Towards Perceptually Grounded Dissimilarity Measures for Computational Design

Ralf P.W. Schmidt[*,†], Rianne Conijn[†], Hèrm Hofmeyer[†] and Pieter Pauwels[†]

Eindhoven University of Technology, De Zaale, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands

Abstract

Computational measures of dissimilarity are increasingly used in generative design and creativity support systems to compare and explore design alternatives. However, dissimilarity in creative contexts is not purely physical or statistical but also perceptual and interpretive, which raises questions about whether computational measures capture the same distinctions humans do. Therefore, understanding the relationship between computational and perceived measures of dissimilarity is important for interpreting how algorithmic metrics relate to human creative judgment. In this work, we conducted an experiment in which participants judged the dissimilarity of 3D geometries and compared these judgments with values produced by a computational measure. Our findings indicate limited alignment between computational and perceptual dissimilarity. However, the computational measure significantly predicted variability in human judgments, suggesting that it captures perceptual salience or agreement rather than the perceived magnitude of dissimilarity. These results highlight the need for perceptually grounded metrics in generative design systems.

Keywords

Computational Dissimilarity, Generative Design, Perception, Evaluation

1. Introduction

Creativity is recognized as a cornerstone of innovation in design, science, and the arts. Yet, it remains difficult to define and measure. A common consensus describes creativity as the production of ideas or artifacts that are both novel and appropriate [1, 2]. Other work by Boden [3] poses that for work to be considered creative, it must be original, surprising, and useful. While these definitions provide some clarity of creativity as a construct, operationalization in empirical research still poses significant challenges.

Creativity is context-dependent and multidimensional, making direct evaluation difficult [4, 5]. While creativity cannot be reduced to a single perceptual or evaluative process, empirical and computational approaches may rely on measurable proxies to approximate aspects of human creative assessment. One widely used proxy for creativity is divergent thinking performance, typically measured by fluency, originality, flexibility, or elaboration scores [6, 7]. Although divergent thinking tasks have some predictive validity regarding creativity, they capture only a subset of creative cognition and are sensitive to scoring procedures and task framing [8, 9].

In computational creativity and generative design, similar proxy-based methods are often used to operationalize concepts such as novelty, diversity, and surprise, which are typically understood in relational terms. For example, an artifact is considered novel or surprising because it differs from prior examples, expectations, or alternatives in a design space [3, 10, 11]. As a result, computational design systems rely on measures of similarity or dissimilarity to estimate how distinct a generated artifact is from others, using these as proxies for creative variation or exploration.

WCDCC 2026: Second Workshop on Computational Design and Computer-aided Creativity 2026, 29 June, 2026, Coimbra, PT

[*]Corresponding author.

[†]These authors contributed equally.

r.p.w.schmidt@tue.nl (R. P.W. Schmidt); m.a.conijn@tue.nl (R. Conijn); h.hofmeyer@tue.nl (H. Hofmeyer); p.pauwels@tue.nl (P. Pauwels)

0000-0003-2667-2473 (R. P.W. Schmidt); 0000-0002-6316-4892 (R. Conijn); 0000-0001-8353-7054 (H. Hofmeyer); 0000-0001-8020-4609 (P. Pauwels)

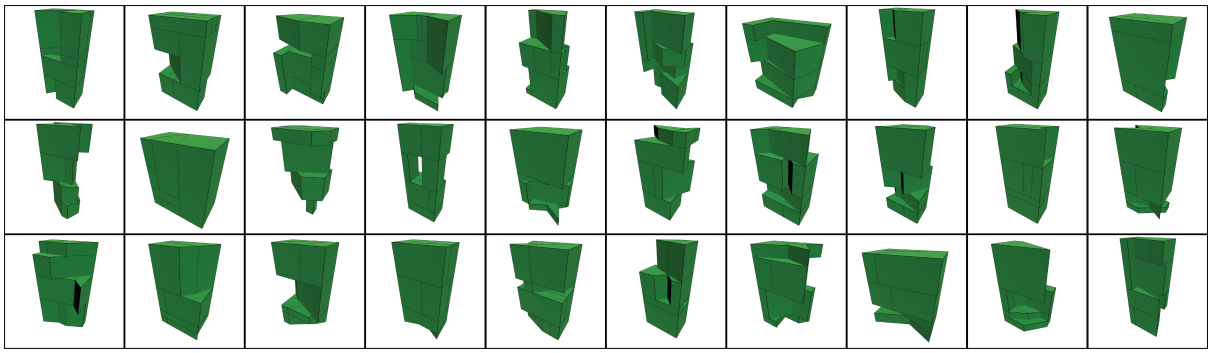

Figure 1: Overview of the 30 individual Building Spatial Designs (BSDs) adopted from Pereverdieva [16]. These designs served as the base set from which all stimulus pairs shown to participants during the experiment were constructed.

Computational design systems, including design process simulations, generative design tools, and Artificial Intelligence (AI) systems, can now produce and transform large numbers of design alternatives, making manual comparison and evaluation impractical. As a result, computational measures are increasingly used because they are scalable, reproducible, and efficient. For example, similarity and alignment metrics such as CLIPScore [12] quantify semantic alignment between images and text, while structure-based or domain-specific measures compute geometric or feature-based distances between design artifacts. However, such metrics are ultimately abstractions of similarity or difference, typically derived from feature spaces, embeddings, or algorithmic distance functions. Empirical validation against human perceptual judgments is therefore essential for safeguarding that computational proxies function as perceptually meaningful indicators rather than purely mathematical abstractions.

An open question, however, is whether such computational measures meaningfully correspond to human perceptual judgments. Human evaluations of similarity and difference are not purely statistical, but are shaped by salience, interpretation, structure, and context [13, 14]. Research in cognitive psychology demonstrates that similarity judgments can be asymmetric and are even sensitive to phrasing or task framing [14]. For example, framing a task as "similarity" versus "dissimilarity," or asking whether object A is different from object B versus asking whether object B is different from object A, can yield different results even when the compared stimuli are identical. This suggests that similarity and dissimilarity judgments can involve different cognitive processes and may not map linearly onto one another [14, 15]. The question is how well these sensitivities in subjective evaluations can be captured in computational metrics. If generative systems rely on computational dissimilarity measures that diverge from human perception, they may optimize for mathematically distinct outputs that are not perceived as meaningfully different by users.

In this work, we investigate the extent to which computational dissimilarity scores of 3D geometries correlate with human perceptual similarity and dissimilarity judgments. In doing so, we seek to inform the design and validation of computational metrics used in design support systems, including design process simulations and creative AI systems, to align algorithmic measures with human creative perception.

2. Method

2.1. Design & Participants

A mixed-design online study was conducted in which each participant evaluated the perceived dissimilarity of pairs of 3D geometries. The stimulus set consisted of 100 pairs of Building Spatial Designs (BSD pairs), sampled from all possible pairwise combinations of 30 individual BSDs taken from Pereverdieva [16]. The BSDs were all of comparable complexity, containing identical numbers of spaces and layers, thereby limiting variability arising from differences in overall design complexity. For each BSD pair,

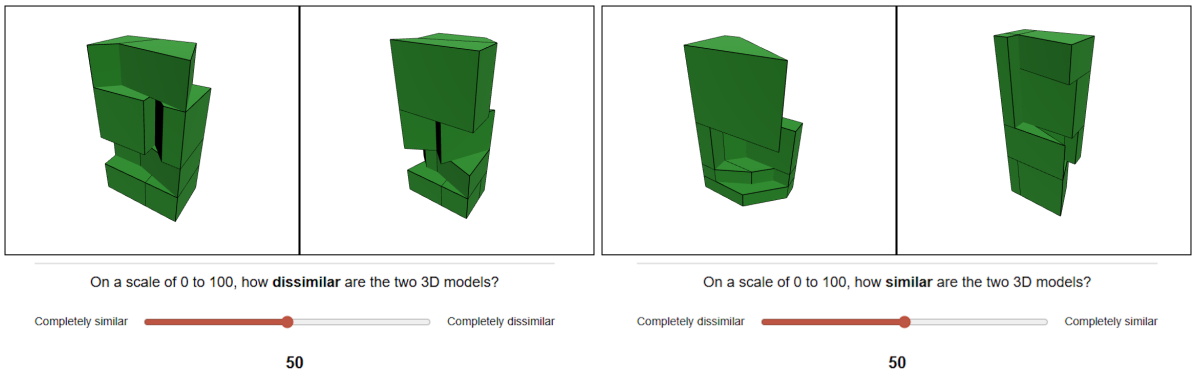

Figure 2: Two example trials shown during the online study. Left: dissimilarity condition. Right: similarity condition.

dissimilarity was precomputed using a dissimilarity measure adapted from Pereverdieva et al. [17] (Figure 1).

Given that human perceptions of similarity are already sensitive to task phrasing [14], two conditions were used: participants were randomly assigned to either the *dissimilarity* or the *similarity* condition, in which participants were asked to judge the perceived dissimilarity or similarity of the pairs. Then, to support the estimation of inter-rater reliability and stable variance component estimation in mixed-effects models [18, 19, 20], approximately 20 ratings per BSD pair across two conditions were needed, which corresponded to 4000 observations. To limit participant fatigue during the online experiment, each participant rated 40 randomly selected BSD pairs. With this setup, a minimum of 100 participants was required to reach the target number of observations. In total, 115 participants were recruited through the Prolific platform, of which 14 participants were excluded due to low data quality. The final sample consisted of 101 participants (48 female, 53 male), between 19 and 60 years old (M = 31.56, SD = 8.09). Ethical approval for conducting this study was obtained from the Ethical Review Board of Eindhoven University of Technology, reference ERB-722, prior to data collection.

2.2. Procedure

Prior to starting the online study, participants were informed about the aim of the study. After providing consent to their participation, participants performed a trial run of rating one BSD pair. Then, each participant evaluated a random subset of 40 out of the 100 BSD pairs on perceived dissimilarity or similarity (depending on the condition) using a slider ranging from 0 to 100, in random order (Figure 2). The experiment was set up in such a way that participants had to rotate both 3D models before continuing to the next set of stimuli. To not influence the participants, they were not informed on how to evaluate (dis)similarity, which would defeat the purpose of this study. After rating 40 BSD pairs, participants were asked which features they took into account or relied on when evaluating (dis)similarity. Finally, participants were asked demographic questions about age and gender. The online study was conducted using the LabJS framework [21] and took approximately 25 minutes to complete. Participants received monetary compensation for their participation after the study concluded.

2.3. Measurements

The computational dissimilarity values were calculated for each BSD pair using a dissimilarity measure proposed by Pereverdieva et al. [17] (Equation 1), which consists of topological dissimilarity d_{GED} and two components of geometric dissimilarity using volumetric overlap d_{VOL} and distance between matching spaces d_{COM}:

$$d(G_1, G_2) = \frac{w_1 d_{GED}(G_1, G_2) + w_2 d_{vol}(G_1, G_2) + w_3 d_{com}(G_1, G_2)}{w_1 + w_2 + w_3} \quad (1)$$

where G_1 and G_2 denote two BSDs with w_1, w_2, $w_3 \geq 1$, ensuring that $d(G_1, G_2) \in [0, 1]$.

The first component, d_{GED} (Equation 2), is calculated using the concepts of edit distance and adjacency graphs via adjacent spaces, using a Graph Edit Distance (GED) implementation following Abu-Aisheh et al. [22]. Each 3D building geometry is represented as a graph $G = (V, E)$, where each node $v \in V$ maps to a space within the 3D geometry, and an edge $e \in E$ exists if and only if two spaces are adjacent (i.e., at least share one point in common). The GED algorithm finds an optimal edit path $\phi \in \Phi$ to transform G_1 into a graph isomorphic to G_2 using a minimal set of node and edge edit operations (insertion, deletion, substitution), yielding a mapping $f : V_1 \rightarrow V_2$. The topological dissimilarity is defined as:

$$d_{GED}(G_1, G_2) = \min_{\phi \in \Phi} \frac{\sum_{e \in \phi} c(e)}{|V_1| + |E_1|} \tag{2}$$

where ϕ denotes an edit path, Φ the set of all possible edit paths, $c(e)$ the cost of edit operation e, and $|V_1| + |E_1|$ the total number of elements in G_1, normalizing the result to [0, 1].

The second component, d_{VOL} (Equation 3), captures the geometric dissimilarity between matching spaces by considering their volumetric overlap. Using the previously obtained mapping $f : V1 \rightarrow V2$, each space $v \in V_1$ is matched to a corresponding space $f(v) \in V_2$. To compare the geometry of a matched pair, the centers of mass of v and $f(v)$ are first aligned. Next, space v is transformed into $4r$ alternative orientations $v_1, v_2, ..., v_{4r}$ by applying r rotations and mirroring over the x- and y-axes. For each transformed orientation, geometric overlap is quantified using the Jaccard distance [23] (Equation 4), which quantifies overlap for optimally aligned spaces. The geometric dissimilarity (bounded between 0 and 1) based on volumetric overlap is then:

$$d_{vol}(G_1, G_2) = \frac{1}{|V_1|} \sum_{v \in V_1} \min_{k \in \{1,...,4r\}} d_J(v_k, f(v)) \tag{3}$$

where the minimum operator selects the orientation that yields the greatest volumetric overlap between a space and its matched counterpart. For each transformed orientation v_k, the Jaccard distance with $f(v)$ is computed as:

$$d_J(v_k, f(v)) = \frac{|v_k \cup f(v)| - |v_k \cap f(v)|}{|v_k \cup f(v)|} \tag{4}$$

where $|v_k \cap f(v)|$ denotes the intersection volume shared by both spaces and $|v_k \cup f(v)|$ denotes their union volume.

The third component, d_{COM} (Equation 5), captures geometric dissimilarity arising from differences in the spatial location of matching spaces within the two 3D geometries. Rather than comparing different spaces that happen to share similar locations, the measure evaluates the average Euclidean distance between the centers of mass of all topologically matched spaces identified by the graph correspondence mapping. The geometric dissimilarity based on center-of-mass distance is then:

$$d_{com}(G_1, G_2) = \frac{1}{|V_1|} \sum_{i \in V_1} \frac{\|com(i) - com(f(i)\|_2}{D} \tag{5}$$

where $\|com(i) - com(f(i))\|_2$ denotes the Euclidean distance between the centers of mass of matched space i and $f(i)$, and D the largest bounding box diagonal of either 3D geometry. The result is bounded between 0 and 1.

Prior to analysis, several bugs identified in the original Python implementation of the dissimilarity measure were corrected. These included a failure to iterate over all spaces when computing the volumetric dissimilarity component (Equation 3), and the use of a fixed rather than pair-specific normalization value D in the distance component (Equation 5). The corrected implementation was rewritten in C++ and used to compute all dissimilarity values reported here.

Due to the computational dissimilarity values observed in the stimulus set being concentrated within a restricted range (min = 0.237, max = 0.516), binning across the full $[0 - 1]$ interval did not yield balanced pair counts. To ensure meaningful coverage of the observed dissimilarity spectrum, pairs

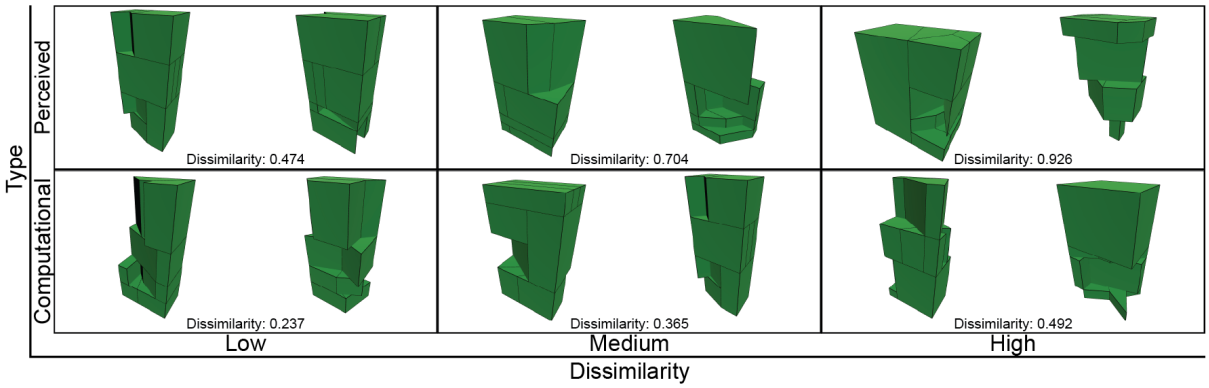

Figure 3: Example pairs of Building Spatial Designs (BSDs). The top row shows perceived dissimilarity, and the bottom row shows computational dissimilarity. From left to right: low, medium, and high dissimilarity.

were smartly chosen to ensure a wide range of computed dissimilarity values. Specifically, 15 pairs were sampled from the lowest 10% and highest 10% of dissimilarity values, and 70 pairs were sampled across intermediate percentiles, resulting in a total of 100 BSD pairs for human evaluation.

3. Results

3.1. Participant Agreement in Perceptual Judgment

To assess the degree of participant agreement, inter-rater reliability was assessed using a variance component approach, which fits the incomplete rating design in which each participant evaluated a subset of 40 out of 100 BSD pairs. Intraclass correlation coefficients (ICCs) at the stimulus-pair level were low in both conditions ($ICC_{similar}$ = 0.19; $ICC_{dissimilar}$ = 0.09), indicating limited agreement among participants in how they ranked the relative (dis)similarity of specific BSD pairs. In this analysis, the ICC reflects the extent to which different participants provide consistent ratings for the same stimulus pair, i.e., whether certain pairs are reliably judged as more or less (dis)similar across observers. The combined ratings showed a slight increase (ICC = 0.14) compared to the *dissimilarity* condition but remained low overall. These results suggest that (dis)similarity judgments were highly subjective, with little consensus across participants.

In contrast, ICCs computed at the participant level were higher ($ICC_{similar}$ = 0.25; $ICC_{dissimilar}$ = 0.38; $ICC_{combined}$ = 0.32). Here, the ICC reflects the consistency of individual participants' response tendencies across stimulus pairs, i.e., whether some participants systematically use the rating scale differently regardless of the specific stimulus pair. The fact that participant-level variance exceeded stimulus-pair-level variance suggests that systematic individual differences in rating style contributed more to the observed variability than the differences between BSD pairs themselves.

Taken together, these results indicate that perceived (dis)similarity is not consistently shared across participants but varies substantially between individuals. This implies that human judgments of dissimilarity cannot be treated as a single, stable ground truth.

3.2. Alignment Between Computational and Perceived Dissimilarity

As a validity check, the relationship between mean pair-level ratings in the *dissimilar* and *similar* conditions was examined. A strong negative correlation was found between the two conditions (ρ = -0.85, $p < .001$), indicating that pairs judged as more dissimilar in the *dissimilar* condition tended to be judged as less similar in the *similar* condition. This suggests that participants in both conditions were sensitive to the same underlying perceptual differences between BSD pairs, and that ratings could be analyzed separately as well as combined. Figure 3 shows six examples of BSD pairs that were computed or rated as low, medium, and highly dissimilar.

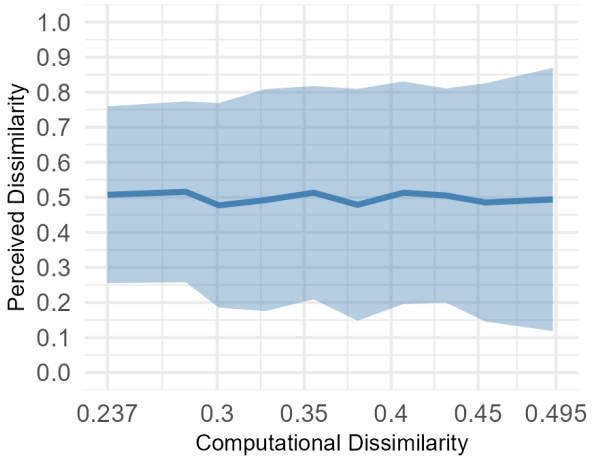

Figure 4: Mean perceived dissimilarity as a function of computational dissimilarity. The shaded area represents the variability in human ratings across participants. The mean rating remains largely stable, while the spread of ratings increases at higher computational dissimilarity values.

The relationship between the computational dissimilarity values and perceived dissimilarity ratings was then examined using Spearman rank correlation coefficients with bootstrapped 95% confidence intervals. Effect sizes were interpreted following conventional guidelines [24]. Results indicated small, statistically nonsignificant correlations between computational and perceived dissimilarity values across both conditions. In the *similar* condition, a small negative correlation was observed (ρ = -0.119, 95% CI [-0.142, 0.137], p = .237), indicating that higher computational dissimilarity values were weakly associated with *lower* similarity ratings, consistent with the expected direction. In the *dissimilar* condition, a small positive correlation was found (ρ = 0.123, 95% CI [-0.146, 0.140], p = .223), again indicating weak alignment. Given the strong inverse relationship between similarity and dissimilarity ratings, similarity ratings were inverted and combined with the dissimilarity ratings. The resulting correlation with computational dissimilarity remained small and non-significant (ρ = 0.121, 95% CI [-0.024, 0.256], p = .087).

To further examine this relationship while accounting for the hierarchical structure of the data, a mixed-effects beta regression was conducted. Perceived dissimilarity values were first transformed to an open interval (0, 1) according to Smithson and Verkuilen [25]. Beta regression is appropriate for modeling continuous outcome variables bounded between 0 and 1. Consistent with the Spearman analyses, computational dissimilarity was not a significant predictor of mean perceived dissimilarity (β = 1.46, SE = 1.085, z = 1.35, p = .179), suggesting limited alignment between the computational metric and the central tendency of human judgments. However, computational dissimilarity significantly predicted the dispersion of perceptual ratings (β = -1.59, SE = 0.433, z = -3.67, p < .001). In beta regression, the dispersion parameter is inversely related to variance; therefore, the negative coefficient indicates that variability in perceptual ratings *increased* with higher computational dissimilarity values (Figure 4). In other words, although the computational metric did not predict how dissimilar participants perceived BSD pairs to be on average, it was associated with how consistently those judgments were made.

Across participants, several recurring strategies for judging (dis)similarity emerged from the qualitative responses, although no single strategy was reported universally by all participants. First, multiple participants indicated relying on the *overall global shape* of the BSDs, suggesting an emphasis on holistic comparisons. Second, several responses highlighted the importance of *size and spatial proportions*, including height, width, and volume, indicating that metric properties were used as key heuristics in similarity judgments. Finally, some participants reported attending to *distinct geometric features* such as “holes”, “cavities”, and “cut-outs”, which were frequently described as particularly diagnostic for distinguishing between the 3D geometries.

Taken together, these findings indicate that the computational dissimilarity metric accounts for only

a small proportion of variance in perceptual judgments, raising questions about the extent to which it captures perceptually relevant differences between the 3D geometries. In addition, participants indicated that (dis)similarity judgments were primarily driven by a combination of global shape perception, metric scaling, and distinct features, which are elements currently not implemented in the computational dissimilarity measure used in this work. Nevertheless, the association with rating dispersion suggests that while the computational metric does not reliably predict perceived dissimilarity itself, it appears to be associated with the degree of perceptual consensus across observers.

4. Discussion

This study examined the extent to which a computational dissimilarity measure aligns with human perceptual judgments of similarity and dissimilarity in 3D geometries. Overall, the results indicate limited alignment between computational and human judgments, while revealing an interesting relationship between computational metrics and perceptual agreement.

A central finding is that correlations between computational and perceived dissimilarity were consistently small and non-significant across conditions. This suggests that the computational measure explains only a limited proportion of variance in perceived dissimilarity, indicating that it does not quite capture the perceptual dimensions that the participants relied on when evaluating the 3D geometries. This finding aligns with insights from cognitive psychology that similarity judgments are not purely determined by physical or statistical features, but are shaped by salience, interpretation, and contextual framing [13, 14]. From this perspective, computational measures that operate on fixed feature representations or embedding spaces may fail to reflect the flexible and context-dependent nature of human judgment.

At the same time, the mixed-effects beta regression revealed a different pattern. While computational dissimilarity did not significantly predict perceived dissimilarity, it significantly predicted the *dispersion* of ratings. Specifically, the variability in human judgments increased when computational dissimilarity increased, indicating lower agreement among participants when evaluating highly dissimilar BSD pairs, and higher agreement for more similar BSD pairs. This suggests that the computational measure captures not *what* people perceive as different, but rather *when* perceptual differences become more salient.

This finding may reposition the use of computational similarity metrics in creativity support systems as well. Rather than directly approximating perceived dissimilarity, computational metrics may function as indicators of perceptual clarity or salience. In other words, they may be more effective at identifying regions of the design space where differences are consistently recognized by humans rather than accurately modeling the magnitude of perceived differences.

The strong negative correlation between similarity and dissimilarity ratings (ρ = -0.85) indicates that participants were broadly sensitive to a shared perceptual structure, but the less-than-perfect (smaller than 1) correlation suggests that similarity and dissimilarity judgments are not fully interchangeable. This finding contributes to ongoing discussions about whether similarity and dissimilarity can be treated as mathematically equivalent in empirical settings and supports the need to consider task framing when designing perceptual studies and evaluation metrics [14].

4.1. Implications for Design Systems, Generative Design, and Creative AI

The findings of this study have implications for the use of computational similarity and dissimilarity measures in design systems and creative AI systems as well. Such metrics are often used to guide exploration, promote diversity, or even evaluate generated output. However, the weak alignment with human perceptual judgment suggests that relying on these measures may lead to unintended outcomes. For example, systems may prioritize variations that are computationally distinct but perceptually similar or overlook differences that are meaningful to users. At the same time, the observed relationship with perceptual agreement suggests another role for computational metrics. Rather than serving as direct proxies for human judgment, they may be used to inform hybrid approaches in which computational

measures are combined with human-in-the-loop evaluations [26], or to point human attention to ambiguous regions of the design space.

These findings also point to the need for the development of more perceptually grounded dissimilarity measures. Rather than relying solely on algorithmic distance functions, future measures may benefit from being explicitly informed by human perceptual mechanisms. This could be achieved, for example, by learning from perceptual data or through iterative human-in-the-loop refinement. Such approaches could help bridge the gap between computational efficiency and perceptual relevance, leading to measures that better reflect how differences between designs are actually experienced by users.

4.2. Limitations & Future Work

Several limitations should be taken into account when interpreting these results. The computational dissimilarity values showed a relatively restricted range (min = 0.237, max = 0.516), which may have reduced sensitivity to detect stronger relationships with perceptual judgment. In addition, the low inter-rater reliability at the stimulus-pair level indicates substantial variability in individual judgments. While this reflects the subjective nature of the task, it also limits the extent to which shared perceptual structures can be inferred for 3D geometries. Furthermore, the BSD corpus used in this study consisted of designs with comparable structural complexity. Although this helped reduce potential confounding effects arising from differences in design complexity, it also limits the generalizability of the findings to more heterogeneous stimulus sets.

These observations suggest several directions for future work. Expanding the range and diversity of stimuli beyond those of Pereverdieva [16] may help to better capture the full spectrum of perceptual differences and increase sensitivity to potential alignment with computational measures. Future studies could also investigate whether variations in design complexity itself influence perceptual agreement. Moreover, participant feedback indicated that perceptual judgments were often based on features not explicitly represented in the current computational measure, including global shape information (e.g., bounding box or volumetric form) and other holistic properties. This suggests that incorporating higher-level or more global geometric features may improve alignment with human perceptual judgment.

5. Conclusion

This work investigated the extent to which a computational measure of dissimilarity aligns with human perceptual judgment in a specific context involving 3D geometries. Across correlation and mixed-effects analyses, the results indicate limited alignment between computational dissimilarity values and human judgments. At the same time, the findings suggest that, in this work, the computational measure captures aspects of perceptual salience or clarity, rather than directly approximating perceived dissimilarity. These findings may contribute to a more nuanced, general understanding of the role of computational dissimilarity in design systems, creativity support systems, and creative AI systems, and highlight the need for approaches that combine computational efficiency with perceptual grounding.

Future work should explore the development of perceptually informed dissimilarity measures and investigate how human judgments can be more effectively integrated into computational design systems. Understanding not only what computational measures capture, but also when they diverge from human perception, is essential for developing systems that meaningfully support human-centered design and creativity.

Acknowledgments

Our thanks go to EAISI, the Eindhoven Artificial Intelligence Systems Institute, since the research for this contribution was made possible by their Exploratory Multidisciplinary AI Research (EMDAIR) grant (Project no. 64).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.2 in order to: Improve writing style, Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. A. Runco, G. J. Jaeger, The Standard Definition of Creativity, Creativity Research Journal 24 (2012) 92–96. doi:10.1080/10400419.2012.650092.

[2] R. J. Sternberg, T. I. Lubart, The Concept of Creativity: Prospects and Paradigms, in: R. J. Sternberg (Ed.), Handbook of Creativity, 1 ed., Cambridge University Press, 1999, pp. 3–15. doi:10.1017/CBO9780511807916.003.

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

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

[5] J. C. Kaufman, J. A. Plucker, J. Baer, Essentials of Creativity Assessment, Essentials of Psychological Assessment Series, Wiley, Hoboken, N.J, 2008.

[6] J. P. Guilford, The Nature of Human Intelligence, McGraw-Hill Series in Psychology, McGraw-Hill, New York, 1967.

[7] E. P. Torrance, Torrance Tests of Creative Thinking: Norms-Technical Manual: Verbal Tests, Forms A and B: Figural Tests, Forms A and B, research ed. ed., Personal Press, Lexington, Mass, 1966.

[8] J. Baer, How divergent thinking tests mislead us: Are the Torrance Tests still relevant in the 21st century? The Division 10 debate., Psychology of Aesthetics, Creativity, and the Arts 5 (2011) 309–313. doi:10.1037/a0025210.

[9] P. J. Silvia, B. P. Winterstein, J. T. Willse, C. M. Barona, J. T. Cram, K. I. Hess, J. L. Martinez, C. A. Richard, Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods., Psychology of Aesthetics, Creativity, and the Arts 2 (2008) 68–85. doi:10.1037/1931-3896.2.2.68.

[10] P. Karimi, M. L. Maher, N. Davis, K. Grace, Deep Learning in a Computational Model for Conceptual Shifts in a Co-Creative Design System, 2019. doi:10.48550/ARXIV.1906.10188.

[11] M. L. Maher, D. Fisher, Using AI to Evaluate Creative Designs, in: Proceedings of International Conference on Creative Design, 2012, pp. 45–54.

[12] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A Reference-free Evaluation Metric for Image Captioning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7514–7528. doi:10.18653/v1/2021.emnlp-main.595.

[13] R. L. Goldstone, The role of similarity in categorization: Providing a groundwork, Cognition 52 (1994) 125–157. doi:10.1016/0010-0277(94)90065-5.

[14] A. Tversky, Features of similarity., Psychological Review 84 (1977) 327–352. doi:10.1037/0033-295X.84.4.327.

[15] A. Markman, D. Gentner, Structural Alignment during Similarity Comparisons, Cognitive Psychology 25 (1993) 431–467. doi:10.1006/cogp.1993.1011.

[16] K. Pereverdieva, Sample for dissimilarity measure testing, 2023. doi:10.5281/ZENODO.10069531.

[17] K. Pereverdieva, M. Emmerich, A. Deutz, H. Hofmeyer, T. Ezendam, T. Bäck, An Alignment Method and a Dissimilarity Measure for 3D Building Spatial Designs, IEEE Access 13 (2025) 195599–195610. doi:10.1109/ACCESS.2025.3626145.

[18] K. A. Hallgren, Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial, Tutorials in Quantitative Methods for Psychology 8 (2012) 23–34. doi:10.20982/tqmp.08.1.p023.

[19] T. K. Koo, M. Y. Li, A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research, Journal of Chiropractic Medicine 15 (2016) 155–163. doi:10.1016/j.jcm.2016.02.012.

[20] P. E. Shrout, J. L. Fleiss, Intraclass correlations: Uses in assessing rater reliability., Psychological Bulletin 86 (1979) 420–428. doi:10.1037/0033-2909.86.2.420.

[21] F. Henninger, Y. Shevchenko, U. K. Mertens, P. J. Kieslich, B. E. Hilbig, Lab.js: A free, open, online study builder, Behavior Research Methods 54 (2022) 556–573. doi:10.3758/s13428-019-01283-5.

[22] Z. Abu-Aisheh, R. Raveaux, J.-Y. Ramel, P. Martineau, An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems, in: Proceedings of the International Conference on Pattern Recognition Applications and Methods, SCITEPRESS - Science and and Technology Publications, Lisbon, Portugal, 2015, pp. 271–278. doi:10.5220/0005209202710278.

[23] P. Jaccard, The Distribution of the Flora in the Alpine Zone, New Phytologist 11 (1912) 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x.

[24] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2 ed., Routledge, 1988. doi:10.4324/9780203771587.

[25] M. Smithson, J. Verkuilen, A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables., Psychological Methods 11 (2006) 54–71. doi:10.1037/1082-989X.11.1.54.

[26] P. Geyer, Multidisciplinary grammars supporting design optimization of buildings, Research in Engineering Design 18 (2008) 197–216. doi:10.1007/s00163-007-0038-6.