 20$) of interior design students, with and without general AI tool (Midjourney), to complete the same chair design brief. Using CPSS, two expert judges evaluated all outcomes and rated the group with AI significantly higher in creativity ($p < .05$). In contrast to Chen *et al.* (2025), Chandrasekera *et al.* (2024) reported a single, averaged CPSS score rather than analyzing each dimension separately. While this approach offers a simplified overview, it differs from the original CPSS protocol established by Besemer & O'Quin's (1999) and later expanded by Besemer (2006) and O'Quin & Besemer (2011), who recommend analyzing each factor independently to reduce the risk of Type II errors – that is, overlooking meaningful differences that may exist within individual creativity dimensions.

In general, the above scholarly discourses and empirical accounts regard collaboration between AI and human designers as the highest levels of creativity. Nevertheless, the key caveat of relying on AI-generated possibilities is decreased originality; a higher quantity often leads to redundant, incremental concepts rather than truly novel ideas (Sarica & Luo 2024; Wang *et al.* 2024). Furthermore, repeated iterations of familiar solutions can reinforce design fixation, particularly harmful for novice designers (Boudier *et al.* 2023; Rahman *et al.* 2024). Yet this caution underscores the importance of human designers in human–AI collaboration, where their abilities to think beyond domain boundaries and make novel connections counteracts the diminishing return of AI.

2. Aims

The author pursued two primary aims to answer whether a machine can be creative as measured against the proximity of its outcomes to an established standard of creativity. First, the author empirically investigated whether significant differences exist in creativity among design solutions generated by AI alone, humans alone and human–AI collaboration. Second, the author contributed insights into ongoing debates regarding whether AI will reach human-level creativity or if collaboration between humans and AI can unlock unprecedented creative possibilities. To achieve these aims, the author tested the following hypotheses:

H1: AI is less creative than a human designer. The author hypothesized that in a real context with specific functional, esthetic and emotional requirements of the Robert Bruce Thompson Lighting Design Competition, AI-generated design solutions will exhibit lower creativity than those of a human designer.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

H2: Human–AI Collaboration is more creative than AI alone. The author hypothesized that in the real context given above, collaboration between humans and AI leads to greater creativity than the sole performance of AI.

H3: Human–AI Collaboration is more creative than a human designer alone. The author hypothesized that in the real context given above, collaboration between humans and AI leads to greater creativity than the sole performance of a human.

The author expected to find empirical evidence of creativity in AI-generated design, supposedly at a lesser degree than solutions produced by a human designer or through human–AI collaboration. Specifically, the author anticipated that human–AI collaboration would excel in the functional, esthetic and emotional requirements of the Robert Bruce Thompson Lighting Design Competition.

3. Significance

This study advances AI research in design via a rigorous creativity assessment across AI-generated, human–AI collaborated and human-designed solutions within the practical setting of a design competition. Distinct from previous studies that relied on subjective ratings, this study employs the established measure of CPSS (Besemer 2006) to provide structured evaluation criteria, thereby reducing uncertainties of subjective criteria, as shown in Amabile's (1982) CAT. Notably, raters often judged work labeled as human-made to be more creative than AI-generated, yet they could not distinguish the difference without labels (Horton Jr, White & Iyengar 2023; Magni, Park & Chao 2024). To avoid this bias, all designs in this study were presented unlabeled, ensuring that crowd-sourced judges evaluated them under fully blind conditions, unaware of whether a design was human-made, AI-generated or human–AI collaborated.

In this study, the CPSS criteria were defined as follows, bridging the Robert Bruce Thompson Lighting Design Competition brief and the creativity literature discussed above: *Novelty* signifies the freshness or newness of light fixture concepts. *Style* means the appearance or craftmanship of light fixtures as perceived via design visualizations. "Unattractive—attractive" semantic pair under the *style* criterion also captures subjective desirability (Kayode *et al.* 2008; Ghasemi *et al.* 2024) of light fixtures. *Resolution* indicates the relevance and logic of light fixtures being mounted devices (to the ceiling or wall). Bradley & Lang's (1994) Self-Assessment Manikin (SAM) offers additional measures for emotional responses to light fixtures, an extended understanding of *resolution* (Norman 2005; Horn & Salvendy 2009), assessing whether their functions and appearances coherently elicit the senses of "comfort" and "engaging" for a pediatric waiting room.

Moreover, this study offers two significant contributions. First, it offers empirical guidance to product and interior designers and educators on strategically integrating AI into the design process, particularly when design fixation limits innovation, by using AI to overcome habitual thinking patterns of fixation and stimulate novel insights. Second, it identifies key points in the design process where AI involvement is most beneficial, promoting effective AI-driven explorations while preserving space for human judgment and agility.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

4. Method

To assess whether creativity differs among AI, Collaboration and Human conditions, the author drew on the Robert Bruce Thompson Lighting Design Competition and compared crowd-sourced ratings of the designs through the ideation, design development and final solution stages. Traditional creativity assessments focus on isolated final products, whereas a more comprehensive approach considers how different social groups perceive creativity across design stages (Sosa & Gero 2005; Cross 2006; Surma-Aho, Björklund & Hölttä-Otto 2022; Raghunath *et al.* 2023). Crowd-sourced evaluations, when sufficiently scaled, align closely with expert ratings, making them an effective tool for assessing design creativity (Foong, Gergle & Gerber 2017; Goucher-Lambert & Cagan 2019). Additionally, evaluating the creativity of a light fixture is well suited for crowd-sourcing, as its moderate complexity allows for manageable, multistage assessments without being overly simplistic or excessively intricate (Galati 2015).

Lastly, while the author used two-dimensional (2D) images for assessments during the ideation and development stages, the final solution stage included virtual waiting rooms featuring three-dimensional (3D) models of the light fixtures. The author selected this approach to align with recent literature in creativity research, where embedding creativity measures within VR's real-time, interactive and immersive environment has been increasingly advocated (Barbot *et al.* 2023; Myszkowski 2024) due to enhanced spatial perception, emotional engagement and accurate assessment of design quality. The author conducted this study under the Institutional Review Board approval #22214 from their institution.

4.1. The design competition

The author chose the established Robert Bruce Thompson Lighting Design Competition as the context for evaluating creativity among AI-generated (AI), human–AI collaboration (Collaboration) and human-designed (Human) solutions. This decision was supported by two key considerations from creativity research literature. First, this annual competition, named after Bruce Thompson, a lighting design expert with over 25 years of experience, provides a distinctive yearly theme and well-defined constraints, creating an ideal ecosystem to assess creativity within clearly articulated, expert-driven parameters. Second, the competition's scope is carefully balanced, centering on designing a single light fixture that is neither overly simplistic nor excessively complex, and which demonstrates novelty and appropriateness within an annually specified context. Original submission rule specifies a four-slide presentation (11″ × 17″) with a concise, 250-word conceptual description; beyond this, interpretation and creativity are left open-ended. There are no requirements regarding the number of iterations or process documentation; instead, the competition emphasizes showcasing the final design outcomes. Previous winning entries usually feature conceptual sketches, explanatory diagrams and visualizations showcasing the fixture in use.

The 2024 competition brief – designing a ceiling- or wall-mounted light fixture for a pediatric waiting room with a 10′ ceiling height that is comfortable and engaging for children – was deliberately chosen by an expert panel attuned to emerging trends and authentic needs within their discipline. This specific context provided a practical, rather than artificial, scenario ideal for assessing creativity, as

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

the designs had to balance functional, esthetic and emotional considerations. The competition thus aligns closely with established frameworks of creativity assessment (Amabile 1982; Besemer 2006), with *novelty* being the conceptual originality, *style* referring to structural refinement as seen via visual representations and *resolution* demonstrating functional execution. Furthermore, *resolution* can be evaluated as emotional responses (Norman 2005; Horn & Salvendy 2009) – comfortable and engaging as described in the brief. Thus, the Robert Bruce Thompson Lighting Design Competition was an ideal context for the nuanced analysis of design quality required in this study. The following sections review relevant literature on creativity involving AI, human designers and human–AI collaboration, providing the foundation for the research question and hypotheses in this study.

4.2. Participants

To obtain consistent evaluations of design creativity, the author conducted two separate experiments on different crowd-sourced platforms. Participants serving as crowd-sourced judges did not require design expertise, reflecting common practice in creativity research where non-expert judges provide a realistic measure of design impact and appeal to a broader audience (Ahmed & Fuge 2018). Experiment I included 120 MTurk workers, whereas Experiment II involved 126 Prolific workers. Participants in both experiments evaluated the creativity of light fixture designs across AI, Collaboration and Human conditions. For Experiment I, the inclusion criteria for MTurk workers were individuals based in the United States (US), aged 18 or older, with Master's qualifications, indicating a record of high-quality performance on the platform. The exclusion criteria included multiple attempts and prior participation in other experimental conditions. In other words, each participant could complete the task only once and under one condition. In Experiment II, the author also recruited Prolific workers based on being in the US and at least 18 years old. As Prolific workers are generally considered to provide higher-quality data (Peer *et al.* 2017), the author required no additional qualifications. However, the same exclusion criteria applied: participants who made multiple attempts or had already taken part in any condition were not eligible. The uses of selection criteria and established platforms (Palan & Schitter 2018) enhanced reliability in data collection, thereby supported the validity of creativity assessments.

The MTurk sample ($n = 120$) was predominantly similar in age, with most participants reporting management positions. Notably, there was no representation from design backgrounds. The MTurk workers in the AI condition were 80% male and 20% female, primarily from 25 to 44 years old (82.5%). Occupation classifications were 35% in management positions, 17.5% each for sales and technical positions, 15% freelance and 12.5% in other roles. Those in the Collaboration condition were 50% male and 45% female, with another 5% prefer not to disclose, also between 25 and 44 years old (67.5%). Occupation classifications were 42.5% in sales, 32.5% in management positions, 12% freelance and 13% in other roles. The rest in the Human condition were 55% male and 45% female, between 25 and 44 years old (75%). Occupation classifications were 25% in management positions, 17.5% in sales, 25% freelance, 10% manual labor jobs and 22.5% in other roles.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

The Prolific sample ($n = 126$) was younger and more gender-diverse than MTurk workers. Most participants reported working in education, management and technical roles, with a small percentage having design backgrounds. The Prolific workers in the AI condition were 69% male, 28.6% female and 2.4% trans female, between 18 and 44 years old (95.24%). Occupation classifications were 21.5% working in education, 21.4% with management positions, 11.9% with technical positions, 9.5% in art and design positions, 7.1% freelance and 28.6% in other roles. Those in the Collaboration condition were 40% male, 55% female and 5% non-binary. Occupation classifications were 21.4% working in education, 14.3% with management positions, 11.9% each for sales and technical positions, 7.2% in art and design positions, 7.1% freelance and 16.7% in other roles. Those in the Human condition were 38% male, 57% female and 5% non-binary. Occupation classifications were 22.5% with management positions, 17.5% working in education, 15% freelance, 7.5% either in sales or technical positions and 30% in other roles.

The author collected data from two different research platforms to ensure data quality. While using MTurk, even among workers with Master's qualifications, several issues emerged, including failed attention checks and bot-generated responses. A notable example occurred when participants were asked to identify the color of a furniture piece in the virtual waiting room where the light fixture models were displayed. Instead of providing a direct response, multiple participants with different IDs submitted identical, irrelevant and AI-generated text, such as:

> An ottoman is a low, upholstered stool, usually with legs. This furniture can be used as a footrest, coffee table, or seat. Many ottomans can also double as hidden storage, with a hollow center that can organize blankets, games.

Due to these inconsistencies, the author removed 172 out of 292 responses (59%) from MTurk workers for poor data quality. To improve reliability, the experiment was repeated with Prolific workers, a platform known for higher data integrity. The results were significantly better: out of 196 responses, only 70 (36%) were removed due to incomplete surveys (e.g., skipped questions), and no bot-generated answers were detected in the attention check.

4.3. Materials

In each condition, the ideation stage included a 2D sketch (lines and forms only, without colors); the design development stage contained a 3D rendering (full colors, materials and lighting); the final solution stage showcased a web-based VR waiting room with a 3D model that is observable from multiple directions (see Figure 1). The author obtained materials for the three conditions as follows:

In the AI condition, the author systematically employed ChatGPT-4, Midjourney, and Kaedim AI – selected specifically for their advanced generative capabilities in text, 2D images and 3D modeling (see Figure 2). The choice of these tools aligns with current literature highlighting their state-of-the-art performance in generating possible solutions for design tasks (Liang, Shan & Chung 2023; Meron & Tekmen Araci 2023; Ghasemi *et al.* 2024). Specifically, at the time this study was conducted in 2024, ChatGPT-4 represented the most advanced model for text generation from OpenAI, significantly surpassing the prior version,

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

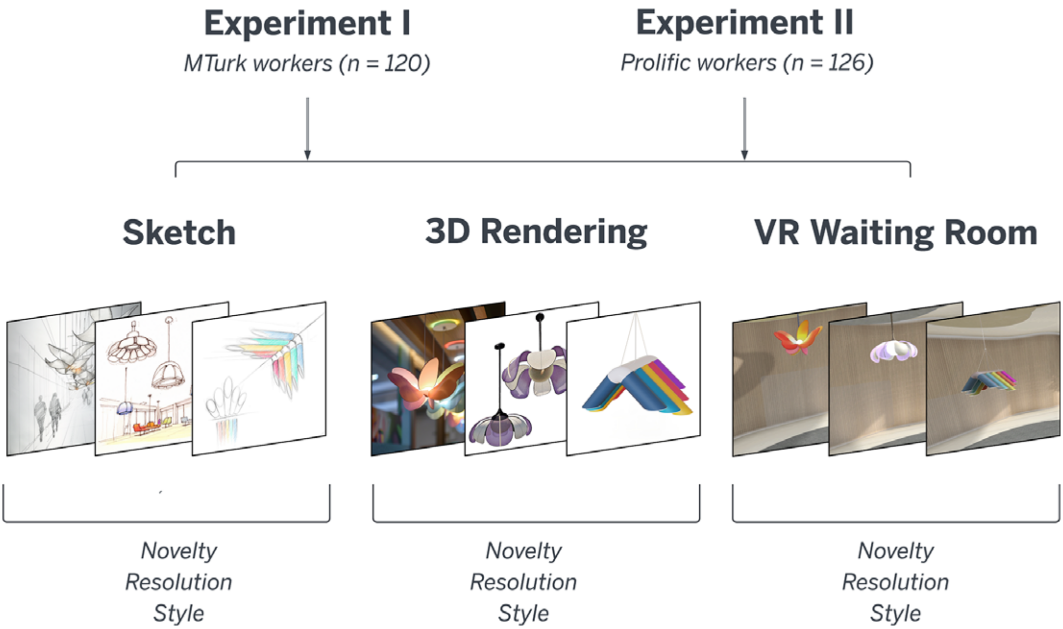

Figure 1. The research design involving two experiments with three conditions and three stages of the design process, measured using CPSS criteria for creativity. Graphical representations were created by the author and Caroline Alleva, using Lucidchart (educational license).

ChatGPT-3.5, in reasoning, coherence and contextual adaptability. To initiate the ideation phase, the author inserted the 2024 Robert Bruce Thompson Lighting Design Competition brief into ChatGPT-4 with the following instruction: “You are now a prompt engineer for generative AI like Midjourney. You have been given the following design problem. Please provide 5 different prompts to respond to the design problem.” The author ran the instructions multiple times, eliminated redundant outputs and ended up with 73 unique textual prompts. Using Midjourney, the author generated four visuals per prompt, resulting in 292 initial concept sketches.

Based on the competition’s verbatim brief – prioritizing novelty and appropriateness for a light fixture in a pediatric waiting room – the author filtered these concept sketches into 19 highly relevant options, ruling out those with extraneous or irrelevant features. The screening process resulted in one sketch with the highest clarity and feasibility for 3D modeling. The author input the sketch back into Midjourney to generate multiple 3D renderings, identifying one that closely matched the original concept among those with random, unexpected details. Finally, the author used Kaedim AI to convert this 3D rendering into a 3D model in 5 hours of computational processing. Throughout this process, the author applied clear, predefined selection criteria to minimize subjective influence on the material selection for the AI condition. Aligning with current literature on AI research that general AI still relies on human judgment to evaluate the appropriateness of its outputs (Babakhani 2023; Abuzuraiq & Pasquier 2024) – and with super AI nowhere yet in sight (Stiefel & Coggan 2023) – this method remains the most practical and feasible approach to addressing the fundamental question: Can a machine be creative?

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

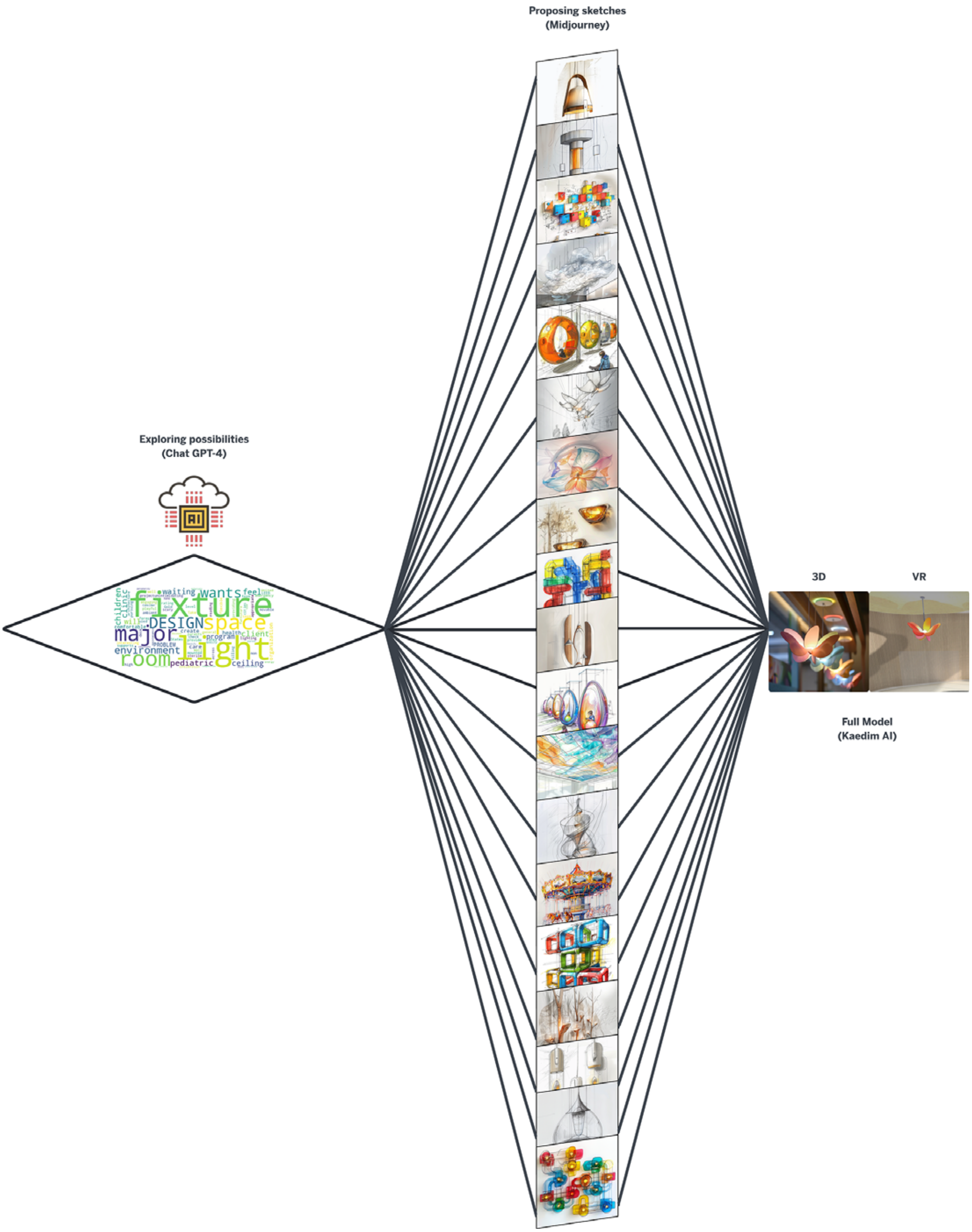

Figure 2. AI condition mapped into the DD framework. Solutions were generated by ChatGPT-4 and Midjourney, and graphical representations were created by the author and Caroline Alleva, using Lucidchart (educational license).

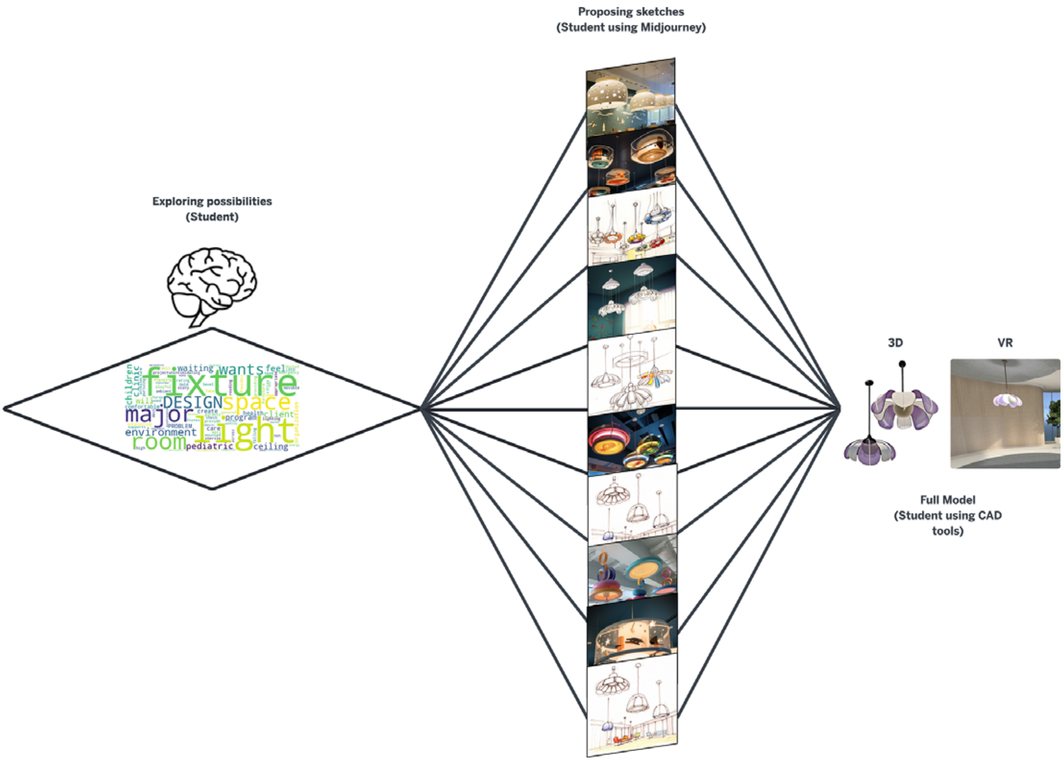

Figure 3. Collaboration condition mapped into the DD framework. Solutions were designed by Hana Pham, and graphical representations were created by the author and Caroline Alleva, using Lucidchart (educational license).

In the Collaboration condition, the author recruited an interior design student who participated in the 2024 Robert Bruce Thompson Lighting Design Competition as part of a class project in fall 2023 (see Figure 3). This student, a junior proficient in design heuristics and computer-aided design (CAD) tools, first generated initial design concepts using Midjourney. The student then refined these AI-generated outputs manually with additional hand sketches to introduce shape variations, details and color enhancements. The student reintroduced these enhanced sketches into Midjourney with tailored prompts, resulting in a collection of 10 sketches specifically addressing the competition theme. The student then selected 3 of the 10 sketches for further exploration and modeling in various CAD tools, ultimately refining one final design into a detailed digital model with materials and finishes. The process took 5 weeks, during which the student independently managed all aspects of AI prompting, selection criteria and manual adjustments of the final light fixture. During instructional hours, the author monitored student work to ensure that it met the institution's interior design program standards and aligned with the competition brief.

Participation in the annual Robert Bruce Thompson Lighting Design Competition is central to the lighting design studio in the interior design program, where this study was conducted. Although the competition itself does not mandate process documentation or specify a required number of concept sketches, the instructor (also the author) required all students, including the

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

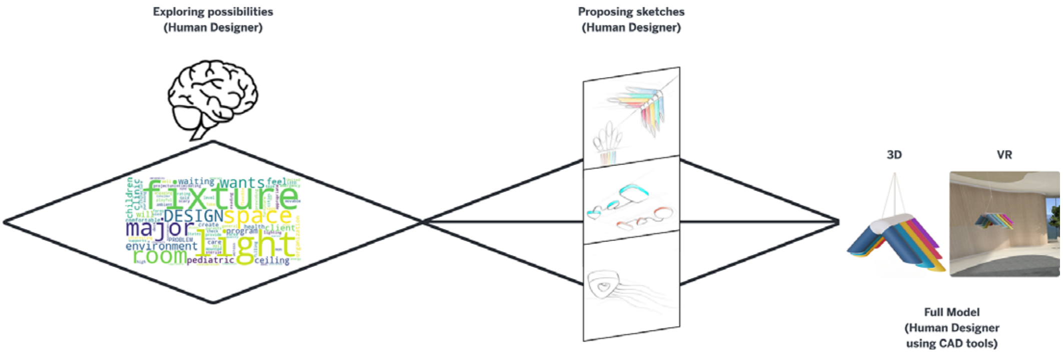

Figure 4. Human condition mapped into the DD framework. Solutions were designed by Slim Z., and graphical representations were created by the author and Caroline Alleva, using Lucidchart (educational license).

volunteer participating in this study, to produce a minimum of 10 initial concept sketches before finalizing their design. This research-informed decision aligns with Kudrowitz & Wallace's (2013) findings that generating at least 10 conceptual ideas significantly correlates with greater creativity in the final design outcome.

In the Human condition, the author hired a freelance designer from Upwork in 4 weeks with $1,200 compensation, selected for their proficiency with design heuristics and CAD tools, but without high-level expertise in designing light fixtures. This approach controlled potential confounding variables related to expertise between the student in Collaboration and the criteria-based selection in AI conditions (see Figure 4). The freelance designer was tasked with responding to the 2024 Robert Bruce Thompson Lighting Design Competition by producing three deliverables, following three stages of creativity assessment: a 2D concept sketch, a 3D rendering and a fully detailed CAD model with materials and finishes. The freelance designer independently created three conceptual sketches, beyond the hiring requirements. The designer retained full creative autonomy in selecting and developing the final design. The author gave no additional instructions besides the initial requirements, allowing the designer complete creative freedom. Throughout the process, the author continuously monitored the quality of the deliverables to ensure that they matched the designer's Upwork portfolio, thereby preserving the study's integrity and experimental rigor.

4.4. Procedure

In Experiments I (Amazon MTurk) and II (Prolific), participants completed an online Qualtrics questionnaire after providing informed consent. Those who agreed to participate continued to the questionnaire, whereas those who declined received a thank-you note, and the survey automatically ended. The questionnaire consisted of four demographic questions (e.g., age, occupation and education level), two commitment prompts encouraging participants to answer the questions thoughtfully, an attention check to ensure that participants visited the virtual waiting rooms and nine items from the CPSS scales measuring novelty, resolution and style (three items each) for the 2D sketch, 3D rendering and VR model,

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

respectively. While reviewing the 2D sketch and 3D rendering within the questionnaire, participants accessed the virtual waiting room via a link that opened in a new window. After that, they returned to the still-open questionnaire to continue. Tables 1 and 2 present the CPSS rating items along with sample paired binaries for each criterion. Moreover, two items from the SAM measured calmness versus excitement and unhappiness versus happiness (Bradley & Lang 1994). These additional self-reported measures complemented the *resolution* criterion of CPSS, as Norman (2005) and Horn & Salvendy (2009) noted that emotional responses to a design solution are also indicators of its creativity. Specifically, the Robert Bruce Thompson Lighting Design Competition requires that the light fixture for the pediatric waiting room be comfortable and engaging. Therefore, the SAM measures helped detect these intangible requirements in the design solutions across conditions.

4.5. Analysis

The author provided descriptive statistics, including means (*M*), standard deviations (*SD*) and percentage (%), to summarize demographic information and self-reported emotions from the two experiments. For hypothesis testing, the author then conducted a series of one-way ANOVA across conditions (AI, Collaboration and Human) for each CPSS criterion (novelty, resolution and style). In Experiments I and II, each condition included 40 and 42 participants, respectively. To assess ANOVA assumptions, the author used Shapiro–Wilk tests for normality and Levene tests for homogeneity of variance. When assumptions were violated, the author conducted Kruskal–Wallis tests as a nonparametric alternative to one-way ANOVA (Yazici & Yolacan 2007; Lock *et al.* 2013). The author conducted statistical analyses in RStudio Desktop (RStudioTeam 2023), confirming significant results at a 95% confidence level.

4.6. Limitations

A primary limitation of this research is its exclusive reliance on crowd-sourced CPSS ratings without direct comparison to expert evaluations. Although the author recruited two sizable crowd-sourced samples (264 non-expert judges) and conducted non-parametric statistical analyses to mitigate demographic biases, this observed divergence suggests that having expert and non-expert evaluations might offer a more comprehensive understanding of creativity differences among AI-generated, human–AI collaborated and human-created solutions.

A secondary limitation of this study is the unavoidable manual involvement by the author in the AI condition – such as inputting the competition brief into ChatGPT-4, transferring generated prompts into Midjourney, screening irrelevant designs and selecting final concepts for 3D modeling in Kaedim AI. The author mitigated potential bias by strictly adhering to objective, predefined selection criteria drawn directly from the competition brief, specifically prioritizing designs depicting clear, wall- or ceiling-mounted fixtures free from irrelevant elements or distracting backgrounds. Nonetheless, fully removing human intervention remains impossible, as current general AI systems lack intuitive judgment and continue to rely on human expertise for evaluating their outputs (Tørresen 2021; Ameen *et al.* 2022; Babakhani 2023; Abuzuraiq & Pasquier 2024). Even the comprehensive studies of Horton Jr *et al.* (2023) and Magni *et al.* (2024) – which involved 2,965 participants

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

(6 experiments) and 2,039 participants (3 experiments), respectively, evaluated the creativity of human versus AI-generated artworks – researchers still manually created AI materials. Runco *et al.* (2024) also monitored the AI-generated outputs for a divergent thinking test, removing inappropriate verbiage to ensure the effectiveness and relevance of the research materials. Until super AI, capable of autonomous decision-making beyond human abilities (Ameen *et al.* 2022), becomes a reality, some degree of human intervention in AI-driven research remains a necessary methodological trade-off.

5. Results

5.1. Creativity ratings and self-reported emotions

5.1.1. Experiment I

For creativity ratings, MTurk workers evaluated 55 semantic pairs for three criteria – novelty, resolution and style – using a 7-point Likert scale. Across three design stages – Sketch, 3D and VR – each worker rated a total of 165 pairs. To mitigate potential fatigue effects, the author implemented attention checks and excluded raters who failed to meet these criteria. Additionally, strong internal reliability across rating pairs ($.84 \le \alpha \le .95$) suggests that cognitive load did not bias the results. CPSS mean scores ranging, as shown in Table 1 ($4.3 \le M \le 5.9$), indicate that judges perceived all design conditions as moderately to highly creative. AI-generated designs received generally the highest scores, particularly in the Sketch and 3D stages. The standard deviations were within one standard deviation, indicating consistent creativity ratings across participants.

Despite demographic discrepancies, all three conditions exhibited similar statistical trends in self-reported emotions. MTurk workers in each condition indicated their emotions on a 9-point Likert scale, representing the spectrums between calmness (1) versus excitement (9) and unhappiness (1) versus happiness (9) – with 5 being the mid-point of this scale. In the AI condition, participants were generally happy as shown via a high mean score ($M = 6.90$), despite slight variations with a moderate standard deviation ($SD = 1.77$). However, their emotions were neutral at a mean score leaning toward the mid-point scale ($M = 5.35$) and quite differed between calmness and excitement with a larger standard deviation ($SD = 2.35$). In the Collaboration condition, participants exhibited a similar sense of happiness via a high mean score ($M = 6.83$) and fewer variations with a lower standard deviation ($SD = 1.52$). While participants indicated a calmer feeling through a mean score lower than the mid-point ($M = 4.63$), their differences showed in a similarly large standard deviation ($SD = 2.36$). In the Human condition, participants reported a similar sense of happiness with few variations via a mean score higher than the mid-point ($M = 6.83$) and moderate standard deviation ($SD = 1.57$). They were also calmer but with differences among individuals, as shown in a mean score lower than the mid-point ($M = 4.78$) and a large standard deviation ($SD = 2.38$). Overall, participants reported a sense of happiness and neutrality, with some differences in excitement levels.

5.1.2. Experiment II

For creativity ratings, Prolific workers also completed 55 semantic pairs over three criteria for each design stage, rating a total of 165 pairs. Like Experiment I, the

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Table 1. Means and standard deviations for novelty, resolution and style across stages and conditions in Experiment I. Created by the author

7-point Likert criterion	Stage	Condition	Mean (M)	Standard deviation (SD)
Novelty (e.g., overused–fresh, predictable–novel and average–revolutionary)	Sketch	AI	5.7	0.7
		Collaboration	4.3	1.6
		Human	5.3	0.7
	3D	AI	5.5	1.0
		Collaboration	5.3	1.0
		Human	5.4	0.8
	VR	AI	5.4	0.9
		Collaboration	5.3	0.8
		Human	5.5	0.8
Resolution (e.g., illogical–logical, useless–useful and ineffective–effective)	Sketch	AI	5.5	0.8
		Collaboration	5.2	1.2
		Human	5.1	1.1
	3D	AI	5.9	0.6
		Collaboration	5.7	0.9
		Human	5.4	1.0
	VR	AI	5.6	1.0
		Collaboration	5.6	0.9
		Human	5.6	1.0
Style (e.g., coarse–elegant, busy–refined and boring–interesting)	Sketch	AI	5.6	0.8
		Collaboration	5.2	1.1
		Human	5.2	1.0
	3D	AI	5.9	0.6
		Collaboration	5.6	0.9
		Human	5.4	0.9
	VR	AI	5.6	0.9
		Collaboration	5.6	0.9
		Human	5.5	1.0

author applied attention checks to mitigate potential fatigue effects by excluding raters who failed to meet these criteria. Again, strong internal reliability across rating pairs ($.87 \leq \alpha \leq .93$) indicates that cognitive load did not bias the results. As shown in Table 2, on a 7-point Likert scale, CPSS ratings ($4.3 \leq M \leq 5.2$) varied from moderate to high across AI, Collaboration and Human conditions. Notably, Collaboration received slightly higher scores for resolution in the Sketch stage and for novelty and style in the 3D stage. Overall, standard deviations were within or slightly above one deviation, reinforcing the consistency in CPSS ratings among participants.

Despite demographic discrepancies, statistical trends across the sample consistently showed positive emotions. Prolific workers in each condition used the same 9-point Likert scale as MTurk workers to report their emotions. In the AI condition, participants reported overall happiness with a mean above mid-point

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Table 2. Means and standard deviations for novelty, resolution and style across stages and conditions in Experiment II. Created by the author

CPPSS criterion	Stage	Condition	Mean (M)	Standard deviation (SD)
Novelty (e.g., overused–fresh, predictable–novel and average–revolutionary)	Sketch	AI	4.9	1.0
		Collaboration	4.3	1.0
		Human	4.9	0.9
	3D	AI	4.8	1.4
		Collaboration	5.3	0.8
		Human	5.0	1.0
	VR	AI	4.4	1.3
		Collaboration	4.9	1.1
		Human	5.1	1.1
Resolution (e.g., illogical–logical, useless–useful and ineffective–effective)	Sketch	AI	4.9	1.1
		Collaboration	5.2	1.0
		Human	4.9	1.0
	3D	AI	5.4	1.2
		Collaboration	5.4	1.0
		Human	5.1	1.1
	VR	AI	5.1	1.2
		Collaboration	5.4	1.0
		Human	5.6	1.0
Style (e.g., coarse–elegant, busy–refined and boring–interesting)	Sketch	AI	5.1	1.0
		Collaboration	5.0	0.8
		Human	5.0	0.9
	3D	AI	5.4	0.9
		Collaboration	5.5	0.8
		Human	5.0	1.1
	VR	AI	5.2	1.0
		Collaboration	5.4	0.8
		Human	5.4	1.0

($M = 6.83$), but the moderate standard deviation ($SD = 1.74$) indicated variations between individuals. They reported being calm with a low mean score ($M = 3.95$), yet there were differences among individuals – as shown in a moderate standard deviation ($SD = 2.00$). In the Collaboration condition, participants displayed a similar trend with a high mean ($M = 6.36$) and a considerable standard deviation ($SD = 1.95$). They shared the feeling of calmness with a slightly lower mean score ($M = 3.88$) and a comparable standard deviation ($SD = 2.03$). In the Human condition, participants were the happiest as shown in a very high mean ($M = 7.10$) and a smaller standard deviation ($SD = 1.65$). They were more excited with the highest mean score ($M = 5.12$) and a similar standard deviation ($SD = 2.20$). In short, participants were happy and calm, with those in the Human condition showing slight excitement, noting individual variations across the sample.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

5.2. H1: AI is less creative than a human designer

The following statistical conventions denote CPSS mean scores (M) under different criteria – novelty (N), resolution (R) and style (S) – in the hypotheses. In the AI condition, the mean scores went as AI_N_M, AI_R_M and AI_S_M. In the Collaboration condition, those scores named Co_N_M, Co_R_M and Co_S_M. In the Human condition, the scores were Hu_N_M, Hu_R_M and Hu_S_M. The author conducted either one-way ANOVA or Kruskal–Wallis tests – depending on whether statistical assumptions were met – to examine the null and alternative hypotheses below at a 95% confidence level ($p < .05$).

$H0_1$: There is no significant difference between AI and a human designer in terms of Novelty, Resolution and Style.

$$AI\_N\_M = \mathrm{Hu\_N\_M},$$

$$AI\_R\_M = \mathrm{Hu\_R\_M},$$

$$AI\_S\_M = \mathrm{Hu\_S\_M}.$$

$H1_1$: AI performs worse than a human designer in terms of Novelty, Resolution and Style.

$$AI\_N\_M < \mathrm{Hu\_N\_M},$$

$$AI\_R\_M < \mathrm{Hu\_R\_M},$$

$$AI\_S\_M < \mathrm{Hu\_S\_M}.$$

Statistical results from Experiment I – MTurk sample ($n = 120, N = 40$) – supported $H0_1$ and rejected $H1_1$. The author conducted Shapiro–Wilk and Levene tests, for normality and homogeneity, on CPSS ratings in all three evaluation stages: 2D sketch (S), 3D rendering ($3D$) and VR waiting rooms (VR). Although most CPSS ratings met the assumption of normal distribution ($p > .05$), those values violated homogeneity of variances, especially in the S and $3D$ stages ($p < .001$). Thus, the author used Kruskal–Wallis tests instead of one-way ANOVA for the MTurk sample.

Statistical results from Experiment II – Prolific sample ($n = 126, N = 42$) – partially supported $H1_1$. Shapiro–Wilk tests showed that most CPSS ratings in the S and $3D$ stages had normal distributions, yet the values in the VR stage mainly violated this assumption. Likewise, Levene tests confirmed homogeneity in the S and $3D$ stages while suggesting violations in the VR stage. To maintain statistical consistency, the author also used Kruskal–Wallis tests instead of one-way ANOVA for the Prolific sample.

5.2.1. *Design stage: sketch*

In the S stage of Experiment I, a Kruskal–Wallis test for *novelty* indicated a significant difference ($X^2 = 18.21, \mathrm{df} = 2$, $p < .001$) in CPSS ratings between the

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

conditions. Nevertheless, a follow-up Mann–Whitney U test with Bonferroni adjustment ($p = .06$) showed no statistical evidence for discrepancies between *AI_N_M* and Hu_N_M. For *resolution* and *style*, Kruskal–Wallis tests indicated no significant differences across conditions ($X^2 = 2.10$, $p = .35$ and $X^2 = 4.24$, $p = .12$, respectively).

In the S stage of Experiment II, a Kruskal–Wallis test for *novelty* indicated a significant difference ($X^2 = 9.37, \mathrm{df} = 2$, $p = .009$) in CPSS ratings between the conditions. However, a follow-up Mann–Whitney U test with Bonferroni adjustment ($p = 1.00$) also yielded no evidence for discrepancies between *AI_N_M* and Hu_N_M. For *resolution* and *style*, Kruskal–Wallis tests were insignificant across conditions ($X^2 = 2.89$, $p = .24$ and $X^2 = 1.26$, $p = .53$, respectively).

5.2.2. Design stage: 3D

In the 3D stage of Experiment I, only the Kruskal–Wallis test for *style* was significant ($X^2 = 6.87, \mathrm{df} = 2$, $p = .03$), whereas those for *novelty* and *resolution* provided no evidence for discrepancies between conditions ($X^2 = 2.54$, $p > .1$ and $X^2 = 3.60$, $p > .10$, respectively). Another follow-up Mann–Whitney U test with Bonferroni adjustment ($p = .02$) revealed that AI scored significantly higher in *style* than a human designer with a medium effect size ($d = .60$), as shown in Figure 5.

Again, in the 3D stage of Experiment II, only the Kruskal–Wallis test for *style* was significant ($X^2 = 6.62, \mathrm{df} = 2$, $p = .04$), and those for *novelty* and *resolution* showed no statistical significance ($X^2 = 2.71$, $p > .1$ and $X^2 = 2.94$, $p > .10$, respectively). However, a follow-up Mann–Whitney U test with Bonferroni adjustment ($p = .08$) revealed no differences between the performance of AI and the human designer.

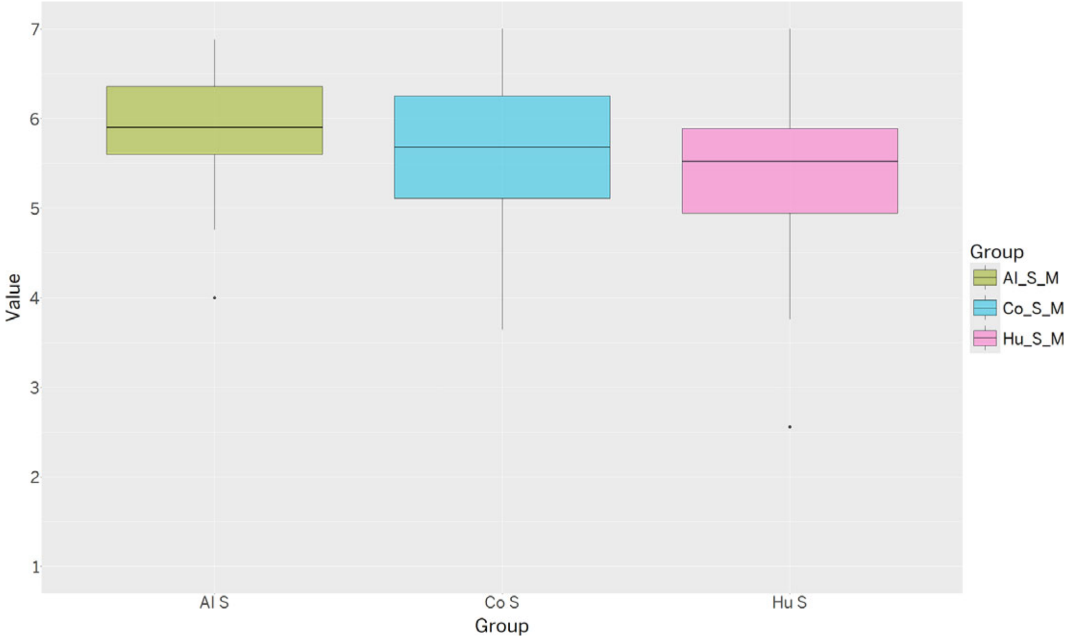

Figure 5. Mean differences in style across three conditions in Experiment I during 3D stage. Data visualizations were created by the author in RStudio.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

5.2.3. Design stage: VR

In the *VR* stage of Experiment I, all Kruskal–Wallis tests were insignificant ($p > .10$). These results supported $H0_1$, showing that AI performed comparably to the human designer and even excelled in *style* during *3D* stage (see Figure 5).

In the *VR* stage of Experiment II, a Kruskal–Wallis test for *novelty* turned out to be significant ($X^2 = 6.59, \mathrm{df} = 2,\ p = .04$) and the follow-up Mann–Whitney *U* test with Bonferroni adjustment ($p = .04$) confirmed that the human designer outperformed AI with a medium effect size ($d = .62$). This key result gave partial evidence for the superiority of the human designer, as indicated in $H1_1$, at least in *novelty* during *VR* stage (see Figure 6). Hence, despite statistical similarities with Experiment I, Experiment II led to an opposite conclusion.

5.3. H2: Human–AI Collaboration is more creative than AI alone

$H0_2$: There is no significant difference between Collaboration efforts and AI in terms of Novelty, Resolution and Style.

$$Co\_N\_M = \mathrm{AI\_N\_M},$$

$$Co\_R\_M = \mathrm{AI\_R\_M},$$

$$Co\_S\_M = \mathrm{AI\_S\_M}.$$

$H1_2$: Collaboration efforts were better than AI in terms of Novelty, Resolution and Style.

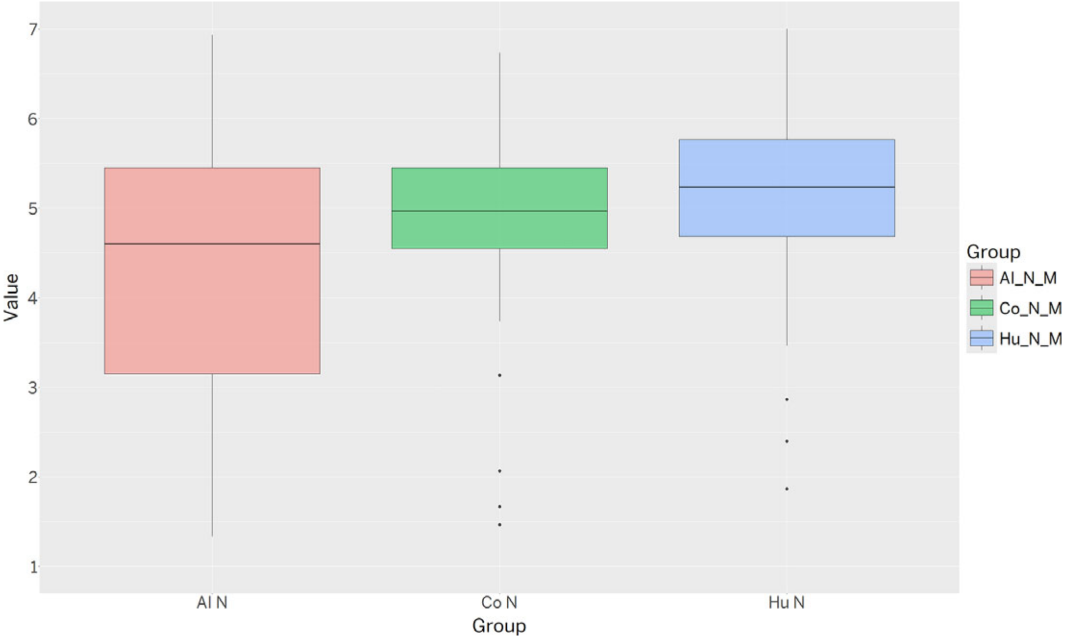

Figure 6. Mean differences in novelty across three conditions in Experiment II during the VR stage. Data visualizations were created by the author in RStudio.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

$$Co\_N\_M > \mathrm{AI\_N\_M},$$

$$Co\_R\_M > \mathrm{AI\_R\_M},$$

$$Co\_S\_M > \mathrm{AI\_S\_M}.$$

Statistical results from Experiment I – MTurk sample – rejected $H1_2$ and provided partial evidence against the expectation that Collaboration efforts would outperform AI. Likewise, Experiment II – Prolific sample – supported $H0_2$, indicating no statistical premises that Collaboration efforts outperformed AI but, in fact, even scored lower in *novelty* during *S* stage.

5.3.1. Design stage: sketch

In the *S* stage of Experiment I, a Mann–Whitney *U* with Bonferroni adjustment following the significant Kruskal–Wallis test for *novelty* showed that AI_N_M was significantly higher than *Co_N_M* ($W = 19.25$, $p < .001$), supported by a large effect size ($d = 1.09$). Kruskal–Wallis tests for *resolution* and *style* in this stage were insignificant. The results supported $H0_2$, indicating that Collaboration efforts had no advantage over AI and, in fact, underperformed in *novelty* during the *S* stage (see Figure 7).

In the *S* stage of Experiment II, following a significant Kruskal–Wallis test for *novelty*, Mann–Whitney *U* with Bonferroni adjustment offered statistical evidence for the superiority of AI_N_M over *Co_N_M* ($W = 17.79$, $p = .02$) with a

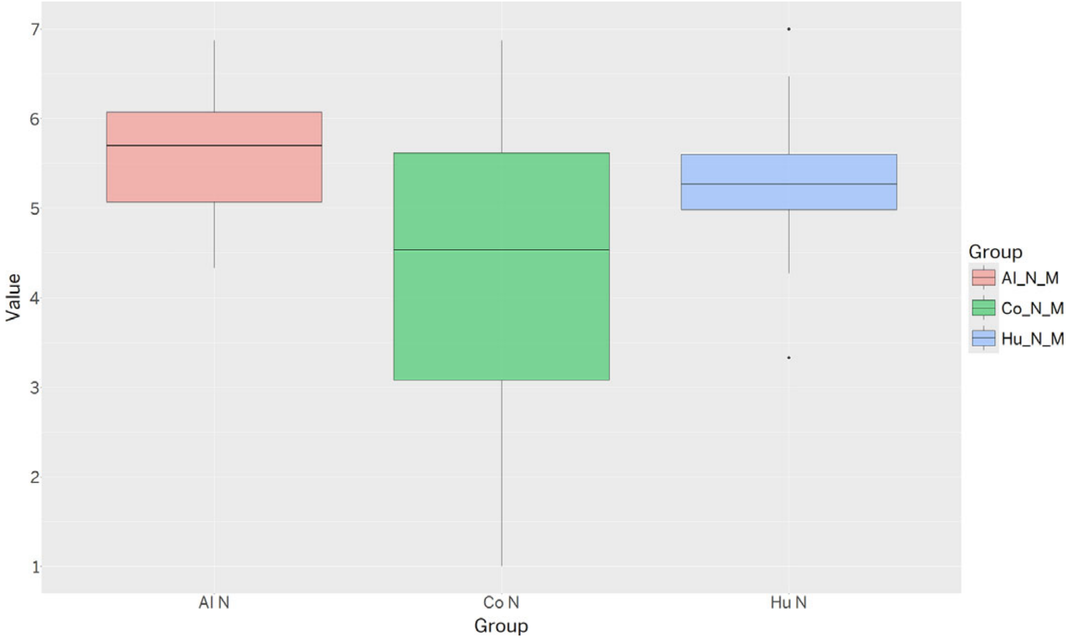

Figure 7. Mean differences in novelty across three conditions in Experiment I during the Sketch stage. Data visualizations were created by the author in RStudio.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

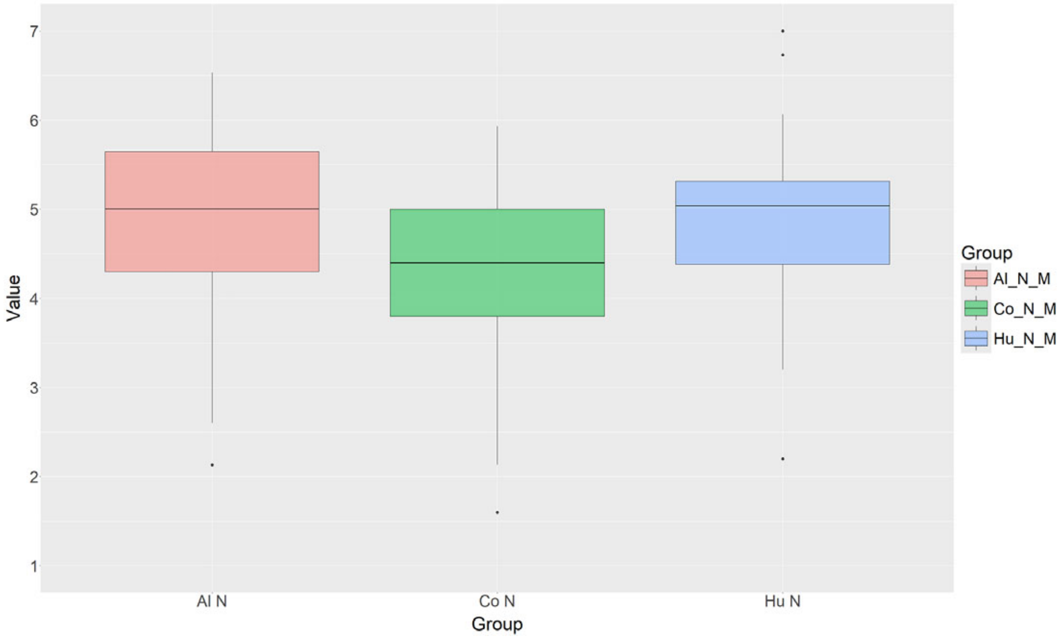

Figure 8. Mean differences in style across three conditions in Experiment II during the Sketch stage. Data visualizations were created by the author in RStudio.

medium effect size ($d = .6$), as shown in Figure 8. For *resolution* and *style* in this stage, Kruskal–Wallis tests showed no significance.

5.3.2. Design stage: 3D

In the *3D* stage of Experiment I, only the Kruskal–Wallis test for *style* was significant, yet the follow-up Mann–Whitney *U* with Bonferroni adjustment yielded no evidence ($p = .6$) regarding whether Collaboration efforts differed from AI. Kruskal–Wallis tests for the rest of the CPSS ratings showed no statistical significance. In the *3D* stage of Experiment II, the Mann–Whitney *U* with Bonferroni adjustment following the only significant Kruskal–Wallis test for *style* once again yielded no statistical result ($p > .1$) that could reject $H0_2$ in favor of $H1_2$.

5.3.3. Design stage: VR

In Experiment I, Kruskal–Wallis tests for CPSS ratings in the *VR* stage showed no statistical evidence for differences between conditions. The same observation occurred in the *VR* stage of Experiment II, as a follow-up Mann–Whitney *U* with Bonferroni adjustment for the significant Kruskal–Wallis test in *novelty* returned no evidence ($p = .35$) for differences between Collaboration efforts and AI.

5.4. H3: Human–AI Collaboration is more creative than a human designer alone

$H0_3$: There is no significant difference between Collaboration efforts and a human designer in terms of Novelty, Resolution and Style.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

$$Co\_N\_M = \text{Hu\_N\_M},$$

$$Co\_R\_M = \text{Hu\_R\_M},$$

$$Co\_S\_M = \text{Hu\_S\_M}.$$

$H1_3$: Collaboration efforts were better than a human designer in terms of Novelty, Resolution and Style.

$$Co\_N\_M > \text{Hu\_N\_M},$$

$$Co\_R\_M > \text{Hu\_R\_M},$$

$$Co\_S\_M > \text{Hu\_S\_M}.$$

Statistical results from Experiment I – MTurk sample – rejected $H1_3$, while offering partial evidence against the expectation that Collaboration efforts would outperform a human designer. Instead, in Experiment II – Prolific sample – the human designer ranked higher than Collaboration efforts in *novelty*, during both *S* and *VR* stages (refer to Figures 6 and 8), rejecting $H1_3$.

5.4.1. Design stage: sketch

In the *S* stage of Experiment I, following a significant Kruskal–Wallis test for *novelty*, the Mann–Whitney *U* with Bonferroni adjustment revealed that Hu_N_M was also higher than *Co_N_M* ($W = 1.34$, $p = .04$), even supported by a large effect size ($d = .8$). In the *S* stage of Experiment II, significant Kruskal–Wallis tests for *novelty* also occurred. A follow-up Mann–Whitney *U* with Bonferroni adjustment showed that, once again, Hu_N_M surpassed *Co_N_M* ($W = 8.45$, $p = .03$), supported by a medium effect size ($d = .60$). Thus, the human designer outperformed Collaboration in *novelty*, at least in the *S* stage (refer to Figure 7).

5.4.2. Design stage: 3D

In Experiment I, Kruskal–Wallis tests were insignificant across CPSS criteria and conditions. Hence, there was no evidence for the superiority of Collaboration efforts toward the human designer. In Experiment II, while the Kruskal–Wallis test was significant, the follow-up Mann–Whitney *U* with Bonferroni adjustment yielded no evidence ($d = .07$) for differences in *style* during the 3*D* stage.

5.4.3. Design stage: VR

In Experiment I, Kruskal–Wallis tests showed no significant differences across CPSS criteria and conditions. On the contrary, in Experiment II, the Kruskal–Wallis test for *novelty* during the *VR* stage indicated that Hu_N_M was higher than *Co_N_M* ($W = 8.60$, $p = .04$), with a medium effect size ($d = .62$). Kruskal–Wallis tests for other criteria were insignificant. Across both experiments, there was no evidence supporting the superiority of Collaboration efforts over the human designer. In fact, in Experiment II, the human designer outperformed Collaboration in terms of *novelty*.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

6. Discussion

This research aimed to answer the overarching question: Can a machine be creative, as measured against the proximity of its outcomes to an established standard of creativity? The two samples of crowd-sourced judges, MTurk ($n = 120$) and Prolific ($n = 126$), showed variations in CPSS ratings, yet these discrepancies remained within one to two standard deviations. The author further employed non-parametric Kruskal–Wallis tests in place of one-way ANOVA based on violations of normality and homogeneity confirmed by Shapiro–Wilk and Levene tests. As normality violations occurred due to demographic variances, Kruskal–Wallis was a robust alternative to ANOVA for comparing AI, Collaboration and Human conditions. This non-parametric approach ranks data across groups, reducing the influence of outliers and unequal variances (Yazici & Yolacan 2007; Lock *et al.* 2013), thereby enhancing the reliability of statistical analyses on CPSS ratings for creativity. Overall, statistical results for CPSS ratings, on *novelty*, *resolution* and *style*, throughout the design stages – S, $3D$ and VR – rejected $H1_1$ in Experiment I yet supported the same hypothesis in Experiment II. However, both experiments endorsed $H0_2$ and $H0_3$ (see Table 3).

Moreover, in both experiments, crowd-sourced judges reported similar emotions across three conditions – AI, Collaboration and Human. According to Norman (2005) and Horn & Salvendy (2009), emotional responses signified a facet of *resolution*, a coherence between function and appearance of light fixtures that elicited the senses of "comfort" and "engaging" suitable for a pediatric waiting room. Overall, participants reported feeling happy ($6.83 \leq M \leq 6.90$; $6.36 \leq M \leq 7.10$) and calm ($4.63 \leq M \leq 5.35$; $3.95 \leq M \leq 5.12$) across experiments, with those in the Human condition of Experiment II being the happiest. While the means for feeling calm had wider intervals, variations among participants remained within one to two standard deviations ($1.52 \leq SD \leq 2.38$; $1.74 \leq SD \leq 2.20$). Thus, in addition to the CPSS ratings, these self-reported measures showed that AI, Human–AI collaboration and the human designer were all able to qualify the criterion of *resolution* by producing designs that evoked desirable emotions as required in the Robert Bruce Thompson Lighting Design Competition's brief.

6.1. Creativity in the sketch stage: collaboration finished last

Although the studies of Zhou *et al.* (2023), Chandrasekera *et al.* (2024) and Chen *et al.* (2025) supported the superiority in creativity of the synergy between humans

Table 3. Statistical results for each hypothesis. Created by the author

Hypothesis	Stage	Criteria	Significance (p and effect size)	Experiment
$H0_1$ supported	$3D$	Style	$AI > Human, p = .02\,(medium)$	I (MTurk)
$H1_1$ supported	VR	Novelty	$Human > AI, p = .04\,(medium)$	II (Prolific)
$H0_2$ supported	S	Novelty	$AI > Co, p = .0002\,(large)$	I (MTurk)
$H1_2$ rejected	S	Novelty	$AI > Co, p = .02\,(medium)$	II (Prolific)
$H0_3$ supported	S	Novelty	$Human > Co, p = .004\,(large)$	I (MTurk)
$H1_3$ rejected	S	Novelty	$Human > Co, p = .03\,(medium)$	II (Prolific)

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

and AI, Experiments I and II demonstrated that both AI and the human designer surpassed Collaboration in *novelty* at the *S* stage. These surprising results offer empirical evidence for the independence of general AI in exercising creative ability. ChatGPT-4 produced 73 textual prompts, and Midjourney subsequently generated 292 sketches. Human oversight was minimal, except for inputting the 2024 Robert Bruce Thompson Lighting Design Competition brief into ChatGPT-4 and filtering irrelevant outputs from Midjourney. Runco *et al.* (2024) also put three general AI tools – Bard, ChatGPT-3.5 and ChatGPT-4 (all generative text applications) – to the test of divergent thinking. Their results indicated that ChatGPT-4's mean score in idea density per word ($M = 0.49$) came close to that of human participants ($M = 0.60$) doing the same divergent thinking test in a separate study. Again, there was minimal intervention from the researchers, only removing irrelevant verbiages. In other words, the machine can generate a myriad of conceptual designs, simulating the divergent phase of the DD framework. Hence, this study's findings further support the evidence that general AI tools can exhibit at least one aspect of creativity: *novelty.*

The underperformance of Collaboration in this study was unexpected. While AI's inherent randomness (Tørresen 2021) highlights the crucial role of human judgment in selecting appropriate solutions for the design problems at hand (Huang *et al.* 2021; Babakhani 2023; Morrison *et al.* 2023; Abuzuraiq & Pasquier 2024), prior research indicates that Collaboration harnesses human ability to link distant or adjacent domains across numerous AI-generated solutions, thereby maximizing creative outcomes (Grilli & Pedota 2024). However, the statistical results of this study contradict that assumption.

A possible explanation lies in the novice status of the student designer in the Collaboration condition; as Sarica & Luo (2024) note, less experienced designers often struggle to manage large volumes of AI-generated outputs and lack the domain knowledge needed to synthesize diverse ideas, potentially reducing originality. Although Chen *et al.* (2025) provided empirical evidence that human–AI collaboration outperformed humans alone in multiple facets of creativity like novelty and conceptual diversity, among novice product design students, it is important to note the timeline difference between their experiment and this study. Participants in Chen *et al.* (2025) completed either a baby chair or tangible music bricks in 20 minutes, whereas the student designer in the Collaboration condition developed the light fixture during 5 weeks. More importantly, Chen *et al.*'s (2025) design briefs were hypothetical and intended for short-term experimental settings, whereas in this study, the Robert Bruce Thompson Lighting Design Competition is a well-established annual event, grounded in real-world design practice. Thus, the 5-week duration of the Collaboration condition was not an arbitrary choice but a direct reflection of the competition's authentic timeline and typical studio design practices, distinguishing this study from short, hypothetical experimental tasks. However, an authentic and extended design timeline may have introduced factors such as shifts in motivation, idea evolution or – particularly – design fixation, as prolonged reflection can lead to overcommitment to a limited set of favored ideas. Henceforth, within the operational context of the DD framework, emphasizing linear, step-by-step progression through divergent and convergent thinking, general AI tools may offer less of an advantage than they do in rapid briefs designed for experimental control.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

A further explanation for the underperformance of Collaboration may lie in prompt engineering proficiency required for effective human–AI ideation (Han *et al.* 2024; Shaer *et al.* 2024; Hwang & Lee 2025). According to Han *et al.* (2024), art and design students ($n = 18$) – including those in product design and interior architecture – who used ChatGPT 3.5 and Midjourney to ideate poem illustrations, found prompting to be burdensome, often discouraging them from further attempts to achieve desired results. Moreover, students in Han *et al.*'s (2024) study preferred collaborating with human designers, citing the rigidity of AI tools such as ChatGPT-3.5 and Midjourney – capturing the literal meaning of prompts while failing to grasp the emotional and contextual nuances that humans could understand. Shaer *et al.* (2024) and Hwang & Lee (2025) also emphasized the importance of prompt literacy in AI–human collaboration, noting student struggles in communicating their ideas to AI tools. Thus, as novice students lack literacy in AI prompting, they often encounter hindrances in productivity and might perform lower in collaboration with this technology. Overall, these studies argue that building prompt literacy is essential for students to fully benefit from AI as a creative collaborator.

The advantage of the human designer in *novelty* challenges the results of Zhou *et al.* (2023) and Chen *et al.* (2025), which suggested that those who collaborated with AI dominated in both quantity and quality of product design solutions. In this research, the interior design student curated 10 sketches from a large AI-assisted pool, whereas the human designer (a generalist) – with basic design heuristics and CAD proficiency but no lighting expertise – manually created three distinct sketches for the entire design process. While the human designer showed variety in just a few design options (see Figure 4), the student exhibited redundant iterations across their AI-generated sketches (see Figure 3). These inconclusive findings highlight the gap in our current understanding of AI's creative potential and the uncertainties surrounding the role of this evolving technology in the design process. Above all, this study underscores the need for more nuanced studies regarding general AI's ability not only in specific facets of creativity but also in relation to its performance during distinct phases of the design process.

6.2. Creativity in the 3D stage: AI's strength in style

MTurk workers rated AI significantly higher in *style* during the 3*D* stage, with a medium effect size, challenging Boudier *et al.*'s (2023) claim that AI's primary contribution is mitigating design fixation rather than enhancing stylistic quality. Such a result also contradicts Chen *et al.*'s (2025) findings that human–AI collaboration outperformed human-alone both in *novelty* and *resolution*, as an umbrella term for feasibility and usability according to Besemer (2006). Therefore, this study provides new evidence of general AI's creative potential, demonstrating a stylistic edge over the human designer in 3D visuals. However, while Prolific workers also rated AI higher, there was no statistical evidence of significant differences compared to the Collaboration and Human conditions. These inconclusive findings both support and challenge assumptions about AI, humans and creativity in the current literature.

One important question derived from this study is whether general AI's leverage in *style* – as reflected via visualizations of design solutions – shapes a perceived superior creativity in human–AI collaboration. For instance, in

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Chandrasekera *et al.* (2024), the AI group produced full-color, fully materialized 3D renderings, whereas the non-AI group submitted only 2D line sketches – with shapes outlined in lines and minimal textual annotations. Given the disparity in visual presentation, the two expert judges may favor the polished outputs of the AI group, potentially awarding higher overall creativity scores as averaged across CPSS criteria. On the contrary, this study evaluates each CPSS dimension independently, allowing for a more nuanced understanding of how general AI performs across different aspects of creativity and reducing the likelihood of Type II errors – that is, missing meaningful differences by collapsing distinct criteria into a single score.

6.3. Creativity in the VR stage: human designers maintain an edge in novelty

Prolific workers favored the human designer in terms of *novelty* during the *VR* stage, while statistical analyses showed no evidence of differences between design conditions in the MTurk sample. Despite the large disparity in the number of concept sketches, the human designer outperformed both AI and human–AI collaboration with a medium effect size. These results support the portion of literature that champions the creativity of human designers over general AI, such as Huang *et al.* (2021), Tørresen (2021), Morrison *et al.* (2023), and Abuzuraiq & Pasquier (2024). This study also echoes what Runco *et al.* (2024) and Wang *et al.* (2024) reported on the human edge in creating “blue sky” ideas as opposed to general AI tools generating pragmatic and repeated concepts. After all, *novelty* remains the foremost facet of creativity that human designers excel in despite general AI’s edge in *style*.

As this study contradicts Chandrasekera *et al.* (2024) and Chen *et al.* (2025), both aiding the superiority of human–AI collaboration in creativity, it suggests that the proximity between solutions generated by general AI, a human designer and a human–AI collaboration might be determine by *novelty* alone. Such a unilateral view reminds us that our understanding of creativity is shaped by human biases – raising awareness of whether humans are gatekeeping creativity by favoring human-generated work while discounting the contributions of general AI, as found by Horton Jr *et al.* (2023) and Magni *et al.* (2024). Above all, our biases might shape not only creativity measures but also definitions of *novelty* – what is deemed new and worthy of recognition. Ultimately, this study calls for a reflective and inclusive approach to creativity assessment in the age of machines.

6.4. Implications for understanding AI’s creative capacity

This study suggests several implications for how we understand creativity in the age of general AI, particularly within product and interior design disciplines. First, research on general AI’s creative capacity should adopt a more granular approach – focusing on specific facets of creativity and performance at distinct phases of the design process (*divergent* and *convergent* as operationalized in the DD framework). Second, AI’s leverage in *style* might influence the perceived superiority in creativity of human–AI collaboration. Third, how humans define and acknowledge *novelty* may represent the final frontier that differentiates AI-generated work from that of human designers. Runco (2024) argued that AI demonstrates discovery and

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

innovation, but not true creativity, as it lacks the authenticity grounded in human experiences and the cognitive processes that define creative thought. While this perspective highlights important distinctions, it is equally important to recognize the current capabilities and limitations of general AI tools. These systems do not possess the full agility or contextual flexibility of the human brain (Ameen *et al.* 2022). More importantly, this study does not compare the internal processes of AI, human–AI collaboration and human designers; rather, it evaluates the creative outcomes using Besemer's (2006) well-established CPSS measure. The empirical findings indicate that, in certain aspects at specific phases – particularly during early ideation (*S* stage) and stylistic development (3*D* stage) – general AI's creative outputs matched or even surpassed those of the human designer.

While prior literature highlights strong convergence between large-scale crowd-sourced assessments and expert judgments (Yuan *et al.* 2016; Foong *et al.* 2017), the current finding – specifically, the lower creativity ranking for Collaboration – differs from results reported by Zhou *et al.* (2023) and Chandrasekera *et al.* (2024), who utilized expert judges and observed superior performance for human–AI collaboration. Table 4 provides a snapshot of general AI's capacity in certain aspects of creativity at specific phases of the design process to better situate this study's findings in current literature.

The unexpected underperformance of the Collaboration condition in a critical measure of creativity – *novelty* – raises questions about the levels of human judgment and expertise required, especially above novice level, in contexts demanding nuanced, context-sensitive solutions. Given the student's dedication to collaboration efforts – drafting text-to-image prompts, selecting and revising options – the consistent advantage of AI in *novelty* also suggests that AI has an edge in executing creative tasks with minimal human oversight. Regardless, further investigation into the role of AI in creative design processes remains crucial. This study's findings suggest that AI is better suited to supporting rapid visualization (Wang *et al.* 2024) or ideation (Zhou *et al.* 2023; Chen *et al.* 2025) rather than acting as a collaboration partner, particularly for novice designers like students. Instead, AI may primarily assist expert designers in quickly visualizing preliminary concept sketches.

This study evaluates the creative capacity of current general AI tools, rather than speculating on the optimal test for AI versus human cognitive abilities, which remain uncertain until the emergence of autonomous super AI. Despite said limitations, this study's findings provide concise and valuable insights into the current capabilities of general AI tools. Guided by the Turing test – a foundational benchmark in AI creativity research (Gonçalves 2023; Wang *et al.* 2023) – these results inform ongoing debates about AI's creative potential. Gonçalves (2023) emphasized that, despite varied interpretations of the Turing test reducing its legitimacy, this benchmark is still an empirical criterion for assessing machine intelligence – or machine creativity in this study – when used in controlled experiments.

Overall, this research is significant in three ways. First, it demonstrates general AI's ability to rapidly generate and visualize preliminary design concepts, offering practical value in early-stage ideation. Second, the results provide empirical evidence supporting the notion that general AI, even without advanced intuitive capabilities, can produce substantial creative outputs, thus contributing to the ongoing, inconclusive discourse regarding AI's creative potential. Third, the

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Table 4. A snapshot of general AI's creative capacity in current literature. Created by the author

Study	Design brief	Discipline	General AI's successful criteria and phases
Zhou *et al.* (2023) *Collaboration*	Luggage	Product	Originality, novelty, practicality *Conceptual*
Gallega & Sumi (2024) *Collaboration*	Material	Interior	72.82/100 CSI (perceived effectiveness) *Conceptual*
Wang *et al.* (2024) *AI*	Residential	Interior	"Good" to "Very Good" (overall evaluation) *Visualization*
Chandrasekera *et al.* (2024) *Collaboration*	Chair	Interior	Overall creativity (averaged across criteria) *Conceptual/Visualization*
Chen *et al.* (2025) *Collaboration*	Baby chair Music bricks	Product Product	Novelty, feasibility, usability, functionality *Conceptual*
This study *AI, Human* and *Collaboration*	Light fixture	Product/ Interior	Style (> Human, > Collaboration) *Visualization* Novelty (< Human, > Collaboration) *Conceptual*

Note: *Conceptual* equals *Sketch* phase, and *Visualization* equals *3D* phase. See Section 1.5 for a detailed discussion on Chandrasekera *et al.* (2024).

findings advocate for broader evaluation methodologies that integrate perspectives from both general audiences and expert evaluators to comprehensively assess creativity across AI-generated, human–AI collaborated and human-designed solutions. Ultimately, these results prompt a reconsideration of creativity itself: while traditionally seen as a social construct (Amabile 1983; Csikszentmihalyi 1988) and measured by established criteria (Besemer 2006; Horn & Salvendy 2006; O'Quin & Besemer 2011), its definition may be evolving into a hybrid concept encompassing human–technology interactions.

7. Conclusions

This research offers new insights into AI's creative outputs, backed by empirical evidence solicited from two prominent crowd-sourced platforms: Amazon MTurk and Prolific. While the findings were mixed – partly confirming $H1_1$ and rejecting $H1_2$ and $H1_3$ – this research challenges prevailing assumptions in design literature regarding AI's creative limitations. In this study, AI demonstrated a capacity for *style* and *novelty* that, in certain design stages, surpassed both Collaboration and Human conditions. These findings position general AI tools as valuable assets for visualization and ideation, particularly in the early process (S stage) of product and interior design.

Above all, this research contributes not only to the question of whether a machine can be creative but also to the ongoing discussion of how we should measure (or ultimately define) this construct. For instance, Barbot *et al.* (2023) and Myszkowski (2024) suggested that VR is a powerful context for assessing the

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

creativity of design outcomes, as this environment facilitates a deeper understanding for the judges. Horton Jr *et al.* (2023) and Magni *et al.* (2024) also highlight the existing bias humans have toward work labeled as AI-generated. Their studies showed that participants could not distinguish between AI- and human-created work, and in Horton Jr *et al.* (2023) case, participants rated lesser-known pieces by famous artists as less creative when they were disguised as AI-generated. Thus, this study evaluates creative outcomes rather than internal cognitive processes and raises awareness of whether humans are gatekeeping creativity and harboring bias against AI.

However, the findings also underscore the continued importance of human insight, particularly in later stages of the design process where context-sensitive judgments and nuanced creativity are critical. For instance, with the human designer outperforming both AI and Collaboration in *novelty* across stages and experiments, the quantity of initial concepts is no longer a reliable indication for divergent thinking in the DD framework. As the findings suggest, AI's diminishing returns can hinder novice designers – as in the Collaboration condition – with repeated iterations of familiar concepts. On the contrary, the human designer's ability to achieve the highest CPSS novelty ratings with the fewest preliminary sketches opens an intriguing research path, suggesting the need for future studies that compare crowd-sourced and expert ratings for creativity.

In addition, future research should examine how prompt engineering practices influence human–AI collaboration. Shaer *et al.* (2024) provide a useful model: students generated ideas individually, then iterated prompts with AI tools like ChatGPT through phases of framing, exchanging, generating, narrowing, enhancing and selecting. Investigating such a structured workflow could clarify how prompt literacy shapes the novelty of AI-assisted designs. These new endeavors should also clarify both the cognitive demands of prompt engineering and identify strategies that enable novice designers to collaborate more effectively with AI tools.

Finally, for product and interior designers, these findings offer clear, actionable insights for integrating AI into their design stages.

- **Leverage AI for Early-Stage Ideation:** Product and interior designers can utilize general AI tools to rapidly generate diverse preliminary concepts, making AI particularly valuable during the early stages of the design process when broad exploration is essential.
- **Use AI to Enhance Stylistic Exploration:** General AI demonstrated strength in producing stylistically compelling outputs, especially in 3D representations. Product and interior designers can employ AI to quickly explore and visualize different style options, refining these outputs as part of their iterative process.
- **Balance AI-Generated Outputs with Human Judgment:** While general AI excels in producing volume and variation, product and interior designers should apply critical judgment to avoid design fixation and ensure novelty and appropriateness. This is especially important in later stages when solutions must be context-sensitive and aligned with user needs.
- **Educators Should Train Critical Evaluation Skills:** Product and interior design educators should guide students in critically assessing AI-generated outputs, teaching them how to identify redundancy, refine concepts manually and integrate human creativity to enhance originality and problem-solving.
- **Develop Prompt Literacy as a Design Skill:** Product and interior designers should practice and refine prompt engineering techniques – such as structured,

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

iterative prompting – to better communicate design intent with AI tools. This skill can enhance the quality of AI-assisted outputs, ensuring that human–AI collaboration becomes a productive part of the design process rather than a hindrance.

- **Prepare for a Hybrid Creative Process:** As general AI tools continue to evolve, designers should anticipate a hybrid model of creativity that combines human intuition and technological capabilities. Understanding how to navigate this interaction will be key to maintaining creative agency and relevance in the design profession.

Ultimately, this study positions general AI not as a replacement for human creativity but as a complementary tool within the design stages, to be used with caution (e.g., considering its impacts on novice vs. expert designers). By providing empirical evidence of AI's creative aspects – *novelty* and *style* – this research encourages designers to move beyond skepticism and embrace AI as a practical resource in their creative endeavors within the design professions. Ultimately, this study calls for a more comprehensive understanding of creativity – one that recognizes where general AI excels, where it falls short, and how its strengths can be meaningfully integrated without overshadowing human novelty. As our definitions of creativity evolve, so too must our measures for it, ensuring that human and machine contributions are understood not in competition, but in collaboration. Finally, the competitive edges of general AI may even stimulate creativity by motivating human designers to polish and differentiate their work. Above all, the rapid advancement of AI technologies marks a new era in design research – one in which questions of creativity, particularly in human–AI collaboration, must be reconsidered in light of machines' evolving capabilities and our ability to harness and manage them.

Acknowledgements

The author thanks Caroline Alleva for her assistance with graphical representations in the Method section and acknowledges the use of ChatGPT-4 for grammar refinement and troubleshooting statistical codes in RStudio.

Financial support

This research was supported by the 2024 Eskenazi Technology Innovation Fellowship, awarded by the Eskenazi Technology Innovation Lab (ETIL) at the Eskenazi School of Art, Architecture + Design, Indiana University Bloomington.

Competing interest

The author declares none.

References

Abuzuraiq, A. M. & **Pasquier, P.** 2024 Towards personalizing generative AI with small data for co-creation in the visual arts. In *CEUR Workshop Proceedings: Proceedings of the 4th Workshop on Human–AI Co-Creation (HAI-GEN 2024)*, pp. 1–14. CEUR-WS. Available at: https://ceur-ws.org/Vol-3660/paper11.pdf (Accessed: 13 September 2025).

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Ahmed, F. & **Fuge, M.** 2018 Creative exploration using topic-based bisociative networks. *Design Science* **4**, e5; doi:10.1017/dsj.2018.5.

Amabile, T. M. 1982 Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology* **43** (5), 997–1013; doi:10.1037/0022-3514.43.5.997.

Amabile, T. M. 1983 The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology* **45** (2), 357–376; doi:10.1037/0022-3514.45.2.357.

Ameen, N., **Sharma, G. D.**, **Tarba, S.**, **Rao, A.** & **Chopra, R.** 2022 Toward advancing theory on creativity in marketing and artificial intelligence. *Psychology & Marketing* **39** (9), 1802–1825; doi:10.1002/mar.21707.

Anantrasirichai, N. & **Bull, D.** 2022 Artificial intelligence in the creative industries: A review. *Artificial Intelligence Review* **55**, 1–68; doi:10.1007/s10462-021-10066-8.

Babakhani, R. 2023 Automatic generation of architectural plans with machine learning. *Technology | Architecture + Design* **7** (2), 183–191; doi:10.1080/24751448.2023.2187324.

Barbot, B., **Kaufman, J. C.** & **Myszkowski, N.** 2023 Creativity with 6 degrees of freedom: Feasibility study of visual creativity assessment in virtual reality. *Creativity Research Journal* **35** (7), 783–800; doi:10.1080/10400419.2023.2213389.

Besemer, S. P. 1998 Creative product analysis matrix: Testing the model structure and a comparison among products – Three novel chairs. *Creativity Research Journal* **11** (4), 333–346; doi:10.1207/s15326934crj1104_5.

Besemer, S. P. 2006 *Creating Products in the Age of Design: How to Improve Your New Product Ideas!* New Forums Press.

Besemer, S. P. & **O'Quin, K.** 1999 Confirming the three-factor creative product analysis matrix model in an American sample. *Creativity Research Journal* **12** (4), 287–296; doi:10.1207/s15326934crj1204_5.

Boden, M. A. 1991 *The Creative Mind: Myths & Mechanisms*. Routledge.

Boden, M. A. 1998 Creativity and artificial intelligence. *Artificial Intelligence* **103** (1–2), 347–356; doi:10.1016/S0004-3702(98)00055-1.

Boudier, J., **Sukhov, A.**, **Netz, J.**, **Le Masson, P.** & **Weil, B.** 2023 Idea evaluation as a design process: Understanding how experts develop ideas and manage fixations. *Design Science* **9**, e9; doi:10.1017/dsj.2023.9.

Bouschery, S. G., **Blazevic, V.** & **Piller, F. T.** 2023 Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management* **40** (2), 139–153; doi:10.1111/jpim.12655.

Bradley, M. M. & **Lang, P. J.** 1994 Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy & Experimental Psychiatry* **25** (1), 49–59; doi:10.1016/0005-7916(94)90063-9.

Chandrasekera, T., **Hosseini, Z.** & **Perera, U.** 2024 Can artificial intelligence support creativity in early design processes? *International Journal of Architectural Computing* **23** (1), 122–136; doi:10.1177/14780771231195312.

Chen, L., **Song, Y.**, **Guo, J.**, **Sun, L.**, **Childs, P.** & **Yin, Y.** 2025 How generative AI supports human in conceptual design. *Design Science* **11**, e9; doi:10.1017/dsj.2025.9.

Cross, N. 2006 *Designerly Ways of Knowing*. Springer.

Csikszentmihalyi, M. 1988 Society, culture, and person: A systems view of creativity. In *The Nature of Creativity: Contemporary Psychological Perspectives* (ed. R. J. Sternberg), pp. 325–339. Cambridge University Press.

Design Council. 2022 What is the framework for innovation? Design council's evolved double diamond. *Design Council* **2022**. Available at: https://www.designcouncil.org.uk/

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

our-work/skillslearning/tools-frameworks/framework-for-innovation-designcouncils-evolved-double-diamond/ (Accessed: 17 July 2024).

Flus, M. & **Hurst, A.** 2021 Design at hackathons: New opportunities for design research. *Design Science* **7**, e4; doi:10.1017/dsj.2021.4.

Fong, C. J., **Warner, J. R.**, **Williams, K. M.**, **Schallert, D. L.**, **Chen, L. H.**, **Williamson, Z. H.** & **Lin, S.** 2016 Deconstructing constructive criticism: The nature of academic emotions associated with constructive, positive, and negative feedback. *Learning and Individual Differences* **49**, 393–399; doi:10.1016/j.lindif.2016.06.013.

Foong, E., **Gergle, D.** & **Gerber, E. M.** 2017 Novice and expert sensemaking of crowd-sourced design feedback. *Proceedings of the ACM on Human–Computer Interaction* **1** (CSCW), 45; doi:10.1145/3134680.

Fraser, C.A., **Ngoon, T.J.**, **Weingarten, A.S.**, **Dontcheva, M.** & **Klemmer, S.** 2017 CritiqueKit: A mixed-initiative, real-time interface for improving feedback. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology*, pp. 7–9. Association for Computing Machinery. doi:10.1145/3131785.3131797.

Galati, F. 2015 Complexity of judgment: What makes possible the convergence of expert and nonexpert ratings in assessing creativity. *Creativity Research Journal* **27** (1), 24–30; doi:10.1080/10400419.2015.992678.

Gallega, R. W. & **Sumi, Y.** 2024 Exploring the use of generative AI for material texturing in 3D interior design spaces. *Frontiers in Computer Science* **6**, 1493937; doi:10.3389/fcomp.2024.1234567.

Ghasemi, P., **Yuan, C.**, **Marion, T.** & **Moghaddam, M.** 2024 DCG-GAN: Design concept generation with generative adversarial networks. *Design Science* **10**, e14; doi:10.1017/dsj.2024.14.

Goel, V. & **Pirolli, P.** 1992 The structure of design problem spaces. *Cognitive Science* **16** (3), 395–429; doi:10.1207/s15516709cog1603_3.

Goldschmidt, G. 2019 Design creativity research: Recent developments and future challenges. *International Journal of Design Creativity and Innovation* **7** (3), 194–195; doi:10.1080/21650349.2019.1675235.

Gonçalves, B. 2023 The Turing test is a thought experiment. *Minds & Machines* **33**, 1–31; doi:10.1007/s11023-022-09621-8.

Görzen, T. & **Kundisch, D.** 2019 When in doubt follow the crowd: How idea quality moderates the effect of an anchor on idea evaluation. In *Proceedings of the 27th European Conference on Information Systems (ECIS 2019)*, Stockholm-Uppsala, Sweden.

Goucher-Lambert, K. & **Cagan, J.** 2019 Crowdsourcing inspiration: Using crowd generated inspirational stimuli to support designer ideation. *Design Studies* **61**, 1–29; doi:10.1016/j.destud.2019.01.002.

Grilli, L. & **Pedota, M.** 2024 Creativity and artificial intelligence: A multilevel perspective. *Creativity and Innovation Management* **33** (2), 234–247; doi:10.1111/caim.12593.

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

Guilford, P. 1967 *The Nature of Human Intelligence*. McGraw-Hill.

Han, Y., **Qiu, Z.**, **Cheng, J.** & **Ray, L. C.** 2024 When teams embrace AI: Human collaboration strategies in generative prompting in a creative design task. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. Association for Computing Machinery. doi:10.1145/3544548.3580984.

Horn, D. & **Salvendy, G.** 2006 Consumer-based assessment of product creativity: A review and reappraisal. *Human Factors and Ergonomics in Manufacturing & Service Industries* **16** (2), 155–175; doi:10.1002/hfm.20046.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Horn, D. & **Salvendy, G.** 2009 Measuring consumer perception of product creativity: Impact on satisfaction and purchasability. *Human Factors and Ergonomics in Manufacturing & Service Industries* **19** (3), 223–240; doi:10.1002/hfm.20198.

Horton Jr, C. B., **White, M. W.** & **Iyengar, S. S.** 2023 Bias against AI art can enhance perceptions of human creativity. *Scientific Reports* **13**, 19001; doi:10.1038/s41598-023-44907-0.

Huang, J., **Johanes, M.**, **Kim, F. C.**, **Doumpioti, C.** & **Holz, G. C.** 2021 On GANs, NLP and architecture: Combining human and machine intelligences for the generation and evaluation of meaningful designs. *Technology | Architecture + Design* **5** (3), 207–224; doi:10.1080/24751448.2021.1945847.

Hwang, Y. & **Lee, J. H.** 2025 Exploring students' experiences and perceptions of human–AI collaboration in digital content making. *International Journal of Educational Technology in Higher Education* **22**, 44; doi:10.1186/s41239-025-00411-3.

Jackman, J., **Ryan, S.**, **Olafsson, S.** & **Dark, V. J.** 2017 *Metaproblem Spaces and Problem Structure*. Taylor & Francis.

Jun, G. T., **Hignett, S.** & **Clarkson, P. J.** 2024 *Design Creativity*. Cambridge University Press.

Kayode, F., **Ojo, B.** & **Sheba, E. A.** 2008 Design, aesthetics and the issue of integrity in the built environment: The Nigerian example. *Indoor and Built Environment* **17** (3), 283–298; doi:10.1177/1420326X08091253.

Kim, Y. S. & **Park, J. A.** 2021 Design thinking in the framework of visual thinking and characterization of service design ideation methods using visual reasoning model. *The Design Journal* **24** (6), 931–953; doi:10.1080/14606925.2021.1988275.

Kudrowitz, B. M. & **Wallace, D.** 2013 Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design* **24** (2), 120–139; doi:10.1080/09544828.2012.676633.

Lamb, C., **Brown, D. G.** & **Clarke, C. L.** 2016. Evaluating digital poetry: Insights from the cat. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016)*, Paris, France.

Liang, J., **Shan, X.** & **Chung, J.** 2023 A study on process of creating 3D models using the application of artificial intelligence technology. *International Journal of Advanced Culture Technology* **11** (3), 346–351; doi:10.17703/IJACT.2023.11.3.346.

Lloyd, D. 2024 What is it like to be a bot? The world according to GPT-4. *Frontiers in Psychology* **15**, 1292675.

Lock, R. H., **Lock, P. F.**, **Morgan, K. L.**, **Lock, E. F.** & **Lock, D. F.** 2013 *Statistics: Unlocking the Power of Data*. Wiley.

Magni, F., **Park, J.** & **Chao, M. M.** 2024 Humans as creativity gatekeepers: Are we biased against AI creativity? *Journal of Business and Psychology* **39**, 643–656; doi:10.1007/s10869-023-09936-x.

Martins Pacheco, N. M., **Geisler, M.**, **Bajramovic, M.**, **Fu, G.**, **Vazhapilli Sureshbabu, A.**, **Mörtl, M.** & **Zimmermann, M.** 2024 Learning by doing? The relationship between effort, learning effect and product quality during hackathons of novice teams. *Design Science* **10**, e9; doi:10.1017/dsj.2024.9.

Meron, Y. & **Tekmen Araci, Y.** 2023 Artificial intelligence in design education: Evaluating ChatGPT as a virtual colleague for post-graduate course development. *Design Science* **9**, e30; doi:10.1017/dsj.2023.30.

Merrotsy, P. 2013 A note on big-C creativity and little-c creativity. *Creativity Research Journal* **25** (4), 474–476; doi:10.1080/10400419.2013.843395.

Miceli, G. & **Raimondo, M. A.** 2020 Creativity in the marketing and consumer behavior literature: A structured review and a research agenda. *Italian Journal of Marketing* **1**, 1–40; doi:10.1007/s43039-020-00006-7.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Morrison, B. W., **Kelson, J. N.**, **Morrison, N. M. V.**, **Innes, J. M.**, **Zelic, G.**, **Al-Saggaf, Y.** & **Paul, M.** 2023 You're not the boss of me, algorithm: Increased user control and positive implicit attitudes are related to greater adherence to an algorithmic aid. *Interacting with Computers* **35** (4), 452–460; doi:10.1093/iwc/iwad016.

Myszkowski, N. 2024 *Item Response Theory for Creativity Measurement*. Cambridge University Press.

Norman, D. A. 2005 *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books.

O'Quin, K. & **Besemer, S. P.** 2011 Creative products. In *Encyclopedia of Creativity*, 2nd Edn (ed. M. A. Runco & S. R. Pritzker), pp. 367–372. Academic Press.

Palan, S. & **Schitter, C.** 2018 Prolific.Ac – A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* **17**, 22–27; doi:10.1016/j.jbef.2017.12.004.

Peer, E., **Brandimarte, L.**, **Samat, S.** & **Acquisti, A.** 2017 Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* **70**, 153–163; doi:10.1016/j.jesp.2017.01.006.

Raghunath, N., **Koronis, G.**, **Karthikayen, R.**, **Silva, A.** & **Yogiaman, C.** 2023 A social science mixed-methods approach to stimulating and measuring creativity in the design classroom. *Design Science* **9**, e34; doi:10.1017/dsj.2023.34.

Rahman, M. H., **Bayrak, A. E.** & **Sha, Z.** 2024 Empirical evidence and computational assessment on design knowledge transferability. *Design Science* **10**, e10; doi:10.1017/dsj.2024.10.

Raman, R., **Kowalski, R.**, **Achuthan, K.**, **Iyer, A.** & **Nedungadi, P.** 2025 Navigating artificial general intelligence development: Societal, technological, ethical, and brain-inspired pathways. *Scientific Reports* **15**, 8443; doi:10.1038/s41598-025-84843-0.

Rhodes, M. 1961 An analysis of creativity. *The Phi Delta Kappan* **42** (7), 305–310. Available at: https://www.jstor.org/stable/20342603.

RStudio Team 2023 *RStudio: Integrated Development for R*. RStudio, PBC.

Runco, M. A. 2024 The discovery and innovation of AI does not qualify as creativity. *Journal of Cognitive Psychology*, 1–10. doi:10.1080/20445911.2024.9999999.

Runco, M. A. & **Jaeger, G. J.** 2012 The standard definition of creativity. *Creativity Research Journal* **24** (1), 92–96; doi:10.1080/10400419.2012.650092.

Runco, M. A., **Turkman, B.**, **Acar, S.** & **Abdulla Alabbasi, A. M.** 2024 Examining the idea density and semantic distance of responses given by AI to tests of divergent thinking. *Journal of Creative Behavior* **59** (3), e1528. doi:10.1002/jocb.659.

Russell, S. J. & **Norvig, P.** 2003 *Artificial Intelligence: A Modern Approach*, 2nd Edn. Prentice Hall.

Sarica, S. & **Luo, J.** 2024 The innovation paradox: Concept space expansion with diminishing originality and the promise of creative artificial intelligence. *Design Science* **10**, e11; doi:10.1017/dsj.2024.11.

Sawyer, R. K. 2012 *The Science of Human Innovation: Explaining Creativity*. Oxford University Press.

Shaer, O., **Cooper, A.**, **Mokryn, O.**, **Kun, A. L.** & **Shoshan, H. B.** 2024. AI-augmented brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–17. Association for Computing Machinery. doi:10.1145/3544548.3580988.

Shi, Y., **Gao, T.**, **Jiao, X.** & **Cao, N.** 2023 Understanding design collaboration between designers and artificial intelligence: A systematic literature review. *Proceedings of the ACM on Human–Computer Interaction* **7** (CSCW2), 368; doi:10.1145/3610891.

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Sosa, R. 2019 Accretion theory of ideation: Evaluation regimes for ideation stages. *Design Science* **5**, e23; doi:10.1017/dsj.2019.23.

Sosa, R. & **Gero, J. S.** 2005 A computational study of creativity in design: The role of society. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **19** (4), 229–244; doi:10.1017/S0890060405050238.

Stein, M. I. 1987 Creativity research at the crossroads: A 1985 perspective. In *Frontiers of Creativity Research: Beyond the Basics* (ed. S. G. Isaksen), pp. 417–427. Bearly.

Stiefel, K. M. & **Coggan, J. S.** 2023 The energy challenges of artificial superintelligence. *Frontiers in Artificial Intelligence* **6**, 1240653; doi:10.3389/frai.2023.1172253.

Stöhr, B., **Koldewey, C.** & **Dumitrescu, R.** 2023 The role of design in interdisciplinary product development – challenges, research approaches and further research needs. *Proceedings of the Design Society* **3**, 3473–3482; doi:10.1017/pds.2023.347.

Surma-Aho, A., **Björklund, T.** & **Hölttä-Otto, K.** 2022 User and stakeholder perspective taking in novice design teams. *Design Science* **8**, e24; doi:10.1017/dsj.2022.24.

Thang, B., **Sluis-Thiescheffer, W.**, **Bekker, T.**, **Eggen, B.**, **Vermeeren, A.** & **Ridder, H. D.** 2008 Comparing the creativity of children's design solutions based on expert assessment. In *Proceedings of the 7th International Conference on Interaction Design and Children (IDC 2008)*, pp. 266–273. doi:10.1145/1463689.1463750.

Timmermans, J. A. & **Van der Rijst, R.** 2023 Interior design as an analogy for academic development. *International Journal for Academic Development* **28** (4), 379–384; doi:10.1080/1360144X.2023.2252467.

Torrance, E. P. 1974 *Torrance Tests of Creative Thinking: Norms-Technical Manual.* Personnel Press/Ginn.

Tørresen, J. 2021 Undertaking research with humans within artificial intelligence and robotics: Multimodal elderly care systems. *Technology | Architecture + Design* **5** (2), 141–145; doi:10.1080/24751448.2021.1945836.

Turing, A. M. 1950 Computing machinery and intelligence. *Mind* **59** (236), 433–460; doi:10.1093/mind/LIX.236.433.

Wang, S.-Y., **Su, W.-C.**, **Chen, S.**, **Tsai, C.-Y.**, **Misztal, M.**, **Cheng, K. M.**, **Lin, A.**, **Chen, Y.** & **Chen, M. Y.** 2024. RoomDreaming: Generative-AI approach to facilitating iterative, preliminary interior design exploration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery. doi:10.1145/3544548.3581262.

Wang, B., **Zhu, Y.**, **Chen, L.**, **Liu, J.**, **Sun, L.** & **Childs, P.** 2023 'A study of the evaluation metrics for generative images containing combinational creativity'. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **37**, e11; doi:10.1017/S0890060423000306.

Wei, X., **Weng, D.**, **Liu, Y.** & **Wang, Y.** 2015 Teaching based on augmented reality for a technical creative design course. *Computers & Education* **81**, 221–234; doi:10.1016/j.compedu.2014.10.017.

White, A. & **Smith, B. L.** 2001 Assessing advertising creativity using the Creative Product Semantic Scale. *Journal of Advertising Research* **41** (6), 27–34; doi:10.2501/JAR-41-6-27-34.

Yazici, B. & **Yolacan, S.** 2007 A comparison of various tests of normality. *Journal of Statistical Computation and Simulation* **77** (2), 175–183; doi:10.1080/10629360600678310.

Yuan, A., **Luther, K.**, **Krause, M.**, **Vennix, S. I.**, **Dow, S. P.** & **Hartmann, B.** 2016 Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported*

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Cooperative Work & Social Computing (CSCW 2016), pp. 1005–1017. Association for Computing Machinery. doi:10.1145/2818048.2820022.

Zhou, C., **Zhang, X.** & **Yu, C.** 2023 How does AI promote design iteration? The optimal time to integrate AI into the design process. *Journal of Engineering Design*, 1–28. doi:10.1080/09544828.2023.2254841.

Zhu, Q. & **Luo, J.** 2022 Generative pre-trained transformer for design concept generation: An exploration. *Proceedings of the Design Society* **2**, 1825–1834; doi:10.1017/pds.2022.185.

Appendix: The CPSS semantic pairs under each criterion used in this study

CPSS criterion	Semantic pair (7-point Likert scale)
Novelty	Over Used–Fresh Predictable–Novel Usual–Unusual Ordinary–Unique Conventional–Original Stale–Startling Customary–Surprising Commonplace–Astonishing Old Fashioned–Shocking Common–Astounding Warmed Over–Trendsetting Average–Revolutionary Old Hat–Radical Uninfluential–Influential Unprogressive–Pioneering
Resolution	Worthless–Valuable Unimportant–Important Insignificant–Significant Inessential–Essential Unnecessary–Necessary Illogical–Logical Senseless–Makes Sense Irrelevant–Relevant Inappropriate–Appropriate Inadequate–Adequate Ineffective–Effective Nonfunctional–Functional Inoperable–Operable Useless–Useful Unworkable–Workable
Style	Disordered–Ordered Disarranged–Arranged Disorganized–Organize Formless–Formed Incomplete–Complete Awkward–Graceful Repelling–Charming

Continued

https://doi.org/10.1017/dsj.2025.10029 Published online by Cambridge University Press

Continued

CPSS criterion	Semantic pair (7-point Likert scale)
	Coarse–Elegant
	Unattractive–Attractive
	Busy–Refined
	Straightforward–Intricate
	Simple–Complex
	Plain–Ornate
	Uncomplicated–Complicated
	Boring–Interesting
	Meaningless–Meaningful
	Mystifying–Understandable
	Unintelligible–Intelligible
	Ambiguous–Clear
	Unexplained–Self-Explanatory
	Bungling–Skillful
	Botched–Well Made
	Crude–Well Crafted
	Sloppy–Meticulous
	Careless–Careful