Proceedings of the Design Society, Volume 5: ICED25
https://doi.org/10.1017/pds.2025.10092

Bridging the gap in engineering creativity evaluations: exploring novice eye-gaze behavior across design modalities

DukHee Ka[1,✉], Sanaz Motamedi[1], Faez Ahmed[2], Farnaz Tehranchi[1] and Scarlett Miller[1]

[1] The Pennsylvania State University, USA, [2] Massachusetts Institute of Technology, USA

✉ dukhee.ka@psu.edu

ABSTRACT: The Consensual Assessment Technique (CAT) is one of the most effective and commonly used design evaluation methods. However, it fails to capture implicit cognitive processes and has mainly been studied in a homogenous design modality. To bridge this gap, the present study investigates the impact of design ideas represented in different modalities (i.e., text-only, sketch-only, text + sketch) on design evaluations for creativity, novelty, and usefulness, and examine human gaze patterns during the evaluation process. Our findings showed that novice raters exhibit higher interrater reliability and greater convergence in visual attention when rating ideas containing sketches compared to text-only design modality, highlighting the value of visual elements in design evaluations.

KEYWORDS: creativity, evaluation, decision making

1. Introduction

Over the last few decades, the advent of many breakthrough methodologies has transformed the way we evaluate creative ideas. These include, but are not limited to, the Consensual Assessment Technique (CAT) (Amabile, 1982), the Creative Product Semantic Scale (Besemer & O'Quin, 1986), and the Shah, Vargas-Hernandez, and Smith methods (Shah et al., 2003). Among these, the CAT is considered to be one of the most effective and commonly used methods for evaluating creative ideas (Anderson et al., 2023; Edwards et al., 2022; Hazeri et al., 2017; Miller et al., 2020). This method operates under the hypothesis that a design idea is creative to the extent that experts in the field independently agree with each other (Amabile, 1982). In practice, a panel of independent experts evaluates design ideas on a Likert scale without having to justify their ratings. Typically, these evaluations include creativity, usefulness, and novelty (Amabile, 2018; Besemer, 1998; Besemer & O'Quin, 1986; Diedrich et al., 2015) to distinguish different facets of creativity.

While the CAT is widely considered a gold standard for evaluating design ideas (Carson, 2006), it is not without its challenges. First, existing design evaluation studies using the CAT have largely focused on ideas represented in homogeneous design modalities, such as sketches (Amabile, 1982; Anderson et al., 2023), music (Hickey, 2001), poems (Baer et al., 2004), and storytelling (Baer et al., 2004). However, engineering design ideas are often complex and multifaceted, ranging from early-phase sketches with text notes to physical prototypes with movable parts or CAD models (Goel et al., 2008; Pahl & Beitz, 2013). These varying levels of detailed design modalities may lead to different levels of comprehension during design evaluation, depending on the amount of information they can transmit (Berni et al., 2020). For instance, text descriptions offer information in sequence through annotation or documentation, whereas sketches provide geometric and topological information (Tversky et al., 2003; Ullman et al., 1990). On the other hand, physical prototypes and CAD models provide concrete and detailed representations, but lack the usability or agility compared to sketches (Alvarado & Davis, 2001; Yang, 2009). Understanding how different design modalities affect idea

evaluations is essential for understanding the CAT method and ensuring more reliable assessments during the design phase.

The second challenge is that the CAT is often criticized as being expensive and time-consuming due to the need to recruit as many experts suitable for a specific domain (Kaufman, 2014; Kaufman et al., 2008). To address this issue, researchers have employed breakthrough AI algorithms to automatically compute ratings from large sets of design ideas, such as Support Vector Machine (Toh et al., 2017), CNN (Patterson et al., 2024), GAN (Elgammal, 2017), and attention-enhanced model (Song et al., 2023). However, these approaches have often focused on ideas represented in homogeneous design modalities and have shown limited predictive accuracy compared to experts, suggesting that these ideas often lack enough evidence for machines alone to make accurate predictions. These challenges stem from our limited understanding of the factors that drive decision-making in design evaluations by humans. Because the CAT relies solely on ratings from humans, it fails to capture the implicit cognitive processes and attention patterns that underlie design evaluations (Said-Metwaly et al., 2017). This is where eye-tracking data becomes invaluable.

Tracking eye movements during design evaluation may be a useful tool for shedding light on raters unconscious behavior that is not articulated by humans but may be crucial for explaining their decisions (Borgianni & Maccioni, 2020; Matthiesen et al., 2013; Tehranchi et al., 2020). Specifically, eye-tracking can help us identify which parts of a design attract attention and how visual attention shifts when evaluating the same ideas in different design modalities such as text or sketch (Berni et al., 2020; Boa et al., 2015; Carbon et al., 2006; Dong & Liu, 2018; Du & MacDonald, 2014). Moreover, by integrating eye-tracking insights with the CAT method, researchers can better understand how raters process design ideas across different design modalities, accounting for both explicit ratings and implicit cognitive processes. This type of insight may be useful in helping train AI to better adapt to accurate assessments of ideas in various design modalities.

Given that previous CAT-based studies have focused on homogenous design modalities without exploring human visual attention, this paper aims to investigate the impact of various design modalities on design evaluation and to understand human eye-gaze behavior during the evaluation process. The results of this paper provide insights into the underlying factors of human decision-making across diverse design modalities revealed through visual attention. This was achieved through an empirical study in which novice raters evaluated engineering concepts represented in various design modalities—text-only, sketch-only, and text + sketch—while monitoring their eye movements.

2. Research questions

The goal of this study was to investigate the impact of ideas represented in different design modalities on design ratings and to understand human gaze patterns during the evaluation process. Specifically, an empirical study was developed to address the following research questions (RQs):

- How does the interrater reliability of ratings vary across design modalities? The first research question (RQ1) was developed to understand whether human ratings exhibit different levels of agreement across various design modalities of the same design concepts. We hypothesized that the design modalities containing more information would result in higher interrater reliability of ratings (e.g., text-only < sketch-only < text + sketch). Ratings for text-only design modality would show lower interrater reliability than the sketch-only modality due to the lack of visual cues such as geometric or spatial information (Tversky et al., 2003; Ullman et al., 1990). Ratings for text + sketch design modality would provide the highest interrater reliability, as this modality integrates both descriptive and visual information, enhancing the clarity of design concepts and enabling consistent understanding among raters (Dong & Liu, 2018; Macomber & Yang, 2011).
- How closely do raters' longest fixation points converge across design modalities? The second research question (RQ2) was constructed to investigate how closely multiple raters' longest fixation points were located and whether this proximity varied by design modality. We hypothesized that raters' longest fixation points would be more closely located when evaluating ideas represented in text + sketch design modality compared to sketch-only modality, followed by text-only modality. This hypothesis was grounded in previous research indicating that longer fixation durations reflect cognitive effort, such as retrieving meaning and constructing mental models (Just & Carpenter, 1980). Therefore, we expected raters' longest fixation points to

converge more in design modalities with richer information, as increased clarity enhances the interpretation (Dong & Liu, 2018).

3. Methodology

To answer these research questions, we conducted an empirical study with novice human raters. The data used in this study were from a larger public dataset of conceptual design ideas for milk frothers (Starkey et al., 2016). Starkey et al. (2016) created this dataset by recruiting undergraduate students in an introductory engineering design course and assigning them the task of developing concepts for an innovative product that froths milk in a short amount of time. Each design idea includes a sketch and a short text description. This dataset has been used in CAT-based research in a variety of settings (Ahmed et al., 2018; Edwards et al., 2022; Song et al., 2023; Toh et al., 2017).

For this study, we used ten conceptual design ideas for milk frothers from previous research by Ahmed et al. (2018) which were randomly selected from a larger milk frother dataset. Each of these ten ideas was then reconstructed into three design modalities: text-only, sketch-only, and text + sketch (Figure 1). The data containing both text and sketch corresponds to the original data while text-only data were created by cropping the text description area from the original dataset, and sketch-only data were created by cropping the sketch area.

Figure 1. A sample milk frother idea represented in three design modalities: text + sketch (left), text-only (middle), and sketch-only (right)

3.1. Participants

This empirical study involved 13 novice raters (8 males and 5 females), who were all graduate engineering students at a large public university. Novices were defined as individuals without experience or expertise in design and creativity assessment (Kaufman et al., 2008; Kaufman et al., 2013), as experts are typically defined as those with domain-specific expertise to make accurate evaluations of creative products (Amabile, 1982).

3.2. Procedure

At the start of the study, a brief description of the background and the purpose of the study was provided, and consent was obtained in accordance with Institutional Review Board (IRB) procedures. All the experiments were conducted in a quiet room with only a novice rater and a researcher present. A novice rater was randomly assigned to one of three rater groups (i.e., A, B, or C), with each group evaluating a total of ten ideas across all three design modalities without overlap (Table 1). For instance, a novice rater in rater group A evaluated three ideas represented in text-only (i.e., idea numbers 5, 8, and 9), three ideas in sketch-only (i.e., idea numbers 1, 2, and 4), and four ideas in text + sketch (i.e., idea numbers 3, 6, 7, and 10). This approach ensured that no rater evaluated the same idea across different design modalities, thereby avoiding reliance on memory during the evaluation process.

Once assigned to a rater group, the novice rater sat at a desk equipped with a 27-inch LCD monitor with a 2560x1440-pixel resolution. A Tobii Pro Fusion screen-based eye tracker was placed beneath the monitor, and the novice rater sat approximately 50 to 80 cm away from the screen. The novice rater completed a 9-point calibration of the eye tracker through Tobii Pro Lab software, which was used to manage the recording and obtain fixation points. To prevent central fixation bias (Tatler, 2007) and minimize head movements, rating sheet was displayed on the left half of the screen, while design idea was on the right half (Figure 2). The rating sheet was created in Excel, where novice raters could use a combo box to input ratings. The combo box required only mouse clicks to avoid looking down at the

keyboard, as this could affect the accuracy of the eye tracking data. Design ideas were presented and randomized using Qualtrics. Each idea was accompanied by a definition of the rating category (i.e., creativity, usefulness, and novelty) to ensure consistent evaluation.

Table 1. Idea configuration for three rater groups (i.e., A, B, or C) and three design modalities without overlapping

Idea number	Design modalities		
	Text-only	**Sketch-only**	**Text + Sketch**
5, 8, 9	A	C	B
1, 2, 4	B	A	C
3, 6, 7, 10	C	B	A

Prior to evaluating the designs of interest in the current investigation, novice raters were provided with written instructions outlining the study procedures and definitions of the three rating categories (i.e., creativity, novelty, and usefulness). For creativity, raters were asked to use their own subjective definitions, in line with the hypothesis of CAT. Usefulness referred to the feasibility and effectiveness of the idea. Novelty referred to ideas only a few people will come up with. Ratings were provided on a 6-point Likert scale, with 1 representing low level and 6 representing high level of each rating category. A trial session was conducted with one sample milk frother idea—which was not included in the ten ideas of interest—to familiarize the raters with the evaluation process, rating sheet, and Qualtrics interface. All raters received the same sample idea rating. After completing the trial session, the study proceeded to the design evaluation phase.

During the design evaluation, each novice rater completed a total of 30 evaluations across three design modalities and three rating categories. Specifically, the rater evaluated ten ideas across the three design modalities, with each idea evaluated three times for its creativity, usefulness, and novelty. The sequence of presenting design modalities and ideas was randomized and balanced.

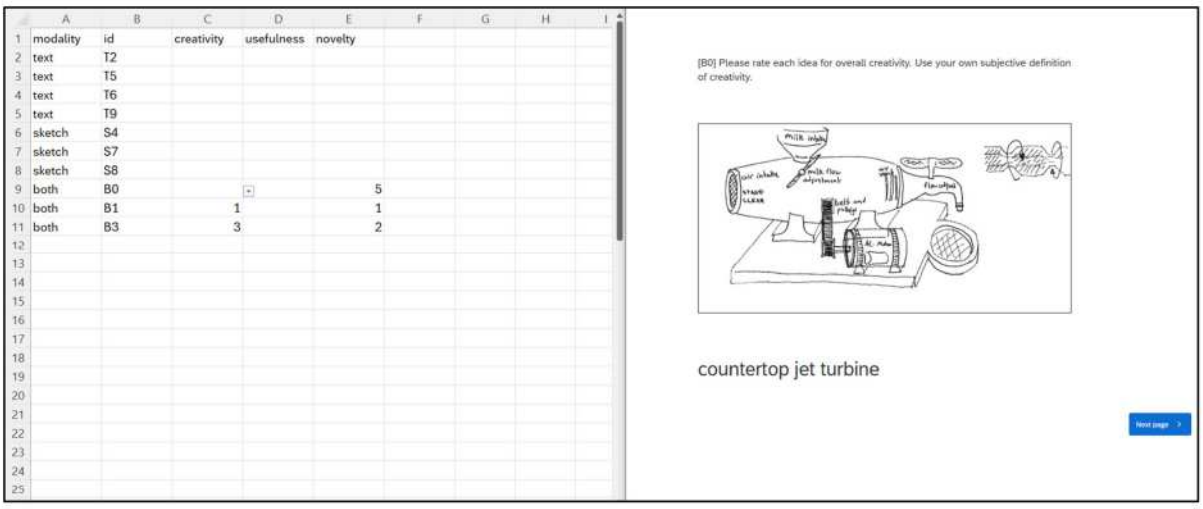

Figure 2. Design evaluation demonstration: providing creativity rating for an idea with text + sketch design modality

3.3. Metric

We measured normalized mean pairwise distance (NMPD) between raters' longest fixation points to assess convergence in visual attention. Specifically, we first identified each rater's longest fixation points within each idea—across design modalities and rating categories—based on XY-coordinates. We then calculated the mean pairwise Euclidean distance between these fixation points among raters within the same rater group. To account for different idea sizes, we normalized these mean pairwise Euclidean distances based on the area of each idea and its respective design modality.

4. Data analysis and results

Prior to the analysis, the eye tracking calibration was verified. One novice rater was excluded from the analysis due to poor calibration accuracy (23.40 mm), which was substantially higher than overall novice raters (M = 5.67 mm, SD = 5.51 mm). As such, a total of 12 novice raters were included in the analysis, with four novice raters in each rater group (i.e., A, B, and C). The remainder of this section presents our results in relation to our research questions.

4.1. RQ1: How does the interrater reliability of ratings vary across design modalities?

RQ1 was constructed to determine whether human ratings exhibit different levels of agreement across different design modalities. Our hypothesis was that design modalities containing more information would result in higher interrater reliability in ratings (e.g., text-only < sketch-only < text + sketch). In order to test this hypothesis, we computed intraclass correlation coefficients (ICC2) of ratings for each design modality (i.e., text-only, sketch-only, and text + sketch), each rating category (i.e., creativity, usefulness, and novelty), and each rater group (i.e., A, B, and C). It is important to note that ICC2 was calculated separately for each rater group since novice raters were assigned to one of the three rater groups and evaluated ideas across three design modalities without overlapping (Table 1) to prevent reliance on memory during the evaluation process. The value of ICC2 ranges from 0 to 1, where 0 means no interrater reliability and 1 means perfect interrater reliability. Typically, an ICC2 value greater than .7 indicates great interrater reliability, a value between .5 and .7 indicates moderate reliability, and a value below .5 indicates poor reliability (Koo & Li, 2016). Table 2 showcases our results.

For Creativity, our results showed that novice raters had higher interrater reliability when evaluating ideas in sketch-only design modality (ICC2, 0.17 – 0.69, M = 0.46), with moderate reliability for two of the three rater groups. Both text + sketch and text-only design modalities had poor reliability. For usefulness, text + sketch (ICC2, 0.17 – 0.45, M = 0.34) had the highest agreement, though all design modalities were considered poor reliability. Finally, for novelty, the sketch (ICC2, 0.12 – 0.81, M = 0.52) and sketch + text (ICC2, 0.18 – 0.90, M = 0.50) design modalities had moderate reliability, while text-only had poor reliability. Overall, novice raters in rater group C exhibited poor reliability in all design modalities.

Table 2. Interrater reliability (ICC2) of ratings for each design modality, rating category, and rater group

Design modalities	Rater groups	Rating categories		
		Creativity	Usefulness	Novelty
Text-only	A	0.00	0.21	0.25
	B	0.21	0.00	0.25
	C	0.00	0.31	0.00
	Average	*0.07*	*0.17*	*0.17*
Sketch-only	A	0.69*	0.10	0.81*
	B	0.52*	0.38*	0.64*
	C	0.17	0.00	0.12
	Average	*0.46*	*0.16*	*0.52*
Text + Sketch	A	0.67*	0.42	0.43
	B	0.16	0.17	0.90*
	C	0.00	0.45*	0.18
	Average	*0.28*	*0.34*	*0.50*

*Note. All were statistically significant at $p < .05$.

4.2. RQ2: How closely do raters' longest fixation points converge across design modalities?

RQ2 aimed to investigate the proximity of multiple novice raters' longest fixation points across design modalities. We hypothesized that raters' longest fixation points would be more closely located when

evaluating ideas in text + sketch design modality compared to sketch-only modality, followed by text-only modality. To address RQ2, we first computed the NMPD of novice raters' longest fixation points for each idea across design modalities and rating categories. We then averaged NMPDs across all ideas within the same design modality, rating category, and rater group. A higher NMPD indicates more dispersion in visual attention, meaning novice raters focused on different areas of ideas, whereas a lower NMPD indicates greater convergence, meaning novice raters focused on similar areas. Table 3 shows our results.

For creativity, our result showed that novice raters' longest fixation points converged the most when evaluating ideas in text + sketch design modality (NMPD, 22.28 – 38.46, M = 28.95), compared to sketch-only modality (NMPD, 40.06 – 79.43, M = 54.14), followed by text-only modality (NMPD = 67.08 – 106.13, M = 82.93). For usefulness, text + sketch design modality (NMPD, 29.90 – 49.15, M = 37.83) had the closest convergence, followed by sketch-only modality (NMPD, 38.88 – 74.78, M = 53.08) modality and text-only modality (NMPD, 55.79 – 94.48, M = 73.81). For novelty, text + sketch design modality (NMPD, 22.28 – 46.30, M = 34.95) demonstrated the best convergence, followed by sketch-only modality (NMPD, 48.55 – 68.33, M = 58.34) and text-only modality (NMPD, 58.41 – 104.66, M = 73.91). Moreover, 94% of the longest fixation points—from all eye-tracking data for the text + sketch design modality—were located in the sketch area rather than the text description area.

Table 3. The average NMPD for each design modality, rating category, and rater group. A higher NMPD indicates greater dispersion in visual attention, while a lower NMPD means greater convergence

Design modalities	Rater groups	Rating categories		
		Creativity	Usefulness	Novelty
Text-only	A	106.13	94.48	104.66
	B	67.08	55.79	58.67
	C	75.58	71.15	58.41
	Average	*82.93*	*73.81*	*73.91*
Sketch-only	A	42.93	38.88	58.15
	B	79.43	74.78	48.55
	C	40.06	45.59	68.33
	Average	*54.14*	*53.08*	*58.34*
Text + Sketch	A	38.46	49.15	46.30
	B	22.28	34.43	22.28
	C	26.10	29.90	36.26
	Average	*28.95*	*37.83*	*34.95*

Figure 3 shows the longest fixation points of novice raters overlaid on a sample milk frother idea evaluated for novelty across three design modalities. Each solid X mark, labeled with R#, represents the longest fixation point of an individual novice rater (R1 – R12). For text-only design modality, all four longest fixation points from rater group A were dispersed (NMPD = 104.66). For sketch-only design modality, two of the four longest fixation points from rater group C were closely located (NMPD = 68.33). For text + sketch design modality, all four longest fixation points from rater group B were clustered within a similar area (NMPD = 22.28).

5. Discussion

We designed this study to provide evidence on how ideas represented in different design modalities impact design evaluation and to understand whether the convergence of novice raters' longest fixation points varies across design modalities. The analysis of data gathered from novice raters revealed the following findings:

5.1. The impact of design modality of ideas in design evaluation

The first finding was that novice raters demonstrated higher interrater reliability when rating ideas represented in sketch-only and text + sketch design modalities compared to text-only modality. This outcome partially supported the hypothesis, as there was insufficient evidence to differentiate interrater reliability when evaluating ideas in sketch-only and text + sketch design modalities. Additionally, the

ICC2 values for ideas represented in text-only design modality were consistently poor. Since text-only ideas contain less information and offer limited clarity about the concept of a milk frother, the ratings of ideas might have diverged across individuals (Dong & Liu, 2018).
Moreover, all ICC2 values for usefulness were below 0.5 or statistically insignificant. The ideas in this study were based on the task of developing concepts for an innovative milk frother, with a primary focus on innovativeness rather than feasibility. Since these ideas differ from the commercially available milk frothers, novice raters might be confused about the feasibility and effectiveness of the ideas (Macomber & Yang, 2011). Overall, these results demonstrate that the inclusion of visual elements like sketches enhances interrater reliability in design evaluation across different rating categories. Specifically, sketch-only and text + sketch design modalities led to a higher agreement among novice raters, particularly in the assessment of creativity and novelty, compared to text-only design modality. This underscores the importance of multimodal approaches in capturing more consistent and reliable evaluations.

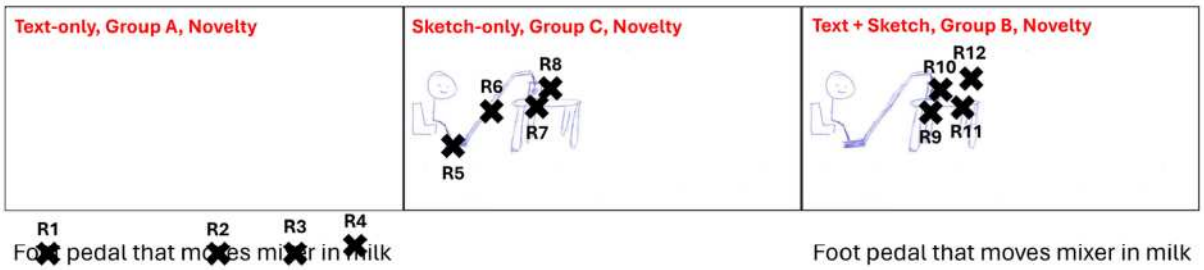

Figure 3. Novice raters' longest fixation points overlaid on a sample milk frother idea evaluated for novelty across design modalities: text-only (left; evaluated by rater group A), sketch-only (middle; evaluated by rater group C), text + sketch (right; evaluated by rater group B). Each solid X mark, labeled with R#, indicates the longest fixation point from an individual novice rater (R1 – R12)

5.2. The contribution of eye-tracking data to understanding the convergence of visual attention

The second finding was that novice raters provided better convergence in their longest fixation points when evaluating ideas represented in text + sketch design modality compared to sketch-only modality, followed by text-only modality. This finding aligned with our hypothesis described in RQ2. A comparison of NMPD values between text-only and sketch-only design modalities suggests that images helped direct attention to specific design components more effectively than solely text descriptions. Considering the increase in ICC2 with sketches compared to the text-only design modality in RQ1, it can be inferred that raters tended to fixate more similarly in closer proximity when evaluating ideas with higher interrater reliability. Additionally, better convergence of visual attention in text + sketch design modality compared to sketch-only design modality indicates that additional text description helped novice raters focus on key aspects of the design, reducing variability in fixation points. This highlights the effectiveness of multimodal representation (text + sketch) in guiding attention and improving visual agreement among raters.
As shown in Table 3, novice raters in rater group C showed a similar level of convergence in their visual attention compared to other rater groups, whereas they showed poor interrater reliability in overall design modalities (Table 2). This contradiction suggests that visual attention does not necessarily translate to agreement in evaluations, implying that raters may focus on similar areas but interpret them differently when rating ideas. Further research may be needed on this topic. Moreover, the variation in rating agreement across rater groups highlights the strong influence of individual differences and the need for a larger number of novice raters.
As expected, novice raters tended to focus on sketches rather than text descriptions in the text + sketch design modality. Once they evaluated an idea for one of the rating categories—such as creativity—some did not even look at the text description when evaluating the same idea for usefulness and novelty. That is, when the two design modalities are presented simultaneously, according to the split-attention effect (Sweller, 2011), raters may select the design modality that offers easier understanding or clearly outlines the idea to reduce cognitive load. However, this does not imply that novice raters neglect the text description in text + sketch design modality. In fact, the additional text description contributed to an increase in ICC2 values for certain rating categories and greater convergence of raters' visual attention.

This suggests that text description seems to serve as a secondary source of information, whereas the sketch plays a more dominant role. However, when it comes to using multimodal representations for AI models, combining various design modalities may create uneven impacts on design evaluation, and these combinations might also pose challenges for machines to interpret the content accurately.

6. Limitations and future work

Despite the valuable findings of this study, there are several limitations. First, we did not use areas of interest (AOIs) to investigate novice raters' visual patterns. Defining AOIs allows researchers to use several commonly used eye-tracking metrics, such as fixation duration on an AOI, time to first gaze on an AOI, the number of revisits to an AOI, etc. However, defining AOI in this study was challenging due to overlapping design components in a two-dimensional environment, unclear borders between design components, and inconsistencies in the number of features across design modalities of the same concept. For future work, it would be beneficial to consider component structures and borders when choosing visual stimuli. Moreover, for ideas with fewer design features, gaze distribution might be sparse, and gaze duration could be lower. To address this, future studies should consider evaluating the complexity of ideas beforehand, allowing for a more balanced selection of ideas to better capture raters' visual attention. These future approaches are relevant to the CAT approach, as they evaluate creativity within the specific pool of ideas.

Second, the NMPD did not take into account the overall size of design modalities. For example, ideas represented in text-only design modality tended to be longer and aligned on a horizontal plane, whereas the sketches were more dispersed. Therefore, the current findings should be interpreted with caution.

Third, we included a small number of ideas and novice raters. The present results should therefore be interpreted with caution, as ICC2 values typically require at least 30 data points as a rule of thumb (Koo & Li, 2016). Additionally, expert raters may have a better understanding of design evaluation and related areas. Future work should include expert raters to compare their interrater reliability with novice raters across various design modalities and relative visual patterns. By analyzing visual attention from these raters, we hope to gain insights into the cognitive factors that influence the validity of their ratings.

Lastly, this study was conducted in a two-dimensional environment with sketches and text descriptions. Throughout the design process, design ideas are often represented in various design modalities such as sketches with notes, CAD models, and physical prototypes with movable parts. In the case of physical prototypes, human raters rely on various senses beyond vision, such as touch and hearing. Therefore, incorporating physical prototypes would require surveys or interviews, as well as various biometric devices such as a glass-type eye tracker and a galvanic skin response sensor.

7. Conclusion

This study was designed to provide insight into how various design modalities of an idea affect design evaluation and to understand the proximity of visual attention during the evaluation process. To achieve this, we conducted a design evaluation study in which novice raters evaluated ideas represented in various design modalities—text-only, sketch-only, and text + sketch—while their eye movements were monitored. As a result, we found that novice raters exhibited higher interrater reliability when rating ideas containing sketches compared to text-only ideas. Additionally, novice raters showed greater convergence in their longest fixation points for ideas represented in text + sketch design modality, compared to sketch-only modality, followed by text-only modality. These findings highlight the value of eye-tracking data in revealing unconscious human behavior during design evaluation. Future studies are needed for further investigation to better understand how novices differentiate rating categories.

Acknowledgements

This work was supported by the National Science Foundation under Grant No. CMMI-2231261.

References

Ahmed, F., Fuge, M., Hunter, S., & Miller, S. (2018). Unpacking subjective creativity ratings: Using embeddings to explain and measure idea novelty. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*,

Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., & Miller, S. (2019). Interpreting idea maps: Pairwise comparisons reveal what makes ideas novel. *Journal of Mechanical Design*, 141(2), 021102.

Alvarado, C., & Davis, R. (2001). Preserving the freedom of paper in a computer-based sketch tool. *Proceedings of HCI international*,

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, 43(5), 997.

Amabile, T. M. (2018). Creativity in context: Update to the social psychology of creativity. *Routledge*.

Anderson, R. C., Beghetto, R. A., Glăveanu, V., & Basu, M. (2023). Is curiosity killed by the CAT? A divergent, open-ended, and generative (DOG) approach to creativity assessment. *Creativity Research Journal*, 35 (3), 380–395.

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity research journal*, 16 (1), 113–117.

Berni, A., Maccioni, L., & Borgianni, Y. (2020). Observing pictures and videos of creative products: An eye tracking study. *Applied Sciences*, 10(4), 1480.

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.

Besemer, S. P., & O'Quin, K. (1986). Analyzing creative products: Refinement and test of a judging instrument. *The Journal of Creative Behavior*.

Boa, D. R., Ranscombe, C., & Hicks, B. (2015). Determining the similarity of products using pairwise comparisons and eye tracking. *DS 80-5 Proceedings of the 20th International Conference on Engineering Design (ICED 15) Vol 5: Design Methods and Tools-Part 1*, Milan, Italy, 27-30.07. 15,

Borgianni, Y., & Maccioni, L. (2020). Review of the use of neurophysiological and biometric measures in experimental design research. *AI EDAM*, 34 (2), 248–285.

Carbon, C., Hutzler, F., & Minge, M. (2006). Innovativeness in design investigated by eye movements and pupillometry. *Psychology Science*, 48(2), 173.

Carson, S. (2006). Creativity and mental illness. In *Invitational panel discussion hosted by Mind Matters Consortium*. New Haven, CT: Yale University.

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of aesthetics, creativity, and the arts*, 9(1), 35.

Dong, Y., & Liu, W. (2018). Research on UX evaluation method of design concept under multi-modal experience scenario in the earlier design stages. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 12, 505–515.

Du, P., & MacDonald, E. F. (2014). Eye-tracking data predict importance of product features and saliency of size change. *Journal of Mechanical Design*, 136(8), 081005.

Edwards, K. M., Peng, A., Miller, S. R., & Ahmed, F. (2022). If a picture is worth 1000 words, is a word worth 1000 features for design metric estimation? *Journal of Mechanical Design*, 144(4), 041402.

Elgammal, A. (2017). Can: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 6, 2017.

Goel, A. K., Davis, R., & Gero, J. S. (2008). Multimodal design: An overview. *AI EDAM*, 22 (2), 83–84.

Hazeri, K., Childs, P., & Cropley, D. (2017). Proposing a new product creativity assessment tool and a novel methodology to investigate the effects of different types of product functionality on the underlying structure of factor analysis. *DS 87-8 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 8: Human Behaviour in Design*, Vancouver, Canada, 21-25.08. 2017,

Hickey, M. (2001). An application of Amabile's consensual assessment technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education*, 49 (3), 234–244.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4), 329.

Kaufman, J. C. (2014). Creativity and mental illness. Cambridge University Press.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton * , J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, 20 (2), 171–178.

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts*, 7(4), 332.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15 (2), 155–163.

Macomber, B., & Yang, M. (2011). The role of sketch finish and style in user responses to early stage design concepts. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*,

Matthiesen, S., Meboldt, M., Ruckpaul, A., & Mussgnug, M. (2013). Eye tracking, a method for engineering design research on engineers' behavior while analyzing technical systems. *DS 75-7: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies*, Vol. 7: Human Behaviour in Design, Seoul, Korea, 19-22.08. 2013,

Miller, S. R., Hunter, S. T., Starkey, E., Ramachandran, S., Ahmed, F., & Fuge, M. (2020). How Should We Measure Creativity in Design Studies? A Comparison of Social Science and Engineering Approaches. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*,

Pahl, G., & Beitz, W. (2013). Engineering design: a systematic approach. Springer Science & Business Media.

Patterson, J. D., Barbot, B., Lloyd-Cox, J., & Beaty, R. E. (2024). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods*, 56 (4), 3619–3636.

Said-Metwaly, S., Van den Noortgate, W., & Kyndt, E. (2017). Approaches to measuring creativity: A systematic literature review. *Creativity. Theories–Research-Applications*, 4 (2), 238–275.

Shah, J. J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies*, 24 (2), 111–134.

Song, B., Miller, S., & Ahmed, F. (2023). Attention-enhanced multimodal learning for conceptual design evaluations. *Journal of Mechanical Design*, 145(4), 041410.

Starkey, E., Toh, C. A., & Miller, S. R. (2016). Abandoning creativity: The evolution of creative ideas in engineering design course projects. *Design Studies*, 47, 47–72.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, 7 (14), 4–4.

Tehranchi, F., Ritter, F. E., & Chae, C. (2020). Visual Attention during E-Learning: Eye-tracking Shows that Making Salient Areas More Prominent Helps Learning in Online Tutors. *CogSci*,

Toh, C. A., Starkey, E. M., Tucker, C. S., & Miller, S. R. (2017). Mining for creativity: Determining the creativity of ideas through data mining techniques. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*,

Tversky, B., Suwa, M., Agrawala, M., Heiser, J., Stolte, C., Hanrahan, P., ... Lee, P. (2003). Sketches for design and design of sketches. *Human Behaviour in Design: Individuals, Teams, Tools*, 79–86.

Ullman, D. G., Wood, S., & Craig, D. (1990). The importance of drawing in the mechanical design process. *Computers & graphics*, 14 (2), 263–274.

Yang, M. C. (2009). Observations on concept generation and sketching in engineering design. *Research in Engineering Design*, 20, 1–11.