Check for updates

Problem Recognition and Intervention Strategies toward AI-Generated Design Outputs between Designers and Non-Designers: AI-Generated Electric Screwdriver Design

Yoonkyeong Lee[1], Sunchul Kim[2*]

[1]Department of Industrial Design, Ph.D. Student, Kyung Hee University, Yongin, Korea

[2]Department of Industrial Design, Professor, Kyung Hee University, Yongin, Korea

Abstract

Background As generative artificial intelligence (AI) enables non-experts to create visually refined design outputs, design expertise is increasingly related to interpretation and evaluation. This study compares how designers and non-designers recognize problems in the same AI-generated product design output.

Methods This study combined in-depth interviews with 12 participants and a survey of 73 valid respondents. All participants evaluated the same AI-generated electric screwdriver design, and the data were analyzed through open coding, descriptive statistics, t-tests, and chi-square analyses.

Results Designers tended to view the AI-generated output as an incomplete design requiring review of structure, usability, materials, and manufacturability, whereas non-designers evaluated the output mainly through visual impression and perceived completeness. Significant group differences were found in problem recognition, judgment criteria, and intervention depth, suggesting that the same rating scale may carry different meanings between the two groups.

Conclusions This study shows that, in AI-integrated design environments, designers contribute by identifying design problems, interpreting outputs in relation to use and production contexts, and proposing appropriate improvement directions.

Keywords Generative AI, Design Expertise, Human–AI Collaboration, Paradox of Scoring, False Premium Perception

*Corresponding author: Sunchul Kim (sunchulkim@khu.ac.kr)

Citation: Lee, Y., & Kim, S. (2026). Problem Recognition and Intervention Strategies toward AI-Generated Design Outputs between Designers and Non-Designers: AI-Generated Electric Screwdriver Design. *Archives of Design Research, 39*(4), 95-109.

http://dx.doi.org/10.15187/adr.2026.11.39.4.95

Received : Jun. 01. 2026 ; **Reviewed :** Jul. 23. 2026 ; **Accepted :** Aug. 07. 2026
pISSN 1226-8046 **eISSN** 2288-2987

Copyright : This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits unrestricted educational and non-commercial use, provided the original work is properly cited.

1. Background and Purpose

The proliferation of generative AI has created an environment in which even non-experts can produce visually sophisticated design outputs. However, producing visually refined outputs does not necessarily ensure their appropriateness within functional, structural, and use contexts. Design scholarship has long regarded judgment as a central component of professional design competence—the capacity to discern what is significant, appropriate, and worth pursuing in complex and indeterminate situations (Nelson & Stolterman, 2012). As AI increasingly participates in creative problem-solving, this capacity takes on renewed importance. Prior research suggests that human design increasingly involves sensemaking—understanding which problems should or could be addressed (Verganti et al., 2020)—and that designers continue to exercise agency in directing AI-assisted creative processes (Guo et al., 2023). Building on these perspectives, the present study examines how design expertise is manifested in the interpretation and evaluation of AI-generated outputs and in the formulation of directions for their improvement. Although related research is growing (Anantrasirichai & Bull, 2022), empirical studies that directly compare how designers and non-designers differ in problem recognition and intervention strategies when presented with identical AI-generated outputs remain limited. Accordingly, this study compares how designers and non-designers differ in recognizing problems in identical AI-generated design outputs and in proposing directions for their improvement. Examining both dimensions enables the study to identify not only whether the two groups perceive different problems, but also how they translate those perceptions into different directions for improvement. In this study, intervention refers not to directly redesigning the output, but to identifying elements requiring modification and suggesting how they could be improved. Non-designers were included not as substitutes for professional designers, but as users who may evaluate AI-generated outputs and suggest revisions within increasingly accessible generative design environments. An electric screwdriver was selected because it is a familiar consumer product whose basic use can be readily understood by non-designers, while also involving professional design considerations such as grip, operability, weight balance, structural configuration, material, and manufacturability.

2. Theoretical Background

2. 1. Generative AI and the Transformation of Design Practice

Earlier research on AI and machine learning provided a conceptual foundation for understanding changes in design practice. However, these studies were grounded in broader conceptions of AI that preceded the widespread public accessibility of generative AI systems in late 2022. Dove et al. (2017) conceptualized machine learning as a new design material, while Verganti et al. (2020) discussed how AI-enabled problem-solving was shifting the role of design toward sense-making. More recent studies have focused specifically on the use of generative AI and text-to-image technologies in design ideation, visual exploration, and

human–AI collaboration. For example, Paananen et al. (2024) showed that text-to-image technologies can support creative divergence and visual exploration during early-stage design ideation.

Recent studies have further examined generative AI in relation to design process models, product ideation, and human–AI collaboration. Na (2024) proposed a generative AI-based design model focusing on applications in design processes and design management, while Kwon et al. (2024) introduced a designer–generative AI ideation process for generating images aligned with designer intent during early-stage product design exploration. Chung and Choi (2025) discussed directions for improving image-generation AI in product ideation, emphasizing the continuing need for designer-led interpretation and refinement. Guo et al. (2023) highlighted the importance of designer agency in AI-augmented co-creation, while Palani and Ramos (2024) documented how creative practitioners' roles and workflows evolve as they interact with and orchestrate generative AI tools. Wadinambiarachchi et al. (2024) further showed that exposure to AI-generated images can increase design fixation and reduce the number, variety, and originality of generated ideas. Similarly, Lee and Kim (2025) found that AI tools are particularly effective in the discovery and definition stages, whereas human designers remain essential in the development and validation stages, where contextual interpretation and design judgment are required.

Collectively, these studies suggest that generative AI is reshaping both the sequence of design activities and the division of roles between designers and AI by introducing iterative cycles of generation, evaluation, and refinement. While AI supports idea generation and visualization, designers increasingly guide the process by interpreting outputs, assessing their appropriateness, and directing subsequent improvement. Accordingly, the critical focus of design expertise is extending beyond the direct production of design outputs toward the evaluation and redirection of AI-generated outcomes.

Although prior research has examined differences between experts and novices in design-related perception and judgment, how such differences emerge in the evaluation of generative AI-produced product designs has not been systematically investigated. Empirical studies directly comparing how designers and non-designers recognize problems and propose directions for improvement in response to the same AI-generated design output therefore remain limited. This study addresses this gap by examining group-based differences in problem recognition and intervention strategies.

2. 2. Design Judgment Differences from the Perspective of Human Augmentation

Augmented Humanity (AH) has been defined as a form of human–computer integration that extends human physical, intellectual, and social capabilities beyond their ordinary ranges (Guerrero et al., 2022). From this perspective, technology is understood as augmenting human capability rather than simply substituting for human agency. Building on the discussion of sensemaking and designer agency in the preceding section, generative AI can expand the capacity for visual generation and alternative exploration, while designers retain responsibility for framing problems, interpreting outputs, assessing their contextual value, and determining subsequent design directions (Guo et al., 2023; Verganti et al., 2020). This interpretation is consistent with Shneiderman's (2020) human-centered AI framework, which advocates reliable, safe, and trustworthy systems that combine advanced automation with

meaningful human control. Recent studies on human–AI collaboration further support this perspective. Song et al. (2024) proposed a human–AI collaboration framework that identifies AI capabilities, interaction-related attributes, and trust enablers as key conditions for effective collaboration. Wang et al. (2025) further demonstrated that the effects of human–AI co-creation may vary according to design experience, indicating that domain expertise remains an important factor in AI-supported creative processes.

Research on expert–novice differences in design judgment provides further grounds for examining how expertise shapes the evaluation of design outputs. Cross (2004) identified expert designers' ability to structure and develop problem and solution spaces as a central characteristic of design expertise. Hu et al. (2023) further demonstrated that experts, quasi-experts, and novices differ in their evaluations of product creativity, indicating that evaluative judgment may vary according to the evaluator's level of design expertise. These findings suggest that exposure to the same AI-generated output does not necessarily produce equivalent interpretations or judgments, because evaluators draw on different forms of knowledge, experience, and internal reference standards.

Based on these theoretical foundations, this study examines how designers and non-designers differ in their recognition, evaluation, and modification of identical AI-generated product design outputs. In particular, Norman's (2013) concept of visceral-level response and the halo effect originally identified by Thorndike (1920) provide a theoretical basis for understanding why visually polished outputs may receive favorable overall evaluations while attracting limited scrutiny of their structural, functional, or manufacturing-related limitations. This study conceptualizes such a response as False Premium Perception. In addition, differences in the internal reference frameworks used by expert and novice evaluators may lead identical rating scales to carry different meanings across groups. Drawing on Kahneman's (2011) account of intuitive and analytical judgment, this study conceptualizes this evaluative discrepancy as the Paradox of Scoring. Together, these concepts support an examination of how professional expertise influences not only the scores assigned to AI-generated designs, but also the problems recognized and the forms of intervention considered necessary.

3. Research Hypotheses and Methodology

This study aims to compare and analyze differences in design judgment between designers and non-designers when evaluating identical AI-generated outputs, with particular focus on their approaches to proposing improvements. Based on the theoretical review presented above, research hypotheses were formulated and a task-based comparative experiment was designed to empirically test them.

3. 1. Research Hypotheses

Prior research indicates that design expertise influences how evaluators recognize problems and assess design outcomes (Cross, 2004; Hu et al., 2023). This difference may be particularly relevant to AI-generated design images. Unlike conventional 3D model-based renderings, which are linked to explicit geometric data even when based on rough or simplified models, image-based generative AI can produce visually coherent outputs

without a corresponding three-dimensional model or consistent geometric information. This gap between visual plausibility and explicit design information increases the importance of sensemaking and interpretive judgment, as discussed above (Guo et al., 2023; Verganti et al., 2020). Thus, visual completeness may not correspond to underlying design validity, potentially leading designers and non-designers to focus on different aspects of the same output. Based on this literature and the characteristics of AI-generated outputs, the following hypotheses were proposed.

H1. Designers and non-designers will differ significantly in the distribution of problem-recognition types when evaluating identical AI-generated outputs.

H1-a. Non-designers will identify a higher proportion of problems related to surface-level visual elements, such as color, form, and texture.

H1-b. Designers will identify a higher proportion of problems related to functional and structural elements, such as structure, usability, and contextual fit.

H2. Designers and non-designers will differ significantly in the distribution of intervention strategies for identical AI-generated outputs.

H2-a. Non-designers will demonstrate a higher proportion of intervention strategies focused on partial modifications to individual design elements.

H2-b. Designers will demonstrate a higher proportion of intervention strategies focused on holistic structural improvements and changes to the overall design direction.

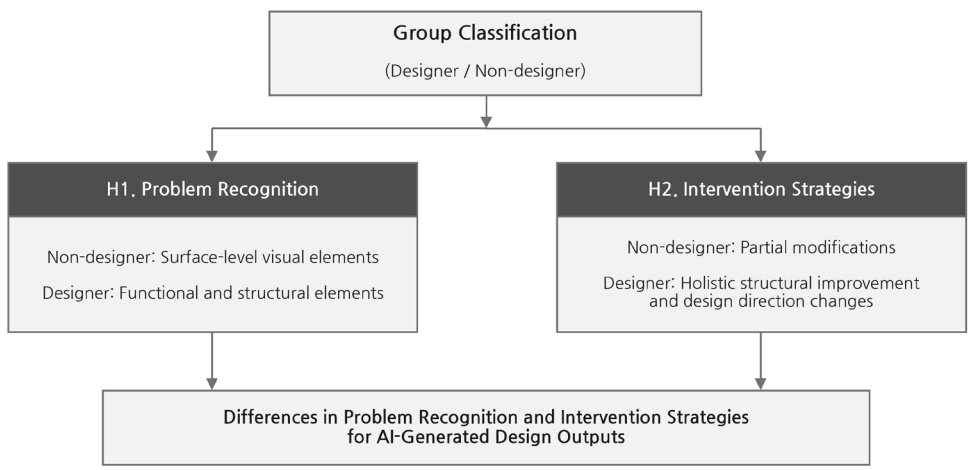

Figure 1 Hypothesized Group Differences in Evaluating AI–Generated Design Outputs

3. 2. Participants and Survey Respondent Characteristics

For the in-depth interviews, 12 participants were recruited, consisting of six industrial designers and six non-designers. The designer group had formal education and professional experience in industrial design, whereas the non-designer group had no design-related academic or professional background. The interview participants ranged from their late twenties to mid-fifties and were all men. The non-designer interviewees came from diverse occupational backgrounds, including sales, marketing, engineering, environmental safety, university teaching in a non-design field, and secondary education.

For the final survey, 75 participants were recruited. After excluding two respondents who had majored in industrial design but had no relevant professional experience, 73 valid responses were retained, comprising 32 industrial designers and 41 non-designers. The sample included 69 men and four women. The predominance of male participants resulted from prioritizing prior experience with electric screwdrivers rather than applying gender quotas. Age and occupation were not systematically collected in the survey because group classification was based on industrial design education and professional experience.

The semi-structured interviews were conducted face-to-face from March 5 to March 19, 2026, in quiet locations selected for participant convenience. A common set of guiding questions addressed initial impressions, perceived problems, reasons for judgment, suggested modifications, and overall evaluation. Each interview lasted approximately 15–20 minutes. The subsequent online survey was administered from March 26 to April 10, 2026. It consisted of 12 main items, including two sub-items under Q6, and included single-response, multiple-response, open-ended, and rating-scale questions and required approximately 10 minutes to complete. The interviews were used to obtain detailed accounts of participants' judgments, while the survey examined whether the observed differences were also evident across a larger sample.

Before participation, all participants were informed of the study purpose and procedures. Interview participants provided prior consent, and survey participants indicated consent after reviewing the online information sheet. Participation was voluntary, and identifying information was removed from the data used for analysis and reporting.

Because this study focuses on differences in interpretation and judgment rather than proficiency in using generative AI tools, AI utilization capability was not included as a direct analytical variable.

Table 1 Survey Participant Characteristics

Group	n	Gender (M/F)	Electric screwdriver use, n
Industrial designers	32	30/2	32
Non-designers	41	39/2	37
Total	73	69/4	69

3. 3. AI-Generated Design Stimulus

The target product for this study was an electric screwdriver. This product type was selected because it requires simultaneous evaluation of both usability and structural judgment — encompassing grip, center of gravity, button operability, and structural stability — making it well-suited for comparing surface-level, impression-based assessment with structurally informed design judgment.

The design stimulus used in this study was generated using Midjourney and selected according to the following criteria. First, the image should not strongly evoke any specific brand identity. Second, the level of visual completeness or visible defects should not be so pronounced as to simplify or predetermine participant judgment. Third, the stimulus should afford interpretation across a range of design elements, including structure, material, interface, and grip. The final stimulus was confirmed through review by two product design experts and was used consistently across both the comparative experiment and the survey.



Figure 2 AI-Generated Design Draft

3. 4. Tasks and Data Collection

Both the interviews and the survey addressed four questions: (1) which element required the most improvement, (2) why it was considered problematic, (3) how the design should be improved, and (4) how complete the design appeared on a 10-point scale.

During the interviews, participants reviewed the same printed image of an AI-generated electric screwdriver design. The researcher asked follow-up questions when necessary to clarify the participants' judgments and reasoning. The participants' responses were documented in writing, while audio recordings were used as supplementary material to verify the written records. The written responses were then used for qualitative coding and comparative analysis between the two groups. In the online survey, the same image was embedded in the questionnaire, and responses were collected through single-response, multiple-response, open-ended, and rating-scale items for group comparison.

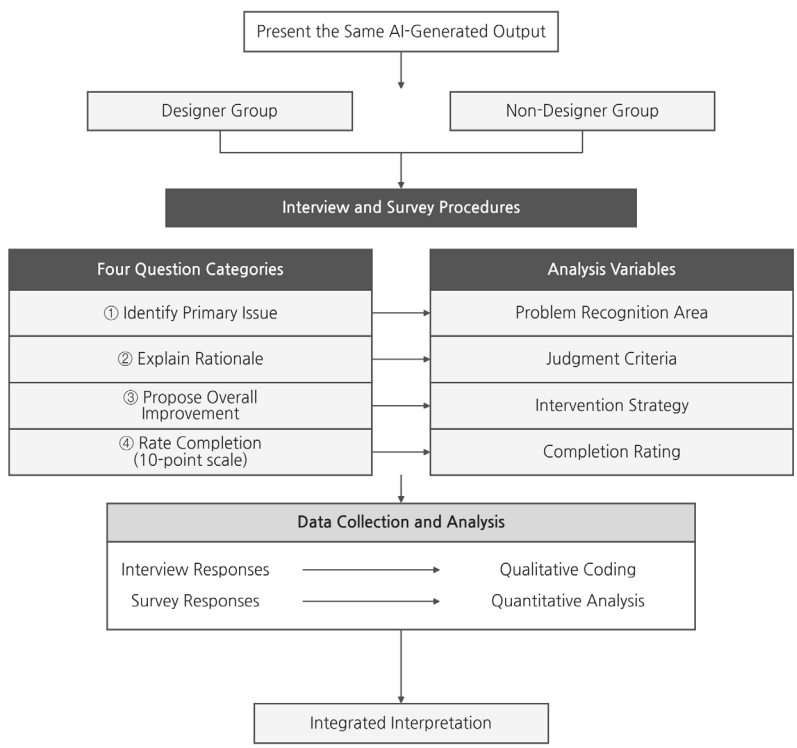

Figure 3 Data Collection Structure of Comparative Experiment and Additional Survey

3. 5. Analysis and Visual Reconstruction

The written records of participants' responses were segmented into meaning units and analyzed through open coding. Coding focused on three dimensions: the target of improvement intervention, the criteria underlying judgment, and the depth of intervention. Through this process, core categories related to types of problem recognition and approaches to improvement suggestions were derived. Coding was conducted by the first author and reviewed by the corresponding author, and any discrepancies were resolved through discussion.

Survey data were analyzed using Microsoft Excel and SPSS 28.0. Descriptive statistics were used to calculate frequencies and percentages, while independent-samples t-tests and chi-square tests were conducted to compare differences between the two groups. T-tests were applied to mean-based items, such as perceived completeness and product feasibility, whereas chi-square tests were applied to categorical response distributions, including problem-recognition areas, judgment criteria, and intervention levels. Statistical significance was determined at $p < .05$. Because Q5, Q6, and Q7 allowed multiple responses, each selected category was recoded as a binary variable, and group differences were examined using chi-square tests. Given the exploratory nature of this study, Bonferroni correction was not applied; therefore, the results should be interpreted as preliminary.

Representative improvement directions for each group were visually reconstructed based on judgment categories commonly identified across both the interview and survey data.

4. Results

This chapter presents an integrated account of findings from both the in-depth interviews and the survey analysis. Group differences were examined based on judgment categories derived from the interviews, and the survey was used to assess whether these tendencies were replicated across a broader sample.

4. 1. Initial Coding Results and Derivation of Judgment Categories

Open coding was conducted by identifying statements describing design issues, evaluation criteria, and suggested improvements. Semantically similar statements were assigned provisional codes and grouped according to the design attributes they addressed. For example, comments on excessive or redundant LEDs were coded as lighting, those on glossiness, slipperiness, scratching, and material weight as material, and those on handle shape, holding comfort, and operational safety as grip and ergonomics. Comments on parting lines, mold configuration, and production constraints were coded as manufacturability, while those on battery volume, lower-body configuration, and standing stability were grouped under structural stability. Through repeated comparison, five categories were derived: lighting, material, grip and ergonomics, manufacturability, and structural stability, as presented in Table 2.

Although both groups commented on some of the same design elements, they interpreted them differently. Designers frequently connected visible features to usability, safety, material performance, structural feasibility, and production constraints. Non-designers evaluated the

design primarily through visual impression, perceived durability, futuristic appearance, and anticipated usability, although some also identified functional concerns such as excessive lighting, battery size, and standing stability. Thus, the principal difference lay not only in what the two groups noticed, but also in the depth and basis of their judgments

Table 2 Summary of Initial Open-Coding Results: Designer Group Judgment Categories

Judgment Category	Key Coding Content	Detail Image
Lighting	Excessive lighting / Meaningless illumination / Risk of visual confusion / Concern over reduced battery efficiency	
Material	Inappropriate use of chrome/glossy materials / Risk of scratches / Risk of slipping / Inconsistency with use environment	
Grip and Ergonomics	Instability when placed upright / Risk of slipping / Insufficient material in hand-contact area	
Manufacturability	Excessive parting lines / Structural complexity / Increased mold fabrication difficulty	
Structural Stability	Unstable lower structure / Lack of stability when placed / Unbalanced center of gravity	

4. 2. Comparison of Judgment Structures Between Designer and Non-Designer Groups

The comparative coding presented in the previous section indicated differences in how designers and non-designers interpreted the same AI-generated design output. Designers tended to connect visible features to usability, safety, material performance, structural feasibility, and manufacturability. In contrast, non-designers relied more strongly on immediate visual impressions, including perceived durability, futuristic appearance, and anticipated usability. Differences were also observed in the suggested improvements. Designers frequently proposed interventions addressing structural or production-related causes, whereas non-designers more often suggested adjustments to color, size, lighting, or visible form. Some non-designers nevertheless identified functional concerns, including battery size, standing stability, and battery replacement.

Participants also provided ratings on a 10-point scale during the interviews. Designers' ratings reflected their assessment of the design's formal completeness, considering proportion, form coherence, and structural plausibility. Non-designers' ratings, by contrast, were generally accompanied by comments concerning visual appeal, futuristic appearance, perceived durability, and overall acceptance. Thus, numerically similar ratings did not necessarily represent the same judgment criteria across the two groups. This divergence is referred to in this study as the Paradox of Scoring.

Table 3 Comparison of Judgment Structures between Designers and Non-Designers

Category	Designer Group	Non-Designer Group
Evaluation Focus	Usability, structural validity, and manufacturability	Appearance, future-oriented imagery, intuitive usability
Problem Recognition Scope	Potential risks across form, material, structure, and interface	Tendency to overlook issues without obvious defects
Judgment Criteria	Context of use, safety, manufacturability, functional hierarchy	Personal experience and intuition-based sensibility
Improvement Suggestions	Structural and logical proposals linking cause and consequence	Partial adjustments focused on appearance and detail
Intervention Level	High-level restructuring of design and functional relationships	Minor adjustments preserving the existing design
Score Interpretation	Indicator of problem recognition and need for improvement	Indicator of overall satisfaction and visual completeness

4. 3. Survey-based Analysis of Group Differences

A total of 73 valid survey responses were analyzed, including 32 designers and 41 non-designers. Among the designers, 87.5% (n = 28) had 15 or more years of professional experience. Given the convenience sampling approach, the estimated margins of error at the 95% confidence level—±11.5 percentage points for the full sample, ±17.3 for designers, and ±15.3 for non-designers—were treated as supplementary indicators.

The survey consisted of 12 main items, including two sub-items under Q6. The survey items were organized as follows: Q1 addressed the feature first noticed; Q2 assessed overall design completeness; and Q3 collected open-ended responses regarding notable characteristics. Q4 asked whether the design required improvement, while Q5 identified the areas requiring improvement. Q6 examined the grounds for participants' judgments, Q6-2 distinguished between intuitive and knowledge-based judgment, and Q6-3 collected specific improvement suggestions. Q7 identified the criteria considered most important in evaluating the design, Q8 assessed the perceived feasibility of realizing it as an actual product, and Q9 examined the type and depth of the proposed intervention. Q10–Q12 collected information on product-use experience, design expertise, and professional experience. Based on these items, the group differences were analyzed across three dimensions: overall value assessment, problem recognition, and judgment approach and intervention depth.

4. 3. 1. Overall Value Assessment of the Product (Q2, Q8)

Analysis of overall completion ratings (10-point scale) and perceived feasibility as an actual product (5-point scale) for the AI-generated electric screwdriver design revealed statistically significant differences in perception between the two groups.

Table 4 Differences in Overall Value Evaluation between Designers and Non-Designers

Analysis Item	Designer (n=32)	Non Designer (n=41)	t	p
Overall Completeness(M)	5.47	7.29	−4.19	.000***
Product Feasibility (M)	2.94	3.80	−3.26	.002**

**p < .01, ***p < .001

The non-designer group assigned relatively high scores on both overall completeness and product feasibility, whereas the designer group provided lower ratings, reflecting greater

consideration of whether the design could be realized as a functional and manufacturable product. This pattern suggests that non-designers evaluated the output primarily on the basis of visual completeness, while designers applied internal design logic—including structural validity and manufacturability—as their evaluative framework.

4. 3. 2. Focus and Interpretive Uncertainty in Problem Recognition (Q4, Q5, Q6–3)

In response to Q4, 63.4% of non-designers either reported no problem (46.3%) or indicated uncertainty (17.1%), failing to arrive at immediate problem recognition. This pattern suggests a state of interpretive uncertainty, in which the visual completeness of the AI-generated image deferred engagement with functional validity assessment. In contrast, 87.5% of designers identified problems, and no instances of judgment suspension were observed in this group. Regarding problem-recognition areas, analysis of the 43 valid responses to Q5 showed that designers selected structural/manufacturability issues (50.0%) and lighting-related issues (50.0%) more frequently than non-designers (13.3% each). Category-level chi-square analyses showed significant group differences in both lighting-related issues and structural/manufacturability issues, $\chi^2(1, N = 43) = 4.16$, $p = .041$ for each category. Open-ended responses (Q6-3) further reflected this divergence: designers articulated technical problems informed by manufacturing processes and use context, while non-designers remained at the level of surface impression-based description.

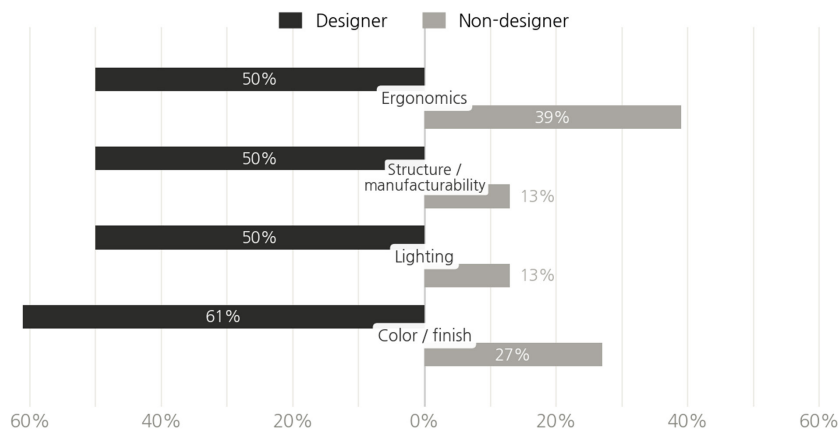

Figure 4 Comparison of Problem Recognition Areas

4. 3. 3. Differences in Judgment Approach and Intervention Depth (Q6–2, Q7, Q9)

The designer group based their evaluations predominantly on design- and engineering-related knowledge (53.6%), whereas the non-designer group relied on intuitive impression-based judgment (60.0%) and personal use experience (40.0%). This difference in judgment approach was statistically significant (Q6-2, $\chi^2(3, N = 43) = 17.60$, $p < .001$). With respect to evaluation criteria (Q7), the designer group frequently cited structural validity (62.5%) and manufacturability (25.0%) as primary standards, while the corresponding rates among non-designers were lower at 9.8% and 0.0%, respectively. Regarding intervention depth (Q9), the designer group centered their responses on fundamental interventions, including structural redesign (34.4%) and design direction redefinition (31.2%), whereas the non-designer

group showed higher proportions of aesthetic adjustment (36.6%) and acceptance without intervention (29.3%). This difference in intervention depth was also statistically significant (Q9, $\chi^2(3, N = 73) = 10.31$, $p = .016$).

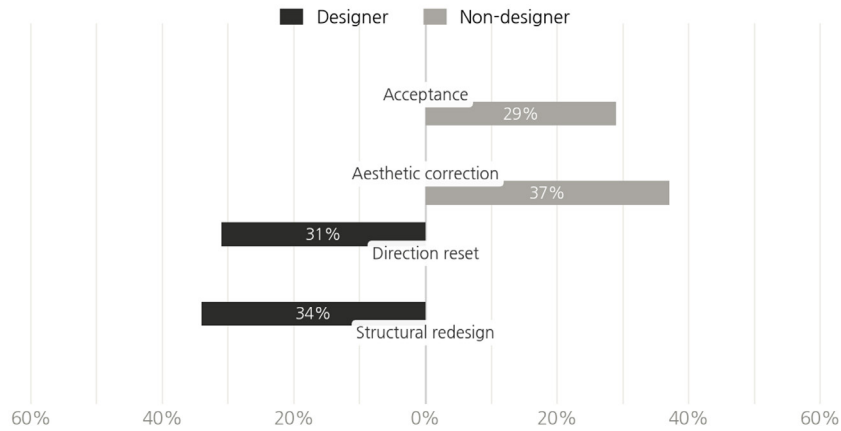

Figure 5 Comparison of Improvement Intervention Levels (Q9)

Taken together, the survey results partially support H1 and support H2. H1 was partially supported by category-level analyses of the multiple-response item Q5. Designers selected lighting-related issues and structural/manufacturability issues more frequently than non-designers, whereas differences in surface-level categories did not reach statistical significance. H2 was supported by a significant group difference in intervention depth (Q9, $\chi^2(3, N = 73) = 10.31$, $p = .016$). In addition, judgment approach differed significantly between the two groups (Q6-2, $\chi^2(3, N = 43) = 17.60$, $p < .001$).

4. 4. Visual Reconstruction of Representative Improvement Proposals by Group

Representative improvement proposals were visually reconstructed by the first author, a product designer with 21 years of professional experience, using ChatGPT-assisted prompting and Adobe Photoshop. The reconstructions were based on judgment categories derived from the interviews and survey responses. The elements were selected according to three criteria: recurrence across responses, validity grounded in design expertise, and depth of problem recognition. The reconstructed proposals were reviewed by two senior product designers with 19 and 26 years of professional experience, respectively. Both reviewers specialized in consumer electronics product design. They reviewed whether the reconstructions appropriately reflected the participants' intended improvements and whether the proposed modifications were plausible in terms of usability, structural validity, and manufacturability.

The designer group's proposal focused on reestablishing structural and functional hierarchy by reducing decorative lighting, clarifying informational elements, and revising materials, parting lines, and lower structural elements in relation to usability, durability, and manufacturability. In contrast, the non-designer group's proposal mainly involved partial modifications to visually salient elements, such as surface appearance and perceived stability, while preserving the overall form and impression of the original AI-generated output.



Figure 6 Design Reconstruction Results Reflecting Group–Specific Improvement Directions

5. Discussion and Conclusion

This study compared how designers and non-designers perceive and interpret identical AI-generated design outputs. The findings show that the two groups differed not merely in surface-level preferences or ratings, but in the cognitive frames through which they interpreted the same output.

Designers recognized the AI-generated output as an intermediate design state requiring further review in terms of usability, structural validity, material appropriateness, interface semantics, and manufacturability. In contrast, non-designers tended to evaluate the output through visual impression, intuitive appeal, and perceived completeness. This difference suggests that designers and non-designers do not simply assign different scores; rather, they apply fundamentally different evaluative logics.

The limited emphasis on aesthetics among designers should not be interpreted as indicating that aesthetic quality was unimportant to them. Rather, because the AI-generated stimulus already conveyed a relatively polished and futuristic visual impression, designers may have prioritized unresolved issues related to usability, structure, material appropriateness, and manufacturability. In addition, the task focused on identifying elements requiring improvement, which may have directed participants' attention toward problem detection rather than general aesthetic evaluation.

This distinction was also evident in the meaning of rating scores. For non-designers, scores functioned primarily as indicators of visual satisfaction and general acceptance, whereas for designers, scores reflected problem recognition and the perceived need for intervention. This phenomenon is conceptualized in this study as the Paradox of Scoring. In addition, the visual persuasiveness of generative AI outputs was found to obscure functional and structural deficiencies, leading to what this study terms False Premium Perception.

In conclusion, as generative AI raises the baseline of visual completeness, the role of designers becomes increasingly defined not by the ability to generate outputs, but by the capacity to critically interpret, evaluate, and redirect them. Generative AI can rapidly produce plausible

and visually refined images, but it cannot independently determine whether an output satisfies the requirements of use context, the organization and prioritization of functions and controls, material appropriateness, and manufacturability as an actual product. Therefore, the core role of the human designer in AI-integrated design environments can be defined as design-critical intervention: a judgment-centered competency that interprets AI-generated outputs within design contexts, identifies problems, and redirects them toward viable and implementable solutions. While non-designer evaluations provide useful insights into user acceptance and visual appeal, productization requires the experience-based judgment and professional intervention of trained designers.

As an exploratory study, this research has limitations in sample size and representativeness. In particular, because the designer group consisted largely of experienced product designers, the findings should be interpreted as reflecting the judgment tendencies of this expert-oriented sample rather than those of all designers. Although prior experience using electric screwdrivers was considered during participant recruitment, such experience alone may not fully capture differences in usage frequency, product familiarity, or manufacturing- and engineering-related knowledge. Future research should further validate the Paradox of Scoring using multidimensional measures of visual completeness, usability, structural validity, manufacturability, and market suitability. Nevertheless, this study contributes to empirically identifying differences in judgment structures between designers and non-designers and to redefining design expertise in generative AI environments as a capacity for interpretive framing and design-critical intervention.

References

1. Anantrasirichai, N., & Bull, D. R. (2022). Artificial intelligence in the creative industries: A review. *Artificial Intelligence Review, 55,* 589–656. https://doi.org/10.1007/s10462-021-10039-7
2. Chung, E. H., & Choi, J. M. (2025). Improvement directions of image generation AI to support product ideation. *Archives of Design Research, 38*(1), 179–196. https://doi.org/10.15187/adr.2025.02.38.1.179
3. Cross, N. (2004). Expertise in design: An overview. *Design Studies, 25*(5), 427–441. https://doi.org/10.1016/j.destud.2004.06.002
4. Dove, G., Halskov, K., Forlizzi, J., & Zimmerman, J. (2017). UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 278–288). Association for Computing Machinery. https://doi.org/10.1145/3025453.3025739
5. Guerrero, G., Mateus da Silva, F. J., Fernández-Caballero, A., & Pereira, A. (2022). Augmented humanity: A systematic mapping review. *Sensors, 22*(2), Article 514. https://doi.org/10.3390/s22020514
6. Guo, X., Xiao, Y., Wang, J., & Ji, T. (2023). Rethinking designer agency: A case study of co-creation between designers and AI. In D. De Sainz Molestina, L. Galluzzo, F. Rizzo, & D. Spallazzo (Eds.), *IASDR 2023: Life-changing design* (Article 170). International Association of Societies of Design Research. https://doi.org/10.21606/iasdr.2023.478
7. Hu, L., Boggess, M., & Shepley, M. M. (2023). Comparing expert, quasi-expert, and novice evaluations of award-winning design products using the consensual assessment technique. *Creativity Research Journal, 35*(4), 730–748. https://doi.org/10.1080/10400419.2022.2157591
8. Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
9. Kwon, J., Jung, E.-C., & Kim, J. (2024). Designer-generative AI ideation process: Generating images aligned with designer intent in early-stage concept exploration in product design. *Archives of Design Research, 37*(3), 7–23. https://doi.org/10.15187/adr.2024.07.37.3.7

10. Lee, Y., & Kim, S. C. (2025). AI-driven product design process: Exploring efficiency and creativity – Effectiveness and limitations of AI tools across design stages. *Journal of Korea Design Forum, 30*(1), 125-136. https://doi.org/10.21326/ksdt.2025.30.1.011
11. Na, D. Y. (2024). *생성형AI를 활용한 디자인 모델 연구: – AI디자인 프로세스 및 디자인 경영 적용 중심 – [A study on the design model by generative artificial intelligence: – Focused on the development of the AI design process & for the design management]* [Doctoral dissertation, International Design School for Advanced Studies, Hongik University].
12. Nelson, H. G., & Stolterman, E. (2012). *The design way: Intentional change in an unpredictable world* (2nd ed.). MIT Press. https://doi.org/10.7551/mitpress/9188.001.0001
13. Norman, D. A. (2013). *The design of everyday things* (Rev. and expanded ed.). Basic Books.
14. Paananen, V., Oppenlaender, J., & Visuri, A. (2024). Using text-to-image generation for architectural design ideation. *International Journal of Architectural Computing, 22*(3), 458-474. https://doi.org/10.1177/14780771231222783
15. Palani, S., & Ramos, G. (2024). Evolving roles and workflows of creative practitioners in the age of generative AI. In *Proceedings of the 16th Conference on Creativity & Cognition* (pp. 170-184). Association for Computing Machinery. https://doi.org/10.1145/3635636.3656190
16. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe, and trustworthy. *International Journal of Human-Computer Interaction, 36*(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118
17. Song, B., Zhu, Q., & Luo, J. (2024). Human-AI collaboration by design. *Proceedings of the Design Society, 4*, 2247-2256. https://doi.org/10.1017/pds.2024.227
18. Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29. https://doi.org/10.1037/h0071663
19. Verganti, R., Vendraminelli, L., & Iansiti, M. (2020). Innovation and design in the age of artificial intelligence. *Journal of Product Innovation Management, 37*(3), 212-227. https://doi.org/10.1111/jpim.12523
20. Wadinambiarachchi, S., Kelly, R. M., Pareek, S., Zhou, Q., & Velloso, E. (2024). The effects of generative AI on design fixation and divergent thinking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 380, pp. 1-18). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642919
21. Wang, N., Kim, H., Peng, J., & Wang, J. (2025). Exploring creativity in human-AI co-creation: A comparative study across design experience. *Frontiers in Computer Science, 7*, Article 1672735. https://doi.org/10.3389/fcomp.2025.1672735