A Ranking-Based Human Validation Layer for Concept Implementation in AI-Generative Design

Manuel Ibáñez-Arnal *, Luis Doménech-Ballester, Víctor García-Peñas

Posted Date: 10 March 2026

doi: 10.20944/preprints202603.0696.v1

Keywords: engineering design; hybrid AI–human evaluation; product semantics; conceptual design; generative AI; ranking methods; Plackett–Luce

 doi:10.20944/preprints202603.0696.v1

Article

A Ranking-Based Human Validation Layer for Concept Implementation in AI-Generative Design

Manuel Ibáñez-Arnal *, Luis Doménech-Ballester and Victor García-Peñas

Department of Mathematics, Physics and Technological Sciences (CEU Cardenal Herrera University)

* Correspondence: manuel.ibanez@uchceu.es

Abstract

Engineering design increasingly uses generative AI to explore large form spaces, yet concept-driven generation is only useful if observers consistently perceive the intended attribute. We propose a ranking-based human validation layer that tests whether AI-generated concept-intensity gradients are interpretable, reliable, and usable. For each Product–Concept pair, a controlled generative workflow produced six variants intended to increase concept expression (A–F). In an online study, 26 design engineers ranked the variants by perceived intensity, with an optional not-applicable (NA) flag when category recognition failed. We analyse rankings with heatmap diagnostics, inter-observer agreement, monotonic alignment with the intended order, and Plackett–Luce aggregation with uncertainty, while using NA trends to bound operational ranges. Across nine pairs, most gradients aligned with the intended direction, but performance depended on the concept and product context, revealing both stable and failure-prone segments. The approach provides an evidence-based gate for concept implementation in AI-generative design.

Keywords: engineering design; hybrid AI–human evaluation; product semantics; conceptual design; generative AI; ranking methods; Plackett–Luce

1. Introduction

Engineering design is an intentional activity in which product form serves not only functional requirements but also communicates meanings, values, and conceptual attributes that shape interpretation in use. During early design stages, designers frequently explore directions that can be operationalised as concept axes, such as aggressivity, elegance, or futurism, long before detailed performance optimisation becomes feasible. The usefulness of such conceptual directions depends on whether observers reliably perceive graded differences in the intended attribute across design variants. Traditional perceptual-evaluation approaches—such as semantic differentials and Kansei-based methods—have long enabled researchers to characterise perceived attributes, but they typically assess impressions of fixed designs rather than the interpretability of controlled conceptual variation. In this work, each target concept is treated operationally as a single conceptual axis for early-stage exploration; this is a practical representation for testing perceptual monotonicity rather than a theoretical claim about the concept's internal semantic structure.

The recent emergence of generative AI has dramatically expanded designers' ability to create large families of visually coherent alternatives at minimal cost. Yet increasing generative capacity introduces a new challenge: the ability to produce variation does not guarantee that this variation is perceptually coherent or monotonic. A sequence intended to progressively intensify a concept may be read inconsistently, non-monotonically, or in ways that diverge from the designer's semantic intent. For instance, a sneaker design intended to appear progressively "more futuristic" may do so initially, but further intensification can distort proportions or silhouette cues to the point where observers no longer recognise the item as belonging to the same product category. When this happens, the conceptual dimension ceases to be usable as a design variable. Validating whether concept-driven manipulations produce a monotonic, interpretable gradient is essential because only

 doi:10.20944/preprints202603.0696.v1

then can a conceptual axis be operationalised as a reliable design variable within generative workflows.

Despite the practical relevance of this question, there is no lightweight, general method for determining whether AI-generated conceptual gradients are perceptually coherent, where ambiguity emerges, or where intensification undermines recognisability. Existing perceptual-evaluation approaches were not designed to diagnose the structure of parametric conceptual variation produced by generative models, and current generative-design research rarely tests whether conceptual manipulations behave predictably for human observers.

This work addresses that gap by proposing a ranking-based human validation layer for concept-driven generative workflows. The approach provides an efficient perceptual check that determines whether a set of AI-generated variants intended to span increasing levels of a target concept is interpreted in the intended direction and with sufficient consistency to support early-stage design decisions. Rather than assessing preference or aesthetic appeal, the method focuses on interpretive directionality and consistency—whether observers perceive a coherent increase (or decrease) along the conceptual axis. In addition to recovering the intended ordering, we examine whether adjacent intensity levels are perceptually separable, identifying segments where local indistinguishability or inversions emerge. To handle extreme cases where conceptual intensification compromises category identity, a simple "not applicable" (NA) option marks recognisability failure, providing an operational signal of where the conceptual dimension ceases to be usable for design purposes.

We evaluate the protocol in an online study with expert design engineers using multiple Product–Concept pairs. Each pair consists of a product category combined with a target conceptual attribute (e.g., "futuristic sneaker," "elegant lamp"), instantiated through a controlled generative workflow that produces a fixed number of intensity levels intended to represent a monotonic increase in concept expression. Rankings are processed to address the following research questions:

- RQ1 (Reliability): Do observers exhibit consistent orderings for a given Product–Concept pair?
- RQ2 (Validity): Does the aggregated human ordering align monotonically with the intended A–F progression?
- RQ3 (Operational range): Do NA responses emerge systematically—e.g., at higher intended levels—indicating bounds within which concept intensification remains recognisable and practically usable?

This paper makes three contributions to engineering design research and practice:

1. Methodological contribution. We introduce a lightweight, reproducible perceptual-validation protocol based on ranking with NA and ordinal–probabilistic analysis. This method enables design teams to evaluate whether a concept-intensification axis generated by AI behaves as a perceptually monotonic, interpretable, and separable gradient—capabilities not addressed by existing product-semantics, Kansei, or generative-AI evaluation frameworks.
2. Conceptual contribution. We propose a general diagnostic framework describing three perceptual regimes—R1 stable, R2 stable-with-weak-segments, and R3 unstable—capturing distinct ways in which AI-generated conceptual gradients manifest perceptually. This framework provides an operational vocabulary for determining whether a conceptual dimension can be treated as a controllable design parameter.
3. Operational contribution. We reconceptualise loss of recognisability (NA) not as noise or missing data but as design-relevant evidence marking the functional boundary of concept intensification. This perspective enables the explicit identification of an operational range within which conceptual manipulation remains recognisable and practically usable in generative design workflows.

The remainder of the paper proceeds as follows. Section 2 reviews related work on product semantics, affective evaluation, ranking-based perceptual assessment, and human–AI collaboration in design. Section 3 details the experimental design, generative workflow, and analysis pipeline. Section 4 presents results across multiple Product–Concept pairs, reporting agreement, directional

 doi:10.20944/preprints202603.0696.v1

alignment, adjacent-level separability, and NA patterns. Section 5 discusses implications for concept-implementation workflows, limitations, and directions for future research.

2. Related Work

2.1. Product Semantics and Conceptual Intent in Product Form

Research in product semantics has long established that product form communicates meaning beyond technical function, positioning products as sign systems through which users infer category, affordances, values, and identity cues. Meaning emerges from the interpretive relation between observer and artefact rather than from form alone, implying that semantic intent must be inferred through perception, interpretation, and use rather than directly observed (Krippendorff, 2006). Within this tradition, concept implementation can be understood as the intentional encoding of semantic cues in form so that a target concept becomes legible to observers. Yet, as Kazmierczak (2003) argues, design operates as meaning-making embedded in culture, which makes the gap between designer intention and user interpretation inevitable. Cross-cultural research further reinforces this variability, showing that the same formal configuration can produce divergent meanings depending on social context (Rau, 2021).

An important implication for engineering design is that concept implementation is rarely binary. Designers often manipulate continuous degrees of an attribute—such as "more aerodynamic" or "more futuristic"—suggesting an underlying perceptual continuum rather than categorical shifts. The challenge is to control semantic cues in a graded and predictable way without compromising functional legibility or recognisability, especially during early design stages when exploration is abundant but interpretive evidence remains limited.

2.2. Measuring Perceived Attributes: Semantic Differential, Kansei, and Affective Evaluation

Efforts to measure subjective meaning in design have evolved through several complementary approaches. The semantic differential, introduced by Osgood, Suci, and Tannenbaum (1957), provided the first systematic tool for mapping impressions of artefacts along bipolar adjective pairs, allowing researchers to quantify perceived connotations in a multidimensional "semantic space." Building on this foundation, Schütte (2005) formalised Kansei Engineering, which seeks to translate affective responses into actionable design parameters, effectively linking user emotion to form features and enabling iterative refinement. These methods have been adapted to a wide range of product categories and cross-cultural contexts.

More recent developments integrate computational and data-driven methods to enhance semantic analysis. Siddharth, Blessing, and Luo (2022) demonstrated how natural language processing can extract affective meaning from textual corpora, enriching the traditional semantic differential framework with algorithmic scalability. Similarly, Feng et al. (2025) proposed an AI-assisted approach that integrates Kansei principles with knowledge-graph reasoning to generate user-centric interface designs, showing how semantic intent can guide generative prompts. Affective evaluation has also expanded toward physiological and behavioural dimensions; Guo et al. (2023) combined psychophysiological signals with perceptual ratings to strengthen empirical grounding in emotional design studies. Yet despite these advances, there remains a persistent need for lightweight, scalable methods capable of capturing meaningful perceptual distinctions in early-stage concept evaluation—without requiring extensive instrumentation.

2.3. Ranking and Comparative Judgement in Perceptual Evaluation

When constructs are inherently subjective or absolute calibration is unreliable, comparative methods such as paired comparisons, sorting, and ranking provide a more natural means of eliciting perceptual information. Ranking-based evaluation aligns closely with the way designers reason about form—making relative judgements such as "A expresses the concept more than B" rather than

 doi:10.20944/preprints202603.0696.v1

assigning arbitrary numeric scores. Empirical evidence suggests that relative methods can reduce scale-use bias and increase sensitivity in perceptual data (Perez-Ortiz & Mantiuk, 2018; Sciandra et al., 2021). From a modelling perspective, probabilistic frameworks such as the Plackett–Luce or Bradley–Terry models (Luce, 1959; Plackett, 1975; Turner et al., 2020) estimate latent "worth" parameters underlying ranked preferences. These models are interpretable, robust to incomplete data, and well-suited for aggregating ordinal information across observers. Importantly, they also handle the common challenge of recognisability breakdown—when extreme design manipulations distort category cues to the point that items become incomparable. In concept-driven design gradients, this property allows analysts to treat perceptual collapse not as error but as a meaningful limit of interpretability.

2.4. Generative AI in Conceptual Design and the Need for Validation

The rise of generative AI has revolutionised conceptual design by enabling the creation of large sets of visually coherent form alternatives at minimal cost. These systems have been shown to augment designers' creativity, broaden the search space, and accelerate iteration (Chen et al., 2024; Marrone, 2023). However, this proliferation of generative capacity introduces a new bottleneck: validation. The essential question is shifting from "Can AI generate alternatives?" to "Do these alternatives reliably implement the intended conceptual change in a way that humans interpret as intended?"

Empirical evidence underscores this challenge. In the AutoSpark study, Chen et al. (2024) integrated Kansei Engineering principles into a generative AI pipeline for automotive appearance design, demonstrating that affective intent can be encoded and evaluated systematically. Liang (2024) extended this line of inquiry to cultural and creative products, proposing adaptive validation frameworks sensitive to audience diversity. Luo (2025) developed a multi-modal evaluation system combining visual, textual, and affective data to verify whether AI-generated forms convey the desired concept with perceptual coherence. Collectively, these works show that generative novelty alone is insufficient for engineering design—what matters is the interpretive fidelity and monotonic control of conceptual expression across levels of manipulation.

2.5. Hybrid AI–Human Validation and Human-in-the-Loop Perspectives

Hybrid AI–human validation frameworks position human judgement not as auxiliary data collection but as an integral mechanism of governance within generative workflows. Human–AI collaboration research consistently emphasises that humans contribute contextual knowledge, interpretive calibration, and correction mechanisms that align model outputs with human goals (Amershi et al., 2019). In design contexts, Fattah Saleh (2025) illustrated how industrial designers integrate subjective reasoning and affective assessment within generative pipelines, while (Wang et al. (2024) proposed a co-annotation framework where human feedback iteratively refines AI-generated aesthetic variations. Theoretically, Comte (2022) argues that this reframing repositions designers as semantic mediators rather than autonomous creators—agents who ensure that AI systems remain accountable to human interpretive logic. Consequently, validation becomes the procedural bridge that transforms algorithmic exploration into evidence-based design knowledge, ensuring that generated artefacts remain both perceptually legible and semantically meaningful.

2.6. Positioning of the Present Work

Building on the traditions of product semantics, Kansei Engineering, and affective evaluation, the present work proposes a hybrid AI–human validation layer for concept-driven generative workflows. The method is designed to empirically test whether generative manipulations produce perceptually monotonic gradients of concept expression while maintaining recognisability. Unlike aesthetic preference studies, which often assess appeal or liking, the focus here is on interpretive consistency—whether a given concept (e.g., "aggressive," "elegant") is perceived as increasing or

 doi:10.20944/preprints202603.0696.v1

decreasing in strength across generated levels. Using ranking-based perceptual evaluation analysed through probabilistic models, the framework enables ordinal aggregation, uncertainty quantification, and cross-concept comparability. Moreover, by explicitly encoding "not applicable" or "unrecognisable" responses, the approach treats recognisability breakdown as an informative boundary condition—marking where concept expression begins to compromise category identity.

This perceptual monotonicity is crucial: if observers do not consistently perceive an increase (or decrease) in the intended concept across levels, then the concept cannot be operationalized as a controllable parameter in generative workflows, and the resulting design space becomes misleading or unusable.

3. Method

3.1. Study Overview and Experimental Design

This study evaluates whether AI-generated conceptual gradients are interpreted by human observers in the intended direction. The experimental unit is a Product–Concept pair, for which a controlled generative workflow produced six visual variants (A–F) representing an ordinal increasing intended expression of a target concept, without assuming perceptual equidistance between adjacent levels. Participants completed a perceptual ranking task, providing an ordered sequence from lowest to highest perceived concept expression.

Six levels were selected as a practical compromise between perceptual resolution and cognitive load. Fewer levels would have provided insufficient granularity to detect local inversions or recognisability breakdown, whereas adding more levels increases the combinatorial complexity of the ranking task and the likelihood of perceptual ties. The A–F scale therefore supports the methodological goals of detecting monotonicity, ambiguity, and recognisability loss without overloading participants.

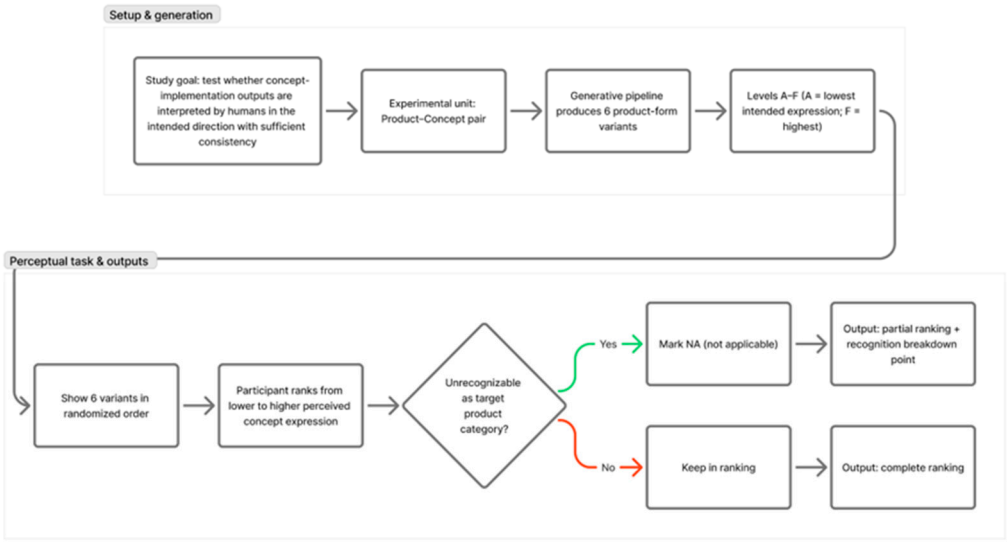

Figure 1. Study protocol and ranking output structure. Top: stimulus setup and generation. Bottom: perceptual ranking task.

An explicit recognisability mechanism was included to identify points at which conceptual intensification undermines product-category identity. For any stimulus that was no longer perceived as a valid instance of the target product category, participants were instructed to assign the label Not Applicable (NA). NA is strictly defined as a categorical judgement of recognisability loss and must

 doi:10.20944/preprints202603.0696.v1

not be used for any other reason—such as uncertainty, difficulty discriminating between adjacent items, or perceived ties. As Fig.1 shows, when a stimulus was marked NA, it was removed from the ordering interface, and participants ranked the remaining items. This mechanism enables the study to characterise the operational range within which conceptual manipulation remains recognisable, and therefore usable, as a design parameter.

3.2. Stimuli Generation

Stimuli were AI-generated product-form images produced under a controlled workflow designed to modulate a target concept across six intended levels (A–F). For each Product–Concept pair, the product category was held constant while the conceptual attribute was parametrically intensified across levels using a consistent base prompt structure and fixed visual constraints (viewpoint, framing, background, and lighting) to support comparability. This yields a monotonic progression along the generative-pipeline axis, with A–F corresponding to successively intensified concept prompts from the system's parametrization.

The generative pipeline was implemented in ComfyUI using a latent-diffusion image-to-image architecture (Figure 2). An input image is VAE-encoded into an initial latent representation; in parallel, positive and negative text prompts are embedded via the CLIP text encoder to provide semantic conditioning and constraint signals. Using an SDXL model checkpoint (UNet + CLIP + VAE), the diffusion sampler (KSampler) performs iterative denoising in latent space (20 steps, CFG = 8, Euler sampler, normal scheduler, fixed seed; and a denoise sweep) to generate a refined latent, which is then VAE-decoded back to image space and saved as the final output. No post-processing was applied beyond standard file export (PNG encoding) by ComfyUI. SDXL was selected as the generative backbone because it is an open-source, locally executable model. This guarantees long-term reproducibility since the results can be reproduced without dependence on commercial APIs.

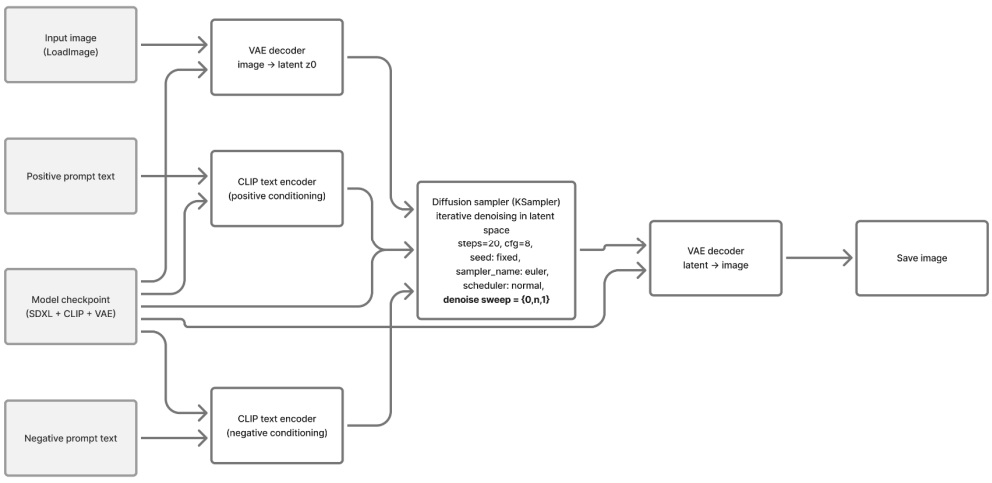

Figure 2. Compact schematic of the latent-diffusion image-to-image workflow implemented in ComfyUI.

The selection of these specific KSampler settings was informed by prior exploratory studies conducted by the authors, in which the principal parameters of the image-to-image pipeline—including CFG scale, step count, sampler type, scheduler, seed strategy, and denoise strength—were systematically varied and their effects on concept attribution assessed. These preliminary calibrations identified which parameters exerted the most significant influence on the perceived intensity and coherence of conceptual attributes in the generated forms, and guided the configuration adopted in

 doi:10.20944/preprints202603.0696.v1

the present study. A detailed account of this parametric exploration is beyond the scope of this paper and will be reported separately.

For each level, multiple candidates were generated under identical sampling settings and a single representative image was selected using predefined curation criteria: (i) product category fidelity, (ii) absence of unintended artifacts, (iii) compliance with the fixed visual constraints, and (iv) plausibility of the target concept intensity for that level. To support full replicability, the complete ComfyUI workflow, the input base images, the prompt specifications for each level, and the generated stimuli used in the study are provided as supplementary material.

The study included nine Product–Concept pairs: *Bottle–Bubble, Bottle–Origami, Vase–Ruby, Lamp–Polygonal, Table–Contrast, Chair–Aerodynamic, Chair–Aerodynamic 2, Sneaker–Silence, and Sneakers–Futuristic*. The input images serving as the base for the image-to-image pipeline were real product photographs selected by the researchers according to deliberate design criteria intended to minimise perceptual confounds. Base images depicted products against neutral backgrounds, exhibiting pure and elementary formal attributes with no salient pre-existing conceptual connotations that could bias observers' subsequent rankings—except in cases where such connotations were deliberately retained for research purposes. Products were required to be highly recognisable exemplars drawn from diverse categories, ensuring broad coverage of the design space. Purely functionalist products were explicitly excluded on the grounds that strong form–function coupling could lead raters to conflate usability cues with aesthetic-conceptual perception, thereby undermining the validity of the ranking task. The selected products were therefore those in which styling plays a prominent role while still departing from minimal, archetypal forms. This selection was conducted by the research team on the basis of the study's hypotheses and objectives; no inter-rater agreement procedure was applied at this stage, as the purpose was to configure the experimental conditions rather than to elicit perceptual judgements. These pairs were selected to span multiple product categories and concept types (e.g., geometric, stylistic, metaphorical), ensuring that the validation approach was exercised across a wide range of scenarios representative of conceptual design practice. Across pairs, the intent was to test the same validation logic under heterogeneous conditions typical of conceptual design practice, rather than to optimise performance for a single object or concept.

The generative pipeline may modify multiple formal attributes simultaneously when intensifying a conceptual prompt, given the entangled nature of latent generative representations. We do not aim to isolate or manipulate a single semantic feature in a controlled factorial sense. Instead, our interest lies in assessing whether the aggregate effect of the prompt intensification results in a perceptually coherent directional gradient as interpreted by human observers. In this framework, any auxiliary or emergent variations introduced by the model—such as changes in curvature, sharpness, colour accents or textural cues—are considered part of the model's expressive mechanism for rendering the specified concept and are therefore accepted as natural within the evaluation protocol.

3.3. Survey Instrument and Procedure

The evaluation was implemented through an online drag-and-drop interface. For each trial, the six stimuli of a Product–Concept pair were displayed simultaneously in a randomised layout. Participants were instructed to order the items from lowest to highest perceived expression of the target concept, forming a strict total order without ties. If two variants appeared similar, participants were asked to select the ordering that best reflected even subtle differences. Figure 3 shows the response collection platform.

 doi:10.20944/preprints202603.0696.v1

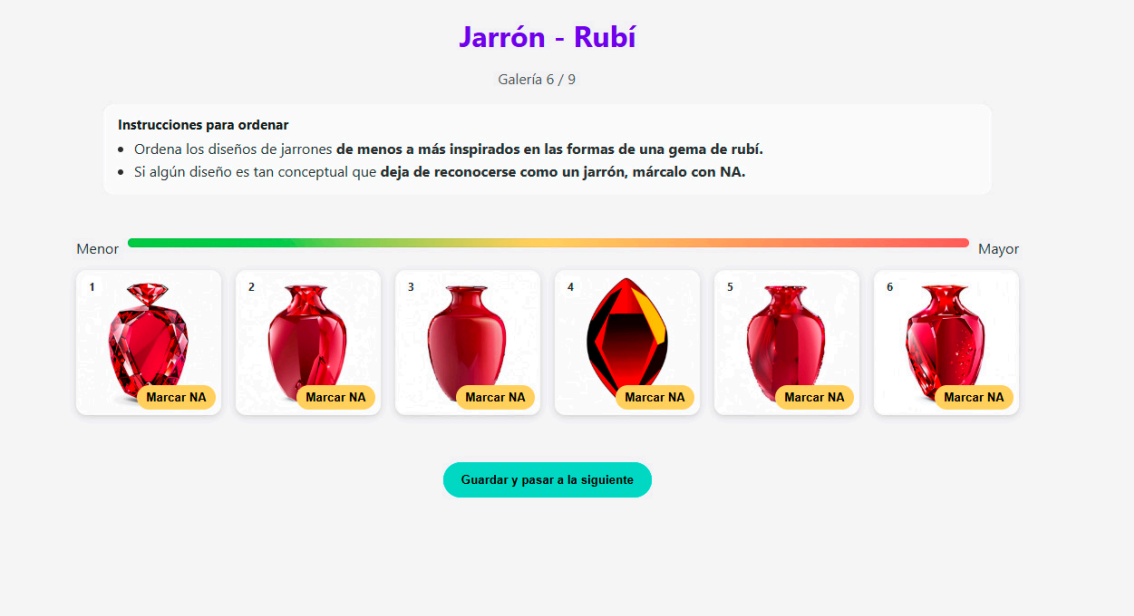

Figure 3. Screenshot of the ranking-based perceptual task interface for the Product–Concept pair "Vase–Ruby". Six AI-generated product-form stimuli are presented simultaneously along a visual lower → higher intensity continuum.

In accordance with the definition provided previously, participants could mark any stimulus as NA when it was not recognisable as belonging to the target product category. NA items were automatically removed from the interface, and the participant completed a partial ranking with the remaining stimuli.

Participants confirmed their ordering before proceeding to the next trial. All instructions were provided in Spanish, matching participants' native language.

3.4. Participants and Data Collection

This study received ethics approval from the Ethics Commitee of the Universidad Cardenal Herrera-CEU (CEEI 25/705).

Participants were qualified design engineers (≥5 years' experience), recruited to ensure familiarity with product-form cues and reduce ambiguity in concept-expression judgments. The final sample comprised N = 26 qualified design engineers. Participant professional experience in engineering design ranged from 5 to 22 years ($M = 12.23$, $SD = 4.04$, $Mdn = 13.5$). Recruitment was conducted via professional networks, alumni networks and industry contacts, and all participants provided informed consent prior to starting the survey. Responses were collected remotely through the online survey, completed on participants' own devices. Participation was voluntary and responses were anonymised prior to analysis. The study was conducted in Spanish with native Spanish-speaking participants; all task instructions and concept labels were presented in Spanish.

The linguistic and cultural homogeneity of the participant sample was a deliberate methodological choice. By ensuring that both the researchers who designed and administered the study and the observers who evaluated the stimuli shared the same native language and cultural background, any potential perceptual or interpretive gap arising from cross-cultural or cross-linguistic differences was controlled for at the design stage. This decision prioritises internal validity: since conceptual labels such as "futuristic," "silence," or "origami" may carry different connotations across languages and cultural contexts, a shared linguistic and cultural frame between emitters and receivers ensures that the rankings reflect differences in concept-form perception rather than differences in semantic interpretation of the task labels. The implications of this choice for external generalisability are acknowledged in the Limitations section.

 doi:10.20944/preprints202603.0696.v1

Twenty-six qualified design engineers participated in the study. Sample sizes of this magnitude are common in expert-panel research in design semantics, Kansei Engineering, and perceptual evaluation, where recruitment is constrained by the need for domain-specific expertise (e.g., Schütte, 2005; Ishihara et al., 2008; Georgiev & Georgiev, 2020). Because each complete ranking implicitly encodes a dense set of pairwise comparisons, ranking-based tasks provide high informational yield per participant, reducing the need for large samples compared to rating-scale methodologies (Perez-Ortiz & Mantiuk, 2017). Ordinal approaches have further been shown to be well suited for subjective perceptual assessment tasks where absolute calibration is unreliable (Sciandra et al., 2021). Moreover, the aim of the present study is methodological validation—evaluating the perceptual structure of concept-intensification gradients—rather than population-level estimation. For this type of objective, expert samples in the range of 20–30 participants are typical and sufficient in prior design-research literature.

3.5. Data Structure and Preprocessing

For each participant and Product–Concept pair, the response is an ordered list of level labels drawn from the set {A, B, C, D, E, F}, assigned to ordinal positions 1–6. NA-labelled items create missing positions, producing partial rankings whose length varies across participants and pairs. Raw survey outputs were validated to ensure internal consistency (e.g., no duplicate assignments, no impossible position allocations), then converted into an ordinal representation suitable for rank-based statistics and probabilistic modelling.

Rankings were treated as directional according to the task instruction (lower→higher). NA responses were retained as an explicit outcome variable rather than discarded. This separation supports later analyses that distinguish disagreements about concept expression from failures of category recognition.

Visual consensus diagnostics: For each Product–Concept pair, a level-by-position contingency matrix was computed, counting how often each nominal level (A–F) was placed in each ranking position (1–6). These matrices were visualised as heatmaps. Concentration along a diagonal pattern indicates convergence toward the intended monotonic ordering, while off-diagonal structure may reflect confusion, inversion, or local indistinguishability between adjacent levels.

3.6. Analysis of NA Responses and Operational Range

In line with RQ3, NA responses were treated as indicators of category-recognition breakdown and potential saturation in the concept-intensification process. For each Product–Concept pair, the proportion of NA responses was computed per nominal level (A–F). Rising NA frequencies at higher levels may be consistent with over-extension, whereas NA at lower levels would signal an early breakdown or floor-effect in which minimal manipulation undermines category identity. This behaviour defines the operational range within which the conceptual axis remains recognisable and practically usable.

To evaluate whether recognisability breakdown occurs consistently across observers, we computed several agreement coefficients on the binary omission variable (NA vs. Non-NA) for each level. The observed agreement (Po) quantifies the raw proportion of coincident omission decisions. Fleiss' κ and Krippendorff's α estimate interobserver agreement beyond chance, with α being robust to the presence of missing values. Because NA can be highly prevalent or extremely rare depending on the Product–Concept pair, we also computed Gwet's AC1, which corrects for prevalence induced biases that typically deflate κ. High agreement indices indicate that observers converge on the same subset of levels as non-recognisable, supporting a well-defined operational boundary. Conversely, low agreement suggests heterogeneous omission criteria or the presence of subgroups operating in different recognisability regimes, in which case no shared operational boundary can be assumed.

In addition to agreement metrics, we evaluated whether omission patterns were systematically related to the intended level. First, a chi-square test of independence (NA vs. Non-NA × A–F) was computed to assess whether NA occurrence was randomly distributed across levels or whether

 doi:10.20944/preprints202603.0696.v1

certain levels concentrated omissions disproportionately. Second, to test for monotonic patterns in omission rates along the A→F progression, both the Cochran–Armitage trend test and Jonckheere-Terpstra test were computed using a two-sided alternative. Directionality (increasing vs. decreasing) was determined from the sign of the standardized Z statistic. A significant trend indicates that NA frequencies vary systematically across levels, revealing where recognisability tends to fail more frequently along the intended axis. Together with the agreement indices, these tests distinguish between (i) homogeneous, collectively shared boundaries of recognisability and (ii) heterogeneous or level-unspecific omission patterns.

3.7. Probabilistic Aggregation and Uncertainty Estimation

To aggregate rankings and estimate uncertainty, a Plackett–Luce (PL) model was fitted for each Product–Concept pair. The PL model estimates a "worth" parameter for each level, reflecting its probability of being ranked above other levels under a probabilistic choice interpretation. This provides an interpretable aggregate ordering together with model-based estimates of uncertainty for each level. Because PL models naturally accommodate partial rankings, NA-induced omissions do not disrupt the estimation process.

Uncertainty estimation is particularly important for engineering design interpretation because adjacent levels may be perceptually close even when the overall ordering is monotonic. To formalise this idea, we refer to local separability (or adjacent-level separability) as the degree to which consecutive intended levels are perceptually differentiated. Whereas global monotonicity concerns the preservation of the overall A→F direction, local separability is a local property of the gradient that captures the potential overlap or ambiguity between neighbouring levels. Within the present framework, local separability is assessed through the uncertainty associated with the Plackett–Luce worth estimates and through the consecutive pairwise comparison probabilities derived from the model.

All analyses were implemented in a reproducible computational workflow. The manuscript reports, for each Product–Concept pair: (i) descriptive response distributions and heatmaps, (ii) agreement metrics, (iii) alignment statistics, (iv) PL worth estimates with uncertainty, and (v) NA patterns across levels. Given N = 26 observers, each pair yields up to 26 rankings, with reduced effective comparisons in cases where NA produces partial rankings.

3.8. Outcome Measures

The outcome analysis in this section addresses RQ1 and RQ2 by focusing on two complementary properties of the ranking data: interobserver reliability and directional validity. Reliability captures whether participants converge on a consistent ordering of the stimuli, whereas validity assesses whether that consensus—when present—follows the intended A–F progression defined by the generative pipeline. Distinguishing these two dimensions is essential, since participants may agree on an ordering that does not match the intended direction, or conversely, they may align with the direction while exhibiting low overall agreement. This separation clarifies the respective roles of reliability as a precondition for interpretability and validity as evidence of perceptual monotonicity.The outcome analysis in this section addresses RQ1 and RQ2 by focusing on two complementary properties of the ranking data: inter-

Reliability: inter-observer agreement. To quantify whether observers share a consistent ordering, agreement among participants was assessed using ordinal-appropriate inter-rater measures (e.g., Kendall's W; Krippendorff's alpha for ordinal data).

Validity: alignment with intended level order. To test whether human judgements align with the intended progression defined by the generative pipeline, rank-based association measures were computed between the aggregated human ordering and the A–F reference order (e.g., Spearman's ρ and/or Kendall's τ). High positive association indicates that participants perceive increasing concept expression in the intended direction; low or negative association may reflect, among other factors, weak interpretability, ambiguity, or systematic inversion of the intended concept axis.

 doi:10.20944/preprints202603.0696.v1

The perceptual constructs examined in this study—agreement among observers, directional alignment with the intended A–F order, monotonicity, and separability between adjacent levels—are conceptually distinct and cannot be fully characterised by any single ordinal statistic. Different indicators capture different aspects of these constructs and rely on different assumptions regarding data structure, scale interpretation, and the presence of partial rankings. For this reason, the analysis incorporates a set of complementary measures rather than relying on a single index. This approach allows each construct to be assessed using tools specifically suited to its properties, without implying that all indicators measure the same phenomenon or that they should lead to identical outcomes.

Anonymised ranking data and analysis scripts may be made available as supplementary material upon publication, in line with institutional and publisher guidelines.

4. Results

Figure 4 displays the level-position heatmaps for all Product-Concept pairs. Each matrix shows the frequency with which levels A-F were placed in positions 1-6 across participants, with darker cells indicating higher counts. The diagonal shading patterns represent how often the intended ordering was recovered, while off-diagonal placements show deviations or local reassignments. The NA row captures items that were excluded from rankings due to loss of recognizability.

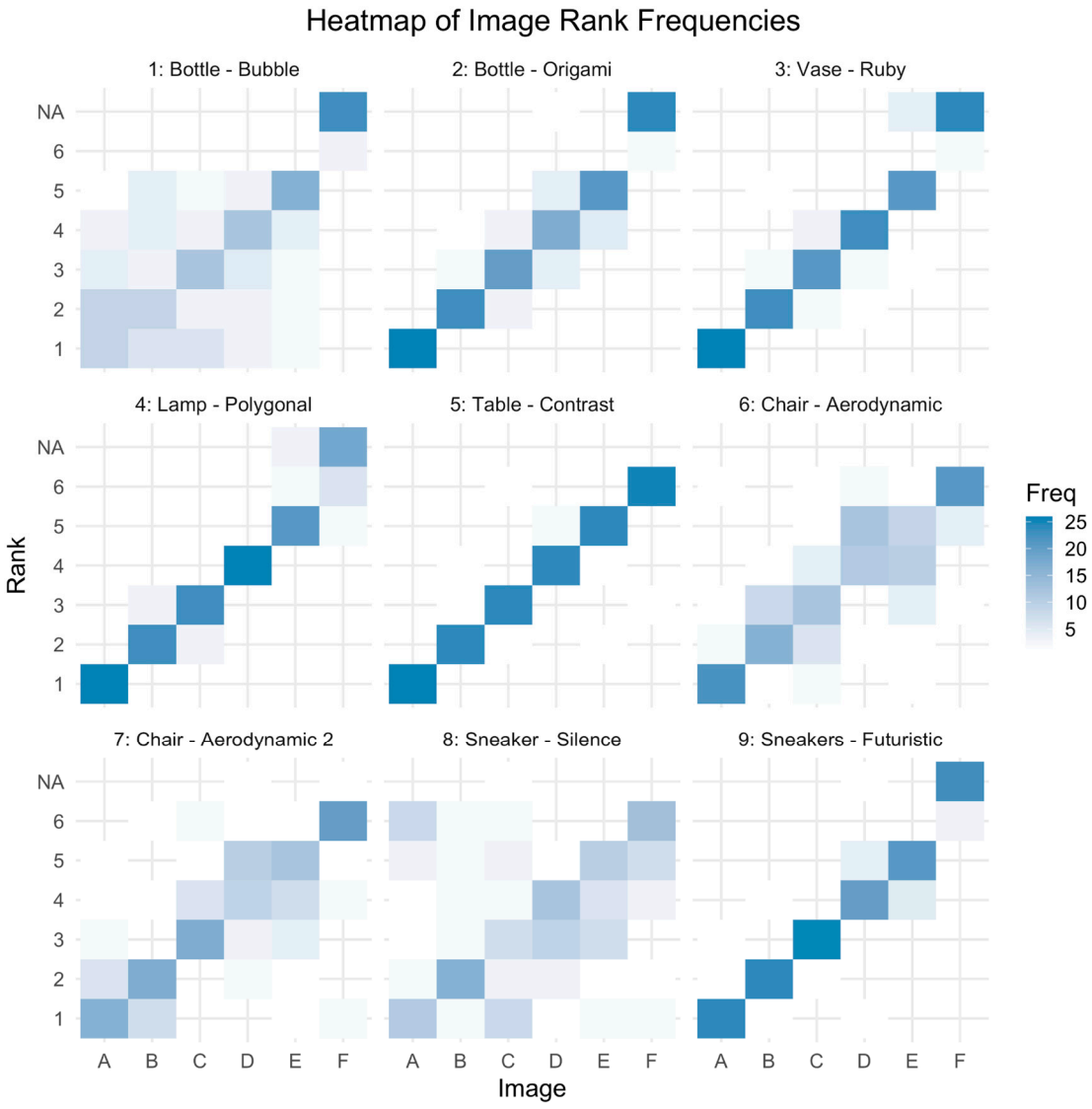

Figure 4. Level–position heatmaps for all nine Product–Concept pairs. Each panel shows how often each intended level (A–F) was assigned to each ranking position (1–6 and NA).

 doi:10.20944/preprints202603.0696.v1

Figure 5 presents the estimated probability (from the ordinal model) that each higher-numbered level was placed above the preceding one (e.g., 2 over 1, 3 over 2). Bars near 1.0 indicate frequent consecutive ordering in the expected direction, whereas values closer to 0.5 reflect reduced local separability between those adjacent levels.

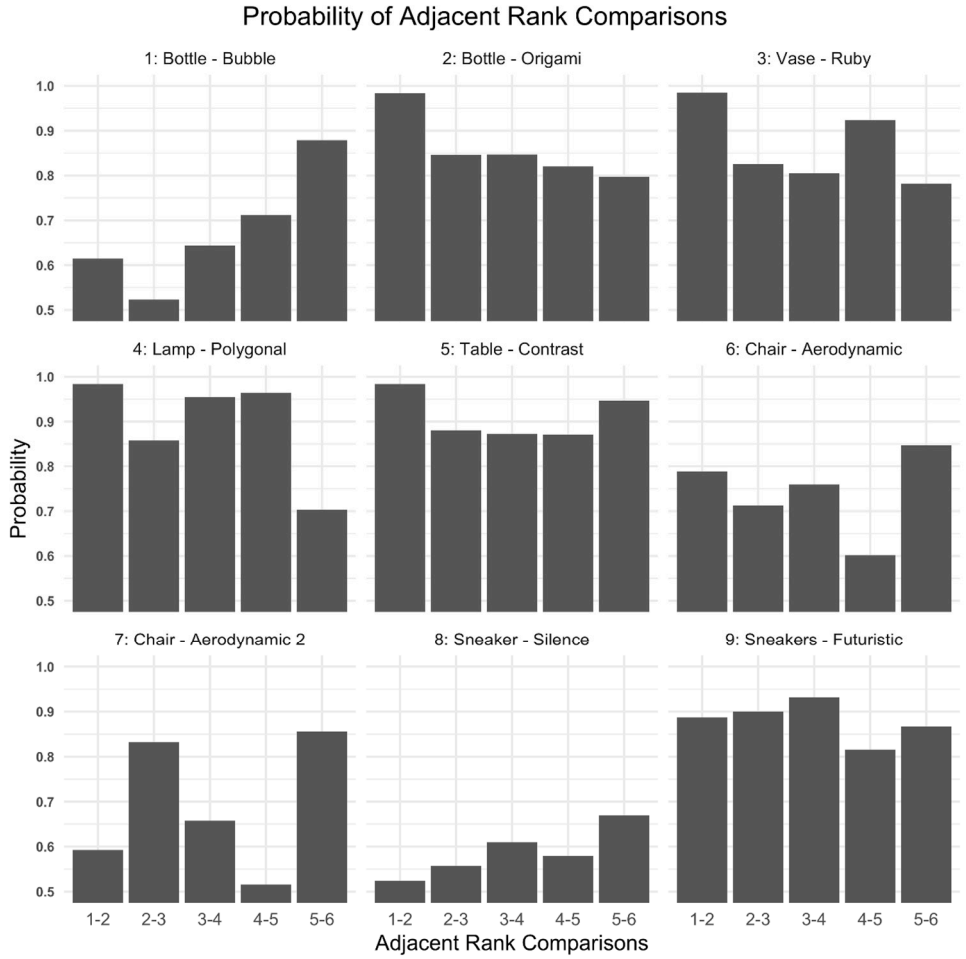

Figure 5. Estimated probability (Plackett–Luce) that the highernumbered level is preferred over the lower-numbered one.

Figure 6 shows the log-worth estimates obtained from the Plackett-Luce aggregation for levels A-F, together with their confidence intervals. The relative horizontal position of the estimates reflects the ordering recovered by the model, while the width and degree of interval overlap indicate the uncertainty associated with each level. Substantial overlap denotes limited separation between levels, whereas non-overlapping intervals correspond to clearer distinctions in the aggregated data.

 doi:10.20944/preprints202603.0696.v1

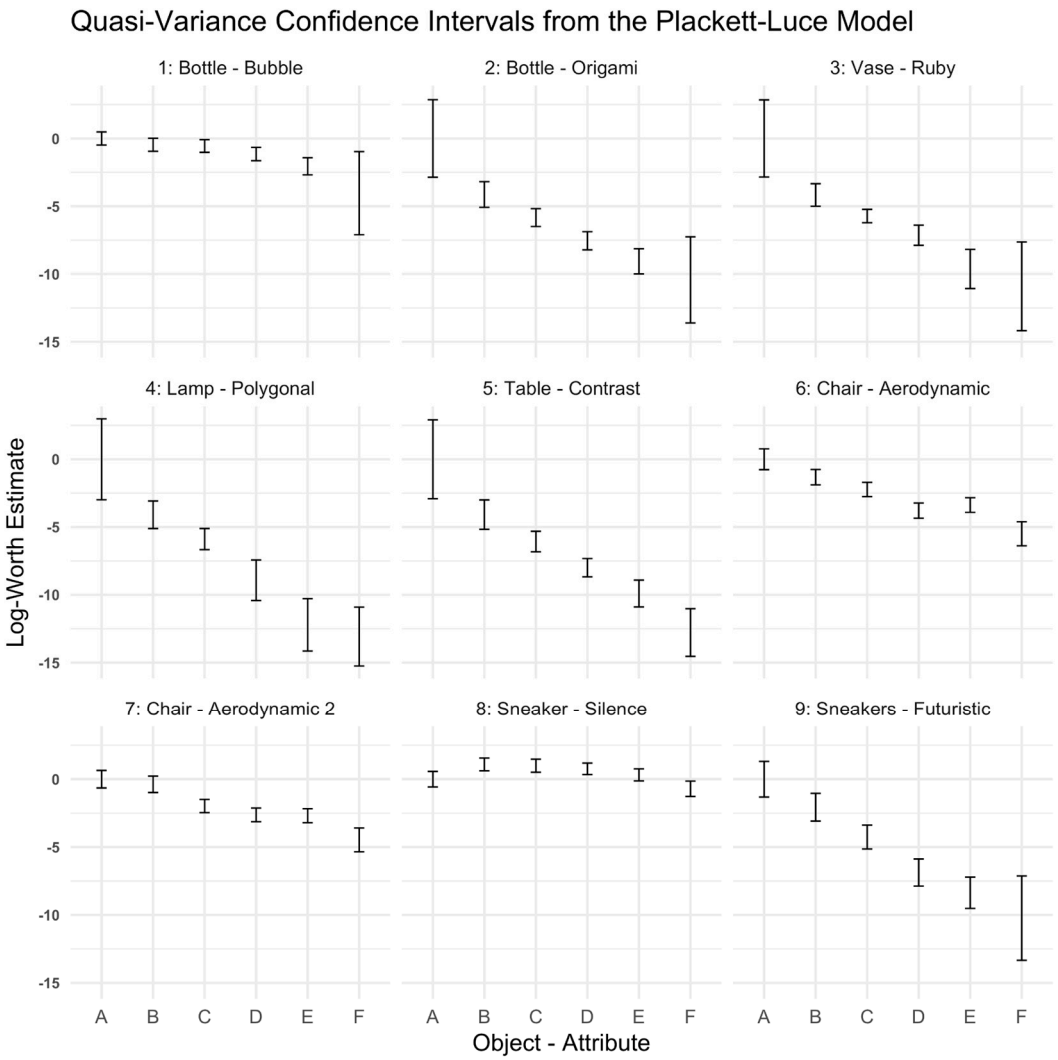

Figure 6. Plackett–Luce log-worth estimates with quasi-variance confidence intervals for each Product–Concept pair.

Figure 7 reports the number of NA responses per participant and Product-Concept pair. Each cell indicates how many items an individual observer judged as not recognisable within the category, and row/column totals summarise overall omission patterns. The distribution of NA responses across participants helps determine whether recognisability issues were concentrated among specific individuals or shared more broadly.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2026 doi:10.20944/preprints202603.0696.v1

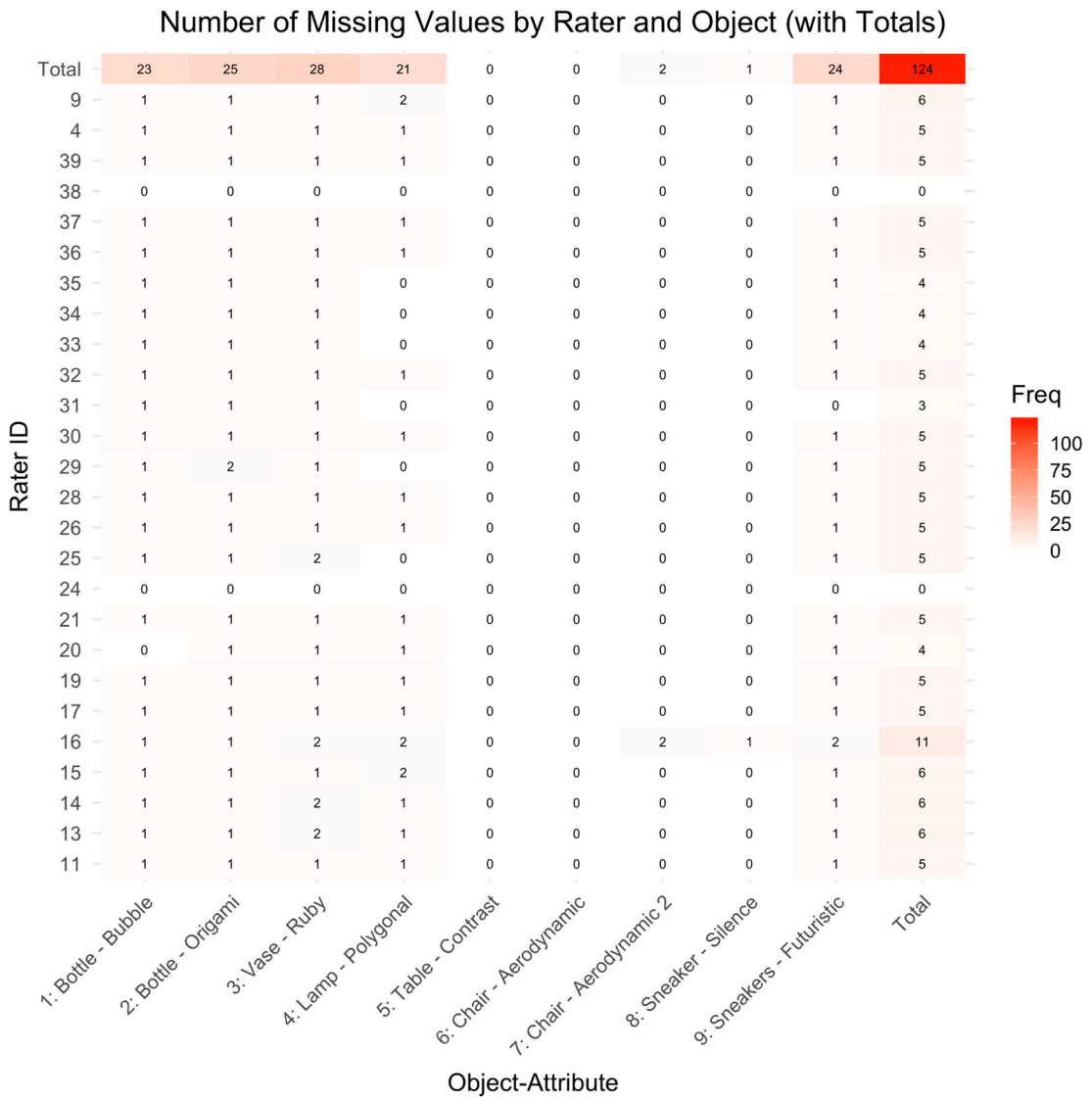

Figure 7. Number of NA responses per observer and Product–Concept pair, with row and column totals.

Figure 8 displays the number of NA responses per intended level across all participants. For each Product-Concept pair, the bars show how often items at levels A-F were excluded from the ranking due to lack of recognisability. Differences in NA frequency across levels reveal whether omissions cluster at particular points of the intended gradient.

 doi:10.20944/preprints202603.0696.v1

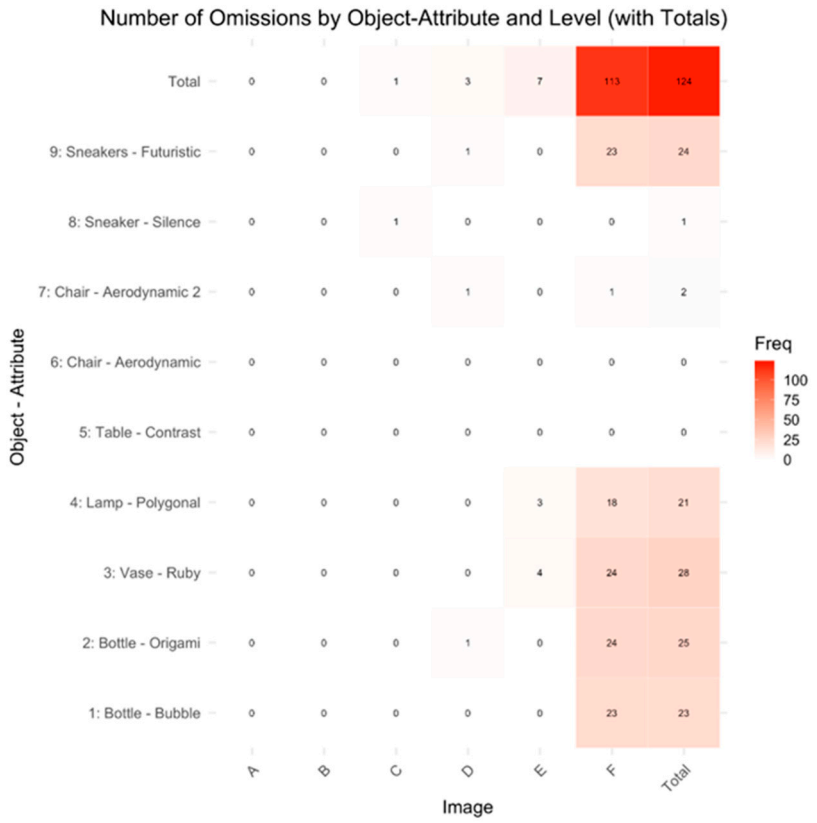

Figure 8. Number of NA responses per intended level (A–F) and Product–Concept pair, with totals.

Table 1 summarises the omission (NA) patterns observed across the nine Product–Concept pairs. Overall, the proportion of coincident NA decisions (Po) was high, indicating that observers largely agreed on which stimuli were no longer recognisable as instances of the target category. Agreement was maximal in Table–Contrast and Chair–Aerodynamic, where Po = 1.00 and no variability in NA responses was recorded. Consequently, Fleiss' κ, χ^2, and other variation-dependent statistics could not be computed for these pairs.

For Bottle–Bubble, Bottle–Origami, Vase–Ruby, Lamp–Polygonal, and Sneakers–Futuristic, agreement coefficients were also high, with Po values ranging from 0.93 to 0.98 and Gwet's AC1 estimates between 0.86 and 0.95. In these pairs, the χ^2 tests produced very low p-values (≈ 0.0002), and both the Cochran–Armitage and Jonckheere–Terpstra trend tests yielded similarly significant results, indicating that NA responses were not evenly distributed across levels A–F and instead showed clear level-dependent structure in their frequencies. The positive sign of the corresponding trend statistics reflects that NA frequencies increased across the intended A→F gradient—that is, higher levels exhibited more omissions, whereas lower levels exhibited fewer.

In contrast, Chair–Aerodynamic 2 and Sneaker–Silence exhibited low and scattered omission rates. This pattern is reflected in κ and α values close to zero or slightly negative, which arise from the minimal variability in NA decisions rather than from disagreement. Correspondingly, χ^2 and trend-test p-values were high (≥ 0.16), indicating that NA responses in these pairs were more uniformly distributed across levels and did not display evidence of a systematic trend.

Table 2 reports the reliability and validity metrics for the perceptual rankings across the nine Product–Concept pairs. Several pairs showed strong and coherent agreement among observers, with high Kendall's W, Krippendorff's α, and Fleiss' κ values—particularly Bottle–Origami, Vase–Ruby, Lamp–Polygonal, Table–Contrast, and Sneakers–Futuristic (W ≥ 0.89). These same pairs also exhibited perfect or near-perfect rank-based correlations with the intended A–F ordering (Spearman ρ = 1.00; Kendall τ = 1.00), indicating that participants consistently recovered the target monotonic progression.

Table 1. Cross-pair summary of NA-based operational-range constraints (nine Product–Concept pairs).

Pair	Po	Fleiss′	Kripp.	Gwet	χ^2 p-value	CA p-value	Z_CA	JT p-value	Z_JT
	κ	α	AC1						
Bottle – Bubble	0.9808	0.8593	0.8593	0.9527	0.00019996	7.44e−09	5.7805740	0.0002	1
Bottle – Origami	0.9808	0.8609	0.8609	0.9488	0.00019996	7.97e−10	6.1454293	0.0002	1
Vase – Ruby	0.9615	0.7632	0.7633	0.9011	0.00019996	5.13e−07	5.0215055	0.0002	1
Lamp – Polygonal	0.9295	0.5312	0.5313	0.8576	0.00019996	1.55e−04	3.7822112	0.0002	1
Table – Contrast	1.0000	NaN	1.0000	1.0000	NaN	1.0	0.0000000	1.0	1
Chair – Aerodynamic	1.0000	NaN	1.0000	1.0000	NaN	1.0	0.0000000	1.0	1
Chair – Aerodynamic 2	0.9872	–0.0130	–0.0127	0.9737	1.0	0.5771	0.5576410	0.1598	1
Sneaker – Silence	0.9936	–0.0065	–0.0062	0.9870	1.0	0.7777	-0.2822883	0.6684	1
Sneaker – Futuristic	0.9744	0.8148	0.8149	0.9348	0.00019996	3.57e−09	5.9030429	0.0002	1

Table 2. Cross-pair summary of agreement, monotonic alignment (nine Product–Concept pairs).

Pair	Kendall's W	Kripp. α	Fleiss κ	Spearman ρ (p)	Kendall τ (p)	Monotonicity	Diagnosis
Bottle–Bubble	0.247	0.268	0.095	1.000 (0.003)	1.000 (0.003)	yes	saturation with low agreement (category drift)
Bottle–Origami	0.914	0.910	0.700	1.000 (0.003)	1.000 (0.003)	yes	saturation / recognisability breakdown at E–F
Vase–Ruby	0.892	0.902	0.742	1.000 (0.003)	1.000 (0.003)	yes	saturation / recognisability breakdown at E–F
Lamp–Polygonal	0.976	0.977	0.858	1.000 (0.003)	1.000 (0.003)	yes	saturation / recognisability breakdown at E–F
Table–Contrast	0.924	0.921	0.865	1.000 (0.003)	1.000 (0.003)	yes	stable monotonic gradient
Chair–Aerodynamic	0.723	0.712	0.350	0.943 (0.017)	0.867 (0.017)	partial	mostly monotonic; minor deviations/ambiguity
Chair–Aerodynamic 2	0.591	0.580	0.271	1.000 (0.003)	1.000 (0.003)	yes	moderate agreement; local ambiguity
Sneaker–Silence	0.203	0.172	0.169	0.429 (0.419)	0.467 (0.272)	no	weak/non-monotonic concept axis
Sneakers–Futuristic	0.898	0.897	0.773	1.000 (0.003)	1.000 (0.003)	yes	saturation / recognisability breakdown at E–F

 doi:10.20944/preprints202603.0696.v1

Moderate levels of agreement were observed for Chair–Aerodynamic and Chair–Aerodynamic 2 ($0.59 \le W \le 0.72$), with Chair–Aerodynamic showing only partial alignment with the intended sequence ($\varrho = 0.94$; $\tau = 0.87$). In these cases, rankings were generally directionally correct but displayed local variability in the relative placement of adjacent levels.

For Bottle–Bubble and Sneaker–Silence, inter-rater agreement was too low to support a meaningful validity assessment. Both pairs showed very weak reliability, with Kendall's $W \le 0.25$, Krippendorff's $\alpha \le 0.27$, and Fleiss' $\kappa \le 0.17$. In these conditions, the dispersion and inconsistency of individual rankings prevent the interpretation of alignment metrics with the intended A–F order, and no conclusions regarding directional validity can be drawn for these pairs.

Across all Product–Concept pairs, the patterns observed in Tables 1 and 2 were consistent with the visual evidence provided by the level–position heatmaps and by the Plackett–Luce log-worth confidence intervals. Pairs with high agreement and clear directional validity showed tight, predominantly diagonal structures in the heatmaps and non-overlapping log-worth intervals across levels. Conversely, the pairs with low agreement (Bottle–Bubble and Sneaker–Silence) displayed highly dispersed heatmaps and extensive overlap in log-worth intervals, mirroring the absence of a stable perceptual structure in their rankings.

5. Discussion

5.1. From Statistical Outputs to Design-Relevant Regimes

The core question for engineering design is not whether a particular estimator (e.g., rank aggregation) behaves well, but whether a conceptual axis can be operationalised as a design variable that behaves predictably for human observers. Across our nine Product–Concept pairs we observe three recurrent regimes of perceptual behaviour for AI-generated gradients:

R1-Stable: The intended A→F progression is recovered with strong consensus and clear adjacent-level separation. Pairs such as Lamp-Polygonal, Table-Contrast or Sneakers-Futuristic exemplify this regime.

R2-Stable with weak segments: the global direction A→F holds, but one or two transitions show local ambiguity or partial inversions. Cases such as Chair-Aerodynamic or Chair-Aerodynamic 2 illustrate this behaviour.

R3-Unstable: no reliable ordering emerges; placements disperse and the intended axis is not read consistently. This pattern appears in Bottle-Bubble or Sneaker-Silence

These regimes are not threshold-based statistical classes, nor do they rely on fixed cut-off values on any single indicator. This is intentional: perceptual structure, reliability, and directional alignment cannot be reduced to a single coefficient, and no universal numerical threshold would be appropriate across product categories, concepts, or generative pipelines.

Instead, each regime is a qualitative pattern that emerges from the joint reading of multiple indicators—agreement (e.g., W, α), directional alignment (e.g., ϱ, τ), separability (PL intervals, pairwise probabilities), and recognisability (NA patterns). These indicators serve as diagnostic evidence, not fixed decision rules. Users of the method may adopt their own preferred statistical criteria, but the framework itself remains deliberately interpretative rather than prescriptive, similar to how p-values or effect sizes are interpreted in context rather than enforced by universal thresholds.

This qualitative structure is what enables the framework to generalise across diverse conditions without imposing arbitrary statistical thresholds, while still offering a reproducible and interpretable diagnostic logic.

Orthogonal to this, NA behaviour marks category recognisability and therefore the operational range of the gradient: NA absent or late (E–F) indicates a broad usable span; NA early (C–D) indicates that intensification quickly undermines category identity. The R-regime answers "is the gradient interpretable and how uniformly?"; the NA pattern answers "how far can intensification proceed before

 doi:10.20944/preprints202603.0696.v1

the design ceases to read as the same product?" Together, R1–R2–R3 × NA provides a diagnostic frame for deciding whether (and how) a concept axis is usable in early-stage, generative workflows.

These regimes emerge from comparing the monotonic generative axis with its perceptual counterpart. They do not demonstrate perceptual monotonicity; rather, they reveal the different ways in which a monotonic IA-driven manipulation can be interpreted—sometimes coherently, sometimes ambiguously, and sometimes not at all. We stress that these regimes are empirical regularities, not universal laws: they summarise how the combined system (pipeline + concept definition + product category) expresses a putative conceptual dimension as read by humans. They are therefore testable across additional concepts, categories, and generative controls.

The perceptual regimes observed in this study (R1, R2, R3 × NA) arise from the specific generative pipeline employed—an SDXL-based latent diffusion model implemented in ComfyUI, with fixed prompts, sampler configuration and seed. As with any generative workflow, alternative models, checkpoints, schedulers or seed strategies may produce gradients with different levels of monotonicity, separability or recognisability. This model-dependence is inherent to current generative-AI systems, and therefore the regimes identified here should be interpreted as characteristic of this pipeline rather than as universal properties of concept-driven image synthesis. Nonetheless, the diagnostic structure we propose—capturing stability, ambiguity and breakdown—is model-agnostic and is expected to remain applicable across evolving architectures. As generative technology continues to improve in prompt adherence and semantic controllability, it is reasonable to anticipate a shift toward more frequent R1 behaviours and fewer unstable cases (R3), although such advances will not eliminate the need for perceptual validation. The method is therefore designed to remain relevant regardless of how generative models progress.

5.1.1. Sources of Regime Behaviour

The three perceptual regimes identified in this study—R1 stable, R2 stable-with-weak-segments, and R3 unstable—should be understood as empirical patterns emerging from the interaction of four elements: the target concept, the product category, the generative model, and the observer population. These factors provide plausible, non-exhaustive explanations that help interpret the behaviours observed in our dataset, rather than offering a universal causal account.

R1 (stable gradients) typically arises when the target concept has shared semantic structure within the participant population and is represented consistently in the model's training distribution. In such cases, intensification reinforces perceptually aligned and category-compatible cues, producing smooth and discriminable transitions. Product categories with high structural tolerance to variation often support this behaviour, enabling conceptual modulation without compromising recognisability.

R2 (stable with weak segments) tends to occur when the concept is interpretable but internally multi-cue, leading the generative model to intensify different attributes at different points along the gradient. This can produce transitions that are globally coherent but locally ambiguous or weakly separable. Sensitive regions within a product category—where small formal changes have disproportionate perceptual effects—can further contribute to these local inconsistencies.

R3 (unstable gradients) is commonly observed when the concept lacks a shared perceptual interpretation, when it interacts poorly with the structural constraints of the product category, or when the generative model does not encode a coherent latent direction for that concept. Under these conditions, intensification may generate heterogeneous or non-monotonic variations, preventing observers from forming a reliable ordering. In such cases, the conceptual axis cannot be operationalised as a controllable design parameter.

Together, these factors clarify that regime behaviour is a diagnostic consequence of the combined system—concept, category, model and human interpretation—rather than a property of the validation method itself.

 doi:10.20944/preprints202603.0696.v1

5.2. Relation to Prior Work in Product Semantics and Kansei

The present findings align with long-standing perspectives in product semantics and Kansei Engineering, where product form communicates meaning through perceptual cues and cultural conventions. Traditional approaches—such as the semantic differential (Osgood et al., 1957) and Kansei Engineering (Schütte, 2005; Ishihara et al., 2008)—provide structured ways to measure affective impressions of fixed designs using rating scales. While these methods are highly effective for mapping connotative meaning, they do not assess whether controlled intensification along a single conceptual axis produces a perceptually coherent, monotonic gradient. Rating-based frameworks characterise impressions of individual artefacts but do not diagnose the interpretability, local separability, or breakdown of concept-driven variation.

Our contribution complements this tradition by shifting the analytical focus from evaluating impressions of discrete designs to assessing whether a generative manipulation behaves as a reliable design parameter. By incorporating local ambiguity and recognisability loss (NA) explicitly—phenomena that rating scales cannot capture—the proposed validation layer extends product-semantics and Kansei methodologies to the parametric, AI-driven manipulation of conceptual expression.

Whereas this situates our contribution within design-specific approaches to meaning and affect, the behaviour of the gradients also reflects more general perceptual regularities, which connects the present work to classical psychophysics.

5.3. Conceptual Link to Psychophysics: Gradients, Local Discriminability, and Breakdown

From a broader perceptual-science perspective, the gradient behaviours observed here parallel well-established properties of psychometric functions. Parametric manipulations typically yield monotonic but locally variable discriminability, with some transitions producing steep, easily perceptible differences and others yielding shallow or ambiguous perceptual changes (Klein, 2001). Under certain conditions—such as noisy cues or competing attributes—local non-monotonicities may also emerge (García-Pérez, 2014). This structure maps directly onto our regimes: R1 resembles a well-formed psychometric function with consistently high discriminability, R2 corresponds to locally shallow or weakly separable segments, and R3 reflects situations in which no coherent perceptual ordering arises from the manipulation.

Similarly, the emergence of NA responses mirrors the collapse of category recognition observed at extreme stimulus values in psychophysical tasks, when distortions or cue degradation prevent observers from maintaining category identity (Klein, 2001). Although we do not model psychometric functions explicitly, this analogy provides a principled foundation for interpreting stability, ambiguity, and breakdown in AI-generated conceptual gradients. It clarifies why certain transitions fail, why some segments remain ambiguous, and why intensification eventually undermines recognisability—insights with direct implications for how concept axes should be operationalised in generative workflows.

5.4. Actionable Guidance for Design Teams

The validation layer serves as an early decision point for determining whether a conceptual axis can function as a controllable parameter within a generative workflow. The combined pattern of perceptual regimes (R1–R2–R3) and NA responses translates directly into design guidance.

R1 (stable gradients) indicate that the intended progression is reliably interpreted; the concept can be used as a parameter, and the usable range extends up to the first NA level.

R2 (stable with weak segments) remain operational but require local adjustments—such as re-spacing levels or reinforcing the cues driving ambiguous transitions—before reuse. Only the problematic region needs re-generation and re-testing.

 doi:10.20944/preprints202603.0696.v1

R3 (unstable gradients) should not be treated as design parameters in their current form. These cases require concept reformulation, additional anchoring cues, or replacement by a more legible conceptual direction.

Across all regimes, NA responses mark the operational boundary of the axis: once recognisability collapses, further intensification becomes impractical regardless of directional clarity. Importantly, this framework evaluates interpretability and control, not aesthetic preference.

Finally, once a concept–category pair has been calibrated, the resulting knowledge can be reused: later iterations can rely on focused micro-panels targeting weak segments or boundary regions, reducing validation cost while maintaining perceptual reliability.

5.5. What This Is (and What It Is Not)

This work does not claim a universal "model of AI understanding". Rather, it proposes a diagnostic frame—R1–R2–R3 × NA—for characterising how AI-driven manipulations of product form manifest perceptually as interpreted by human observers. The novelty lies in:

Reconceiving the problem as engineerability of concept axes (legibility + local separability + operational range),

Showing regularities (the three regimes) that are testable beyond our nine cases, and

Making recognisability breakdown an explicit design constraint rather than an artefact to discard.

5.6. Limitations

A key limitation of the present study is its ecological validity. The perceptual judgements were made based the use of static, uniformly presented AI-generated images. While this controlled presentation isolates formal cues, it does not capture how products are interpreted in real contexts involving materiality, scale, motion, interaction, ambient lighting or social setting. As such, the perceptual gradients observed here should be understood as formal legibility under controlled visual conditions rather than as full accounts of in-situ interpretation.

The participant sample consisted exclusively of experienced design engineers, whose training enables fine-grained and calibrated readings of product-form cues. This expertise is appropriate for assessing the interpretability of concept gradients in a design workflow, but it may not represent how non-experts or general consumers would perceive the same variations. Consequently, the stability or instability observed for some gradients may differ in broader or more heterogeneous user populations.

Several of the concepts used in this study—such as "futuristic," "silence," or "origami"— culture-dependent interpretations. Although we selected terms that are widespread in design discourse and likely well represented in the training data of contemporary generative models, their perceptual meaning may vary across cultural contexts, user populations or design traditions. As such, the observed perceptual gradients should be interpreted within the cultural and disciplinary frame of the participant sample.

Although the nine Product–Concept pairs span heterogeneous categories and concept types, they do not exhaust the combinatorial space of possible semantic manipulations—and no finite sample could. The regimes observed in this study should therefore be interpreted as empirical regularities rather than universal patterns, with the understanding that additional categories, concepts and generative strategies may reveal further variants of gradient behaviour

5. Conclusions

This study introduced a ranking-based human validation layer designed to assess whether AI-generated conceptual gradients are perceptually coherent—that is, whether observers consistently interpret the intended direction of concept intensification across a series of product-form stimuli. The

 doi:10.20944/preprints202603.0696.v1

approach was evaluated through an online study with 26 qualified design engineers across nine Product–Concept pairs spanning diverse product categories and concept types.

Regarding RQ1, results showed that inter-observer reliability varied markedly across pairs. Five pairs—Bottle–Origami, Vase–Ruby, Lamp–Polygonal, Table–Contrast, and Sneakers–Futuristic—exhibited strong agreement among observers, while Chair–Aerodynamic and Chair–Aerodynamic 2 showed moderate consensus. Bottle–Bubble and Sneaker–Silence yielded insufficient agreement to support meaningful interpretation of the rankings.

Regarding RQ2, the pairs with strong reliability also displayed near-perfect monotonic alignment with the intended A–F progression, confirming that observers perceived the conceptual intensification in the direction defined by the generative pipeline. In moderate-agreement cases, the global direction held but local ambiguities emerged at specific transitions between adjacent levels. For low-agreement pairs, no directional conclusion could be drawn.

Regarding RQ3, NA responses proved to be systematically associated with higher intended levels in most pairs, marking the point at which conceptual intensification undermined product-category recognisability. Agreement on which items warranted exclusion was high, indicating that recognisability breakdown is not idiosyncratic but a shared perceptual boundary.

These findings converge into a diagnostic framework comprising three perceptual regimes—R1 (stable), R2 (stable with weak segments), and R3 (unstable)—combined with the NA-based identification of the operational range. Together, these dimensions provide design teams with actionable information: R1 gradients can be directly operationalised as design parameters; R2 gradients require local refinement at ambiguous transitions; and R3 gradients call for concept reformulation before reuse. In all cases, NA responses delimit the functional boundary beyond which further intensification becomes impractical.

The methodological contribution lies in demonstrating that a lightweight, reproducible protocol based on ranking with explicit recognisability marking and ordinal–probabilistic analysis can capture the perceptual structure of concept-driven generative manipulations—a capability not addressed by existing product-semantics, Kansei, or generative-AI evaluation frameworks. The conceptual contribution resides in the regime taxonomy itself, which offers a transferable vocabulary for characterising gradient behaviour across diverse concept–category combinations. The operational contribution is the reconceptualisation of recognisability loss as a design-relevant constraint rather than noise, enabling explicit identification of the usable span of a conceptual axis.

By shifting the focus from output volume to interpretable structure, this work positions human perceptual validation as an integral governance mechanism within AI-driven generative design workflows, bridging the traditions of product semantics and Kansei Engineering with the parametric control demands of contemporary generative systems.

These conclusions should be interpreted in light of the study's scope. The participant sample was culturally and linguistically homogeneous, the stimuli were static AI-generated images presented under controlled visual conditions, and the nine Product–Concept pairs, while heterogeneous, represent a finite subset of the possible concept–category space. Future work should explore the generalisability of the regime framework across broader and more diverse observer populations—including non-expert users and cross-cultural samples—as well as under richer stimulus conditions involving materiality, scale, or interactive presentation. Extending the validation protocol to alternative generative architectures and to concepts drawn from other design traditions would further test the robustness of the R1–R2–R3 taxonomy. Finally, longitudinal studies examining how calibrated gradients perform across iterative design cycles, and whether micro-panel strategies can effectively reduce validation cost without compromising diagnostic accuracy, constitute promising directions for integrating this approach into routine generative design practice.

 doi:10.20944/preprints202603.0696.v1

References

Amershi, S., Weld, D. S., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human–AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3290605.3300233

Amiri Manesh, S. (2025). Quality assessment of generative AI-based compression. University of Padua. Retrieved from https://thesis.unipd.it/retrieve/c2f0bb2f-39ef-4247-a4a3-1e1c03f212fa/Amirimanesh_Sana.pdf

Anandan, S., Maier, J. R. A., Bapat, V., & Bettig, B. (2007). Design semantics: A clear definition based on a knowledge view. Retrieved from https://www.researchgate.net/publication/255568485_Design_Semantics_A_Clear_Definition_Based_on_a_Knowledge_View

Arlitt, R. M., & Van Bossuyt, D. L. (2019). A generative human-in-the-loop approach for conceptual design exploration using flow failure frequency in functional models. Journal of Computing and Information Science in Engineering, 19(3), 031001. https://doi.org/10.1115/1.4042913

Arlitt, R., & Van Bossuyt, D. (2022). Human–AI collaboration in conceptual design: Opportunities and challenges. Design Studies, 78, 101–118.

Chen, L., Jing, Q., Tsang, Y., Wang, Q., Liu, R., & Xia, D. (2024). AutoSpark: Supporting automobile appearance design ideation with Kansei engineering and generative AI. ACM Digital Library.

Chen, L., Song, Y., Guo, J., Sun, L., Childs, P., & Yin, Y. (2025). How generative AI supports humans in conceptual design. Design Science. https://doi.org/10.1017/dsj.2025.2

Chen, Y., Tan, J., Zhang, A., & Yang, Z. (2024). On softmax direct preference optimization for recommendation. NeurIPS 2024 Proceedings. https://proceedings.neurips.cc/paper_files/paper/2024/file/30732ddb12d9faf7180f5d0e8b5b5da7-Paper-Conference.pdf

Comte, J. (2022). AI vs. Human in the Loop: Prototype designs and expert evaluation of semiautomated consistency checking. Retrieved from https://www.diva-portal.org/smash/get/diva2:1696863/FULLTEXT01.pdf

Demirel, H. O., Goldstein, M. H., & Li, X. (2024). Human-centered generative design framework: An early-design framework to support concept creation and evaluation. International Journal of Human–Computer Interaction, 40(7), 657–674. https://doi.org/10.1080/10447318.2023.2171489

Fattah Saleh, H. (2025). Enhancing the industrial design process with generative AI. Theseus Repository.

Feng, X., Du, H., Ma, J., Wang, H., & Zhou, L. (2025). Crafting user-centric prompts for UI generations based on Kansei engineering and knowledge graph. Advanced Engineering Informatics. ScienceDirect.

García-Pérez, M.A.(2014). Adaptive psychophysical methods for nonmonotonic psychometric functions. Atten Percept Psychophys 76, 621–641. https://doi.org/10.3758/s13414-013-0574-2

Georgiev, G. V., & Georgiev, D. D. (2020). Semantic analysis of engineering design conversations. Proceedings of the Design Society: Design Conference. https://doi.org/10.1017/dsd.2020.294

Gorantla, S., Bhansali, E., & Deshpande, A. (2023). Optimizing group-fair Plackett–Luce ranking models for relevance and ex-post fairness. arXiv preprint. https://arxiv.org/abs/2308.13242

Guo, F., Qu, Q. X., Nagamachi, M., & Duffy, V. G. (2020). A proposal of the event-related potential method to effectively identify Kansei words for assessing product design features in Kansei engineering research. International Journal of Industrial Ergonomics, 75, 102891. https://doi.org/10.1016/j.ergon.2020.102940

Han, J., Sarica, S., Shi, F., & Luo, J. (2022). Semantic networks for engineering design: State of the art and future directions. Journal of Mechanical Design, 144(2), 020802. https://doi.org/10.1115/1.4052148

Ishihara, S., Nagamachi, M., Schütte, S., & Eklund, J. (2008). Affective meaning: The Kansei engineering approach. In Advances in Applied Ergonomics (pp. 523–530). https://doi.org/10.1016/B978-008045089-6.50023-X

Jin, J. (2013). Information mining from online reviews for product design. Hong Kong Polytechnic University. Retrieved from https://theses.lib.polyu.edu.hk/handle/200/7198

Kazmierczak, E. T. (2003). Design as meaning making: From making things to the design of thinking. Design Issues, 19(2), 45–59. https://doi.org/10.1162/074793603765201406

 doi:10.20944/preprints202603.0696.v1

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. Perception & Psychophysics 63, 1421–1455. https://doi.org/10.3758/BF03194552

Krippendorff, K. (2006). The semantic turn: A new foundation for design. CRC Press. https://doi.org/10.4324/9780203299951

Li, J., Cao, H., Lin, L., Hou, Y., Zhu, R., & El Ali, A. (2024). User experience design professionals' perceptions of generative artificial intelligence. CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3613904.3642114

Liang, J. (2024). The application of artificial intelligence-assisted technology in cultural and creative product design. Scientific Reports, 14(2). Nature.

Lopez, C. E., Zhao, Z. V., & Tucker, C. S. (2019). Semantic network differences across engineering design communication methods. ASME IDETC-CIE 2019. Retrieved from https://sites.lafayette.edu/lopezbec/files/2019/12/Lopez_et_al_IDETC_DRAFT-1.pdf

Luce, R. D. (1959). Individual choice behavior: A theoretical analysis. Wiley.

Luo, J. (2025). Development of an artificial-intelligence-driven product design evaluation model using multi-modal data. PolyU Theses Repository.

Ma, J., Yi, X., Tang, W., Zhao, Z., & Hong, L. (2021). Learning-to-rank with partitioned preference: Fast estimation for the Plackett–Luce model. Proceedings of Machine Learning Research, 130, 1–15. http://proceedings.mlr.press/v130/ma21a/ma21a.pdf

Marrone, A. (2023). Optimizing product development and innovation processes with artificial intelligence. Polito Webthesis.

Oosterhuis, H. (2021). Computationally efficient optimization of Plackett–Luce ranking models for relevance and fairness. Proceedings of the 44th ACM SIGIR Conference on Research and Development in Information Retrieval. https://doi.org/10.1145/3404835.3462830

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. University of Illinois Press.

Pearce, M., & Erosheva, E. A. (2022). A unified statistical learning model for rankings and scores with application to grant panel review. Journal of Machine Learning Research, 23, 1–40. Retrieved from http://www.jmlr.org/papers/v23/21-1262.html

Perez-Ortiz, M., & Mantiuk, R. K. (2017). A practical guide and software for analysing pairwise comparison experiments. arXiv preprint. https://arxiv.org/abs/1712.03686

Plackett, R. L. (1975). The analysis of permutations. Applied Statistics, 24(2), 193–202.

Rosen, B. G., Eriksson, L., & Bergman, M. (2016). Kansei, surfaces and perception engineering. Surface Topography: Metrology and Properties, 4(3), 033001. https://doi.org/10.1088/2051-672X/4/3/033001

Schäfer, D., & Hüllermeier, E. (2018). Dyad ranking using Plackett–Luce models based on joint feature representations. Machine Learning, 107(6), 943–978. https://doi.org/10.1007/s10994-017-5694-9

Schütte, S. (2005). Engineering emotional values in product design—Kansei engineering in development. Linköping University. Retrieved from https://www.diva-portal.org/smash/get/diva2:20839/FULLTEXT01.pdf

Schütte, S., Eklund, J., & Axelsson, J. R. C. (2004). Concepts, methods and tools in Kansei engineering. Theoretical Issues in Ergonomics Science, 5(3), 214–231. https://doi.org/10.1080/1463922021000049980

Sciandra, A., Dufour, F., & Caruso, G. (2021). Ordinal scaling in subjective image quality assessment. Frontiers in Psychology, 12, 715042.

Sciandra, M., Fasola, S., Albano, A., & Di Maria, C. (2024). Discrete Beta and Shifted Beta–Binomial models for rating and ranking data. Statistical Papers. https://doi.org/10.1007/s10651-023-00592-5

Siddharth, L., Blessing, L., & Luo, J. (2022). Natural language processing in-and-for design research. Design Science, 8, e16.

Tang, Y., Zhang, R., Guo, J., De Rijke, M., & Chen, W. (2024). Listwise generative retrieval models via a sequential learning process. ACM Transactions on Information Systems. https://doi.org/10.1145/3653712

Turner, H. L., van Etten, J., Firth, D., & Kosmidis, I. (2020). Modelling rankings in R: The PlackettLuce package. Computational Statistics, 35(3), 1027–1054. https://doi.org/10.1007/s00180-020-00959-3

 doi:10.20944/preprints202603.0696.v1

Verma, D. (2025). Is generative AI a successor to human-in-the-loop perception and cognition experiments in urban design and planning? Journal of Urban Design, 30(2). https://doi.org/10.1080/13574809.2025.2514574

Wang, P., Peng, D., Li, L., Chen, L., & Wu, C. (2019). Human-in-the-loop design with machine learning. Proceedings of the International Conference on Engineering Design (ICED19). https://doi.org/10.1017/dsi.2019.264

Wang, Y., Zhang, J., Shen, C., Yu, H., & Luo, S. (2024). Generative AI aids product design iteration framework based on personalized style and aesthetic preferences. SSRN.

Zhang, D. C., & Lauw, H. W. (2021). Topic modeling for multi-aspect listwise comparisons. Proceedings of the 30th ACM International Conference on Information and Knowledge Management. https://doi.org/10.1145/3459637.3482398