A Comparison Study of Human and Machine Generated Creativity

Liuqing Chen
a. College of Computer Science and Technology,
Zhejiang University, Hangzhou, 310058, China
b. Zhejiang – Singapore Innovation and AI Joint Research Lab,
Zhejiang University, Hangzhou, 310058, China
e-mail: chenlq@zju.edu.cn

Lingyun Sun
a. International Design Institute of ZJU,
Zhejiang University, Hangzhou, 310058, China
b. Zhejiang – Singapore Innovation and AI Joint Research Lab,
Zhejiang University, Hangzhou, 310058, China
e-mail: sunly@zju.edu.cn

Ji Han[1]
INDEX, Business School,
University of Exeter, Exeter, EX4 4PU, UK
e-mail: j.han2@exeter.ac.uk

[1] Corresponding author: Ji Han, j.han2@exeter.ac.uk

ABSTRACT

Creativity is a fundamental feature of human intelligence. However, achieving creativity is often considered a challenging task, particularly in design. In recent years, using computational machines to support people in creative activities in design, such as idea generation and evaluation, has become a popular research topic. Although there exist many creativity support tools, few of them could produce creative solutions in a direct manner, but produce stimuli instead. DALL·E is currently the most advanced computational model that could generate creative ideas in pictorial formats based on textual descriptions. This study conducts a Turing test, a computational test and an expert test to evaluate DALL·E's capability in achieving combinational creativity comparing with human designers. The results reveal that DALL·E could achieve combinational creativity at a similar level to novice designers and indicate the differences between computer and human creativity.

Keywords*: Artificial Intelligence, Computer Aided Design, Human Computer Interfaces/interactions*

1. Introduction

Creativity has attracted great research interest in psychology, cognitive science, computer science, engineering, and design fields for many years, and has a profound impact on society [1]. It is defined as ‘the process by which something so judged (to be creative) is produced’ [2], which is an essential skill to be successful in the current complex and interconnected world [3]. In the past decades, several methods and approaches, also known as creativity tools, are developed to support the generation of creative ideas. Brainstorming, six thinking hats [4], SCAMPER [5], morphological analysis [6], and TRIZ [7] are the most often used ones. Most of these conventional tools were not developed specifically for design. Design-focused tools, such as the WordTree method [8], 77 design heuristics [9], and bio-inspired design [10, 11], are thereby developed specifically for supporting creative design idea generation. However, many designers still prefer not to use these non-computational tools due to lack of knowledge and experience, difficulties in mastery, and seemingly cumbersome steps which could cause additional work [12].

In recent years, a number of computational design support tools have been explored to tackle these limitations. For example, Han et al. [13] came up with an analogical reasoning tool for supporting idea generation by employing aspects of ontology and producing a corresponding image mood board; Sarica et al. [14] developed a technology semantic network based on patent data, which could support ideation by knowledge discovery; Siddharth et al. [15] proposed an engineering knowledge graph, containing < entity, relationship, entity > triples extracted from patent database, to support inference and reasoning; Obieke et al. [16] came up with a computational framework that explores new engineering design problems for creativity. Most of the existing so-called computational creativity tools do not generate creativity in a direct manner, but produce stimuli instead, such as texts and images, to prompt designers' creative minds.

Combinational creativity involves unfamiliar combinations of familiar ideas, which is the easiest approach for humans to achieve creativity [17]. Producing combinational creativity is a natural feature of humans' associative memory system, while it is challenging for computers, due to issues such as the need for a rich store of knowledge, the ability to form various combinations, and the competence to evaluate combination outputs [17-20]. However, the rapid advancements in the field of artificial intelligence, such as deep learning based computer vision and natural language processing (NLP), have provided new and better approaches to enable computers to produce combinational creativity. To the best of the authors' knowledge, no studies to date have compared the performance between humans and computers in producing combinational creativity. This leads to a debatable question that whether computational machines (computers) can outperform humans in achieving combinational creativity.

Evaluating combinational creativity is challenging, and there is no widely adopted method for such evaluation. In the field of design creativity, a variety of creativity assessment methods have been proposed, which generally require human raters to judge the quality of generated creativity [21], such as the Consensual Assessment Technique method [38], Creative Product Semantic Scale [22], Product Creativity

Measurement Instrument (PCMI) [23], Creative Solution Diagnosis Scale (CSDS) [24], and using creativity metrics [25, 26]. In the field of artificial intelligence, the common computational metrics for evaluating generative models involve Inception Score (IS) [27] and Frechet Inception Distance (FID) [28], which are quantitative and calculated based on probability distribution. In the interdisciplinary research between artificial intelligence and human study, Turing test is a basic and widely adopted method [30-32], as it can provide an overall impression of how a machine performs. With consideration of the advantages of the evaluation methods in these three areas, this study applies a combined research approach by conducting a CAT based expert test, a computational test and a Turing test, and then synthesizes the results to elicit useful findings.

Therefore, the aim of this paper is to compare the combinational creative performance of machines and human designers, and explore the differences between human designers and computers in generating creativity. This is the first study that compares the performance between novice designers and machines regarding combinational creativity, which employs a combined research approach integrating a Turing test, a computational test and an expert test. This study will shed light on the research of computational creativity evaluation and artificial intelligence applications in design. The following section provides the theoretical background of this study. The methodology of the study is described in Section 3, and followed by the implementation of the Turing test, computational test and expert test in Section 4. In Section 5 and 6, the results of the tests are presented, analyzed and discussed. The paper is then concluded in Section 7.

2. Theoretical Background

Combinational creativity is claimed to be one of the best approaches for fully utilizing nowadays abundant data, including texts, images, concepts, sounds and so on [29], to achieve creativity [30]. A number of studies have explored combinational creativity in the context of design, particularly in idea generation. For instance, Nagai et al. [31] proposed three types of concept-synthesizing processes, namely property mapping, concept blending, and concept integration in thematic relation, for generating new concepts based on three interpretation methods of combinational phrases respectively. Han et al. [32] indicated that associating far-related ideas for forming combinational ideas could lead to outcomes that are more creative in comparison with linking closely-related ones. Han et al. [33] investigated how combinational creativity is formed in design, focusing on conventional noun-noun combinations. It was revealed that a noun-noun combinational idea is produced by associating a base idea and an additive idea. The base idea refers to the basic idea of the combinational idea, while the additive idea could be a problem-solving idea, a similar representational idea, or an inspirational idea. For example, the famous Juicy Salif is an example of associating a basic idea (a manual juicer) and an inspirational additive idea (a squid). This study has thereby laid a theoretical foundation for our paper exploring human and machine generated combinational creativity.

Although Han et al. [19] and Chen et al. [34], [35] have employed pictorial data to form combinational images to facilitate users in combinational creativity, these combinational images are produced independently from semantic contexts. For instance, the Combinator [19] produces a compound phrase of ‘flower glass’ and a corresponding combinational image of merging a ‘flower’ and ‘glass’. Without semantic context, the combinational image produced could represent a ‘flower’ made out of ‘glass’, a piece of ‘glass’ in the shape of a ‘flower’, or a piece of ‘glass’ with printed ‘flowers’. This might cause potential distractions and affect users’ creative performance.

In recent years, several computational models are developed to transform texts into images, such as LeicaGAN [36] and Semantic-Spatial Aware GAN [37]. These models could exploit text information for producing semantically consistent realistic images. Among them, DALL·E [38] is one of the most advanced ones, which employs GPT-3 [39] trained on a set of text-image pairs data for producing images based on text descriptions. As introduced by OpenAI [40], DALL·E has distinguishing capabilities, such as creating anthropomorphized versions of animals and objects. Moreover, it seems to have achieved a certain level of creativity. Specifically, the model could create pictorial combinations of unrelated concepts in plausible ways, even producing fantastical objects that do not exist in reality, according to textual descriptions. Thus, DALL·E is considered one of the most powerful systems capable of generating combinational creativity in pictorial formats within the constraints of texts. In this study, we perform a thorough performance benchmark evaluation comparing DALL·E with novice designers regarding combinational creativity, involving a Turing, a computational, and an expert test.

3. Methodology

To compare the performance between human novice designers and machines regarding combinational creativity, we first create two datasets for evaluation: the machine dataset and the human dataset. As shown in Figure 1, the input for both DALL·E and novice designers are the same textual prompts which contain combinational design ideas. The outputs are images matching the corresponding textual prompts. After selections, the same amount of data sets are saved as the machine dataset and the human dataset respectively. This is then followed by three tests: a Turing test, a computational test, and an expert test, in which the human and machine data are evaluated employing corresponding approaches.

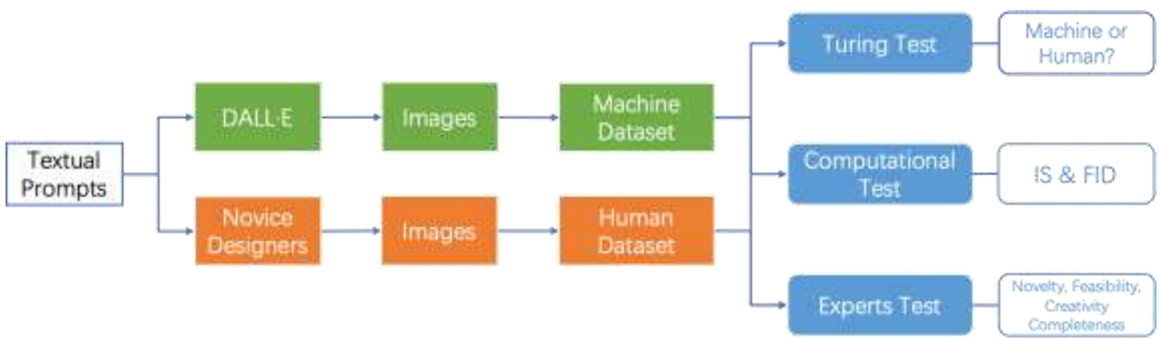


Figure 1. The workflow of the proposed research approach

3.1. Data Source – Machine and Human Datasets

Only a partial code of the DALL·E model was released on Github, it is thereby impossible to run DALL·E to generate images due to missing training codes and data. Thus, the performance of DALL·E is evaluated based on the presented outcome from OpenAI's official blog, in which the published data is representative and of high quality. In the blog, sets of textual descriptions and the corresponding generated images by

DALL·E are presented. Three designers with over three years of experience were invited to judge whether the textual description in each set is a combinational idea. Prior to the judgement, the authors have well explained the definition of combinational creativity and showed some practical cases to the designers. If a set was judged as combinational creativity based, then five corresponding top-ranked images produced by DALL·E were collected. In total, eight sets, with five images in each set, are collected as the machine generated combinational creativity dataset. All the input texts and one corresponding machine-produced image sample in each set are shown in Table 1.

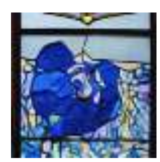
Seven novice designers were employed to create a human dataset. They are either postgraduates or employees in companies with less than three years of working experience. They all hold a bachelor's degree in design disciplines, and have at least two years' experience in product design and graphic design. Since the human dataset is associated with combinational creativity, prior to the creation of data, each designer was informed of the definition of combinational creativity and related design cases, especially the meaning of 'base' and 'additive'. Each designer was required to produce a drawing for each of the textual descriptions as indicated in Table 1 by using familiar computer-aided design software within one hour. The designers were required to use white backgrounds and not to include any textual annotations to be in line with the ones of the machine dataset. Besides, the quality of drawings should be as high as possible, which is measured from three aspects:

1) Novelty: The drawing should be new, unusual, original and attractive.
2) Usefulness: The drawing should be feasible, reasonable and appreciable.

3) Creativity completeness: The drawing should match the corresponding textual description, and combined concepts could be visible to recognize.

As a result, eight sets of data involving seven images each are produced. Three designers were then employed to select the top five images within each set. The eight sets of corresponding image samples produced by human designers are shown in Table 1.

Table 1. An overview of the machine and human data

Group No.	Input	Machine Output	Human Output
1	a pentagonal green clock. a green clock in the shape of a pentagon		
2	a capybara made of voxels sitting in the field		
3	a stained-glass window with an image of a blue strawberry		
4	a snail made of harp. A snail with the texture of a harp		
5	an armchair in the shape of an avocado. an armchair imitating an avocado		
6	a giraffe imitating a turtle. a giraffe made of turtle		

7	a cube made of porcupine. a cube with the texture of a porcupine		
8	a professional high-quality emoji of a lovestruck cup of boba		

3.2. Evaluation Methods

3.2.1. Turing Test

A Turing test [41] is conducted in this study to explore whether DALL·E can achieve combinational creativity at the human level. In the test, participants were required to identify whether an image, within our mixed machine and human datasets, is produced by machine or human, providing the image's corresponding textual background. The test is consistent with the studies and arguments by Boden [42]; Pease and Colton [43]; Peter Berrar and Schuster [44]. The test is specific and blinded, and contains necessary contextual information. Though DALL·E is encouraged to produce realistic images in accordance with texts, it is not exclusively encouraged to exhibit creative behaviors. Therefore, the machine dataset, which can reflect DALL·E's capability of combinational creativity, was exclusively constructed to avoid possible trickery behaviors. For instance, instead of selecting the most realistic images generated by DALL·E to cheat human observers, we required that the images should first match their textual combinational ideas.

3.2.2. Computational Test

Given a deep learning based model for image generation, such as VAE [45] and GANs [46] based, the most common metrics for evaluating its capability are Inception Score (IS) and Frechet Inception Distance (FID). IS concerns the realism and diversity of generated images when evaluating a specific model. Specifically, IS calculates the KL divergence between the probability distribution of every generated image and the overall average of all generated images [27]. As shown in $Equation\ (1)$, given N classes, KL divergence is calculated between the conditional probability $p(y|x)$ in which a generated image x is classified into a particular class y, and the average probabilities for all the images in the class group $p(y)$ which is also called marginal distribution. High diversity of the generated images' categories and high certainty of the arbitrary image's category indicate high KL divergence, which means high IS and a better corresponding model, and there is no maximum value for IS.

$$\mathrm{IS}(G) = \exp\left(\frac{1}{N}\sum\nolimits_{i=1}^{N} D_{KL}\left(p\left(y \mid \mathbf{x}^{(i)}\right) \parallel \hat{p}(y)\right)\right) \quad (1)$$

FID is proposed to perform better in terms of discriminability, robustness and computational efficiency and to address the limitations of IS [28]. It calculates the distance of two multidimensional normal distributions based on the mean (μ) and covariance (∑) of the vectors extracted from both real (with the subscript r) and generated images (with the subscript g), as shown in the $Equation\ (2)$. Ideally, the FID can be zero if the generated data is identical to real data, while higher FID value corresponds to low quality of generated images. Considering the popularity of these two metrics in generative models' evaluation, we calculate both values for our machine and human datasets respectively, and then compare them.

$$\mathrm{FID} = \left\|\mu_r - \mu_g\right\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right) \quad (2)$$

3.2.3. Expert Test

The Turing test can estimate the overall appreciation of DALL·E's performance compared with humans by subjective evaluation, while the computational test can quantitatively and objectively compare machine and human performance but lack detailed and interpretable criteria. Hence an expert test is necessary to deeply investigate the difference between the two groups and provide interpretable results. In this study, a Consensual Assessment Technique (CAT) based method [47] is adopted in the expert test for creativity evaluation.

Novelty, quantity, quality and variety are the four metrics often used in design research for evaluating creativity [25]. In the expert test, a modified version of the metrics was adopted. Novelty, feasibility and creativity completeness were used to measure a single image, and variety was used to measure a group of images generated by either a human designer or machine. The combinational creativity images are generated based on textual descriptions, thus novelty originates from the creation of combining the 'base' elements with the 'additive' elements, such as the novelty of the creation of combining 'armchair' with 'avocado' in an imagery format. On the other hand, creativity completeness is an essential metric for evaluating the transformation quality from textual description to imagery visualization, instead of focusing on evaluating creation results (novelty). Since some of the combinational ideas are imaginary rather than physical, such as ‘a giraffe imitating a turtle’, feasibility is chosen as the metric instead of quality and utility. The meanings of novelty, feasibility and creativity completeness are identical to the descriptions for ranking drawings in the human dataset indicated in the preceding. Variety refers to the diversity of a set of images, which measures the differences between images.

4. Evaluation

4.1. Turing Test

The Turing test is conducted by developing a website where all web pages are completely customized to minimize distractions. Participants were asked to read the instructions, agree with the test protocols, and provide demographic information before starting the test. Eight groups of questions in total, corresponding to eight groups of data in our datasets, are provided to the participants. Each group contains ten questions that are randomly ordered for mixing the human and machine generated data, while five questions are from the human dataset and another five are from the machine dataset. This fact is not revealed to the participants to avoid introducing any potential bias. This would not influence participants’ choices since they could feel free to make decisions without restrictions. There is only one question on each webpage consisting of a question serial number, a short textual description, an image which is either from the human dataset or the machine dataset, and two buttons indicating ‘human’ and ‘machine’ for participants to choose, as shown in Figure 2. The participants were required to spend at least three seconds on each question before moving to the next one.

Figure 2. A question webpage in the Turing test

After a successful pilot test, the test was distributed across multiple channels, including university BBS, social media, and personal contacts. Each participant was invited for an interview voluntarily when completing the test. Three questions were asked in the interview:

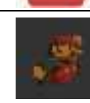
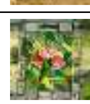
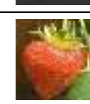
1) How difficult do you think this test is?
2) What is your method for distinguishing human and machine?
3) What is your feedback about this test?

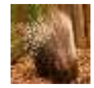
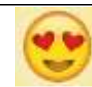
Answers of the interviews were collected and analysed in a qualitative way and the results were reported in the '5. Results and Analysis' section.

4.2. Computational Test

Two rounds of computational tests were conducted. In the first round, we implemented the algorithms of IS and FID by following Zhu et al. [48] and calculated IS and FID scores. FID calculation needs a reference distribution for comparison, so the mean and co-variance of COCO datasets [49] were used. However, it is found that some concepts in our datasets are not covered by COCO datasets, which might weaken the fairness of comparison. Therefore, we performed a second round of tests by comparing our data with a new reference dataset. As indicated in the preceding, a combinational idea consists of a base and an additive. Hence, we randomly collected 25 images for each base and additive in every group from the Internet, which results in 400 images in total. The 25 images for each base or additive were further equally divided into five reference groups in order to validate that no significant bias in image collection was introduced into the test. An overview of our reference data is shown in Table 2.

Table 2. An overview of the reference data

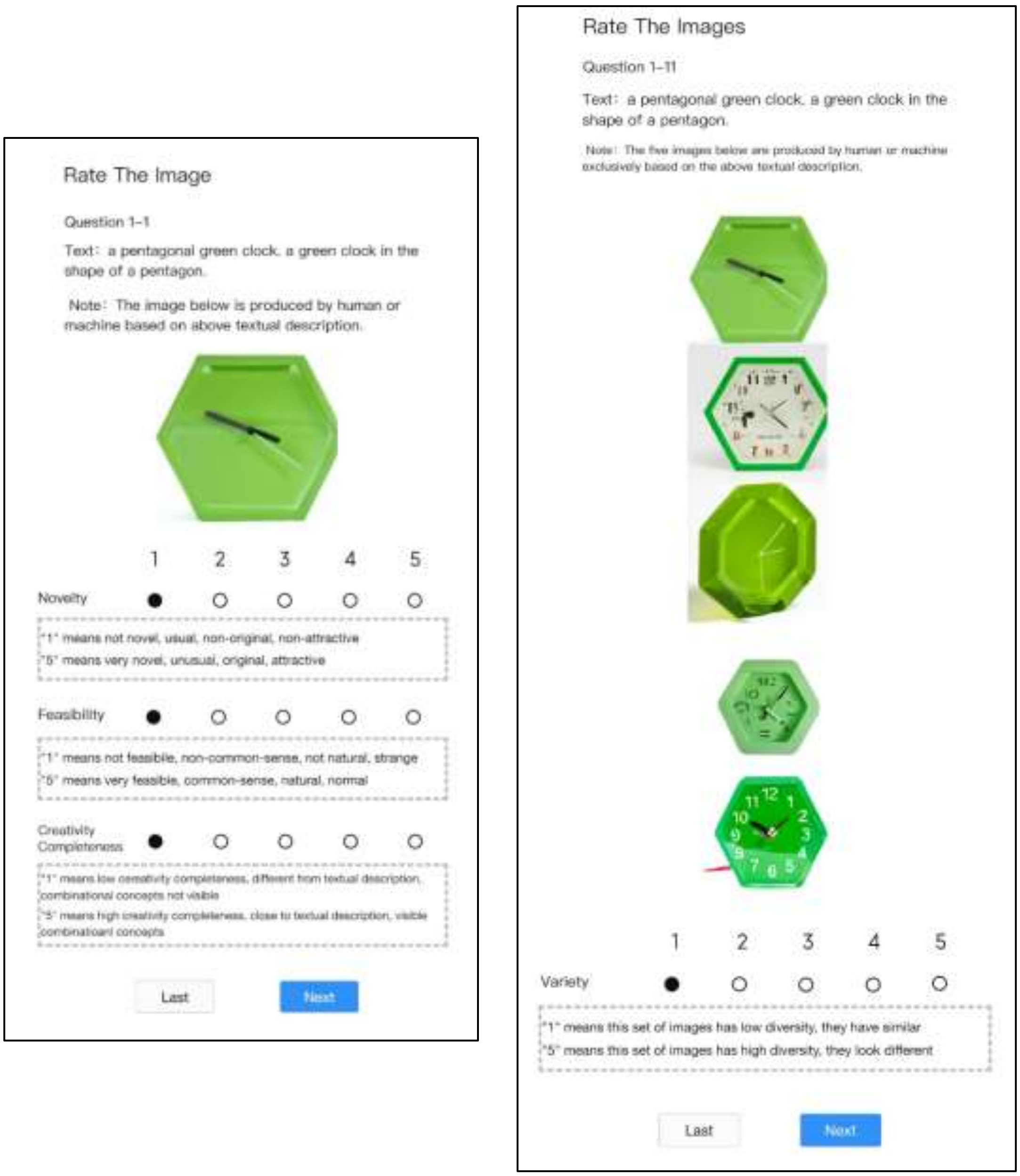
Group	Base	Sample-Base	Additive	Sample-Additive
1	Clock		Pentagonal	
2	Capybara		Voxels	
3	Glass		Strawberry	
4	Snail		Harp	
5	Armchair		Avocado	

6	Giraffe		Turtle	
7	Cube		Porcupine	
8	Cup		Emoji of lovestruck	

In the second round, we further calculated the IS of all five reference groups as a reference to the IS of the human and machine dataset. The new FID scores were calculated by comparing each reference group with the human and machine dataset respectively. Since each generated image is based on a combinational idea and contains concepts of base and additive, it is useful to investigate the FID by comparing the base and additive data to the human and machine dataset. Therefore, the five reference groups were further divided into base and additive sub-groups, and were used to calculate base-FID and additive-FID.

4.3. Expert Test

The expert test was also conducted via a customized website. There are eight groups of questions, and each contains twelve questions. In each group, the first ten questions are single image based, of which a textual description and corresponding image are provided in each question and participants are required to rate the image using a 5-point Likert scale regarding three metrics: novelty, feasibility and creativity completeness, as shown in Figure 3(a). The ten images are randomly selected from the human or machine datasets. The last two questions in each group are five-image based, in which a textual description and corresponding five images (merged in a vertical sequence) are shown. Participants are informed that all five images were generated by humans or machines exclusively, and they are required to rate the variety of the five images using a 5-point Likert scale, as shown in Figure 3(b).



(a) (b)

Figure 3. Webpages of two question examples in the expert test

Before starting the test, participants were asked to read the instructions and test protocols, and provide their demographic information. The explanation of four evaluation metrics (novelty, feasibility, creativity completeness and variety) was provided within the webpage, and further assistance was provided as well when experts had questions. There was no time limit for each question, and more than 30 seconds of rest time was provided in the test when experts completed half of the questions.

5. Results and Analysis

5.1. Turing Test

All ten images in each group shared the same textual description, and participants were not informed how many images of the ten are from the human or machine dataset, which means participants' judgement based on a single image is independent. Among a total of 100 received submissions, there were 97 participants who validly participated in this test by answering the 'human or machine' questions, while three submissions were considered invalid as it was reported by the participants that some machine generated images in the test were seen previously. The mean accuracy of each question within each group was calculated, as well as the mean accuracy of every group. The overall accuracy was obtained by averaging the accuracy of eight groups, which is 55.9%, as shown in Table 3. Furthermore, group-8 achieved 42.4% which is below 50% and the accuracy of group-6 is also very close to 50%.

Accuracy concerns whether a question is correctly answered or not, rather than which answer is more often answered. Given a classification problem, human or machine classes in our case, three metrics are widely applied when measuring the performance of a classification machine learning model: precision, recall and F1 score. The formulas of the three metrics are given in *Equation* (3), (4), (5) respectively, where TP represents True Positive and similarly FN represents False Negative. In our calculation, Positive means the answer is ‘human’ while Negative indicates ‘machine’. The results of precision, recall and F1 score of the two classes (human and machine) are presented in Table 3. As shown in the table, the precision between human and machine is very close (56.1% versus 55.6%), but the recall between human and machine are noticeably different. The recall of the machine class is higher than the human class by 7.6%, which is due to high TN and high FN. Besides, the F1 score of the machine dataset is higher than human by 3.3%.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

Table 3. Results of the Turing test

		Mean	1	2	3	4	5	6	7	8
Accuracy		55.9%	60.8%	55.2%	61.9%	60.3%	54.2%	51.1%	61.1%	42.2%
Machine	Precision	55.6%	60.6%	54.4%	63.5%	60.8%	54.2%	51.1%	59.0%	42.2%
	Recall	57.9%	61.9%	64.1%	55.7%	57.9%	54.2%	53.4%	73.0%	42.7%
	F1	56.7%	61.2%	58.8%	59.3%	59.3%	54.2%	52.2%	65.3%	42.5%
Human	Precision	56.1%	61.1%	56.3%	60.6%	59.8%	54.2%	51.2%	64.6%	42.1%
	Recall	53.8%	59.8%	46.2%	68.0%	62.7%	54.2%	48.9%	49.3%	41.6%
	F1	54.9%	60.4%	50.7%	64.1%	61.2%	54.2%	50.0%	55.9%	41.9%

It is also useful to explore the variance of accuracy among questions and groups when investigating machine and human classes respectively. Therefore, the statistics of minimum and maximum accuracy in each group in terms of human and machine classes are collected and presented in Table 4. As indicated in the table, both humans and machines have very high variance throughout all groups, while the variance in the human class is higher than the machine class. The highest accuracy in the human class (93.8%) is higher than the machine class (86.6%) while the lowest accuracy in the human class (16.5%) is lower than the machine class (26.8%), which corresponds to the value of *(Max – Min)* between human and machine. The difference between the maximum and minimum accuracy in the human class is higher than the machine class with 17% on average.

Table 4. Variance of accuracy in different groups

		Min	Max	Max-Min	Difference
1	Human	37.1%	90.7%	53.6%	23.7%
	Machine	51.5%	81.4%	29.9%	
2	Human	19.6%	74.2%	54.6%	18.6%
	Machine	40.2%	76.3%	36.1%	
3	Human	43.3%	79.4%	36.1%	-2.1%
	Machine	38.1%	76.3%	38.1%	
4	Human	45.4%	93.8%	48.5%	26.8%
	Machine	46.4%	68.0%	21.6%	
5	Human	16.5%	89.7%	73.2%	45.4%
	Machine	43.3%	71.1%	27.8%	
6	Human	27.8%	77.3%	49.5%	7.2%
	Machine	32.0%	74.2%	42.3%	
7	Human	40.2%	63.9%	23.7%	-5.2%
	Machine	57.7%	86.6%	28.9%	
8	Human	25.8%	69.1%	43.3%	21.6%
	Machine	26.8%	48.5%	21.6%	
Overall	Human	16.5%	93.8%	77.3%	17.5%
	Machine	26.8%	86.6%	59.8%	
Mean	Human	32.0%	79.8%	47.8%	17.0%
	Machine	42.0%	72.8%	30.8%	

Twenty participants accepted the interview and answered questions after completing the Turing test. Concerning the method of distinguishing human and machine, the participants indicated that they believe the human-generated images have ‘more clear details’, ‘a unified style (such as sketches)’, and ‘high resolutions’, while the machine-generated images are ‘unreal’, ‘blurred’ and have ‘unhuman combination logics’ and ‘cut and paste by Photoshop patterns’. In terms of the difficulty of the task, the participants suggested that natural or physical subjects are easy to make ‘human’ or ‘machine’ selections, as well as images employing sketch styles. The interview results are a supplement to the Turing test, and can potentially explain the Turing test results and help understand the reasons underpinning the choices made by the participants. This is in line with other similar studies. For example, Sarica et al. [50] interviewed twenty-five participants to understand their choices of the best computational representation of a specific design, and Zhu [51] interviewed ten engineers regarding their views towards a set of computationally generated design concepts.

5.2. Computational Test

The computed results of the Inception Score (IS) are shown in Figure 4 where the IS of five reference groups are presented together for reference purposes. The machine group has a higher IS than the human group by 4.6%. The IS of the five reference groups are much higher than the machine and human datasets with an average IS of 7.65 ($\sigma = 0.27$). The computed FID scores including reference groups are presented in Figure 5. When comparing with COCO datasets, the FID of the machine dataset is higher than the human dataset by 6.7%. All the FID scores in comparison with reference groups are lower than COCO datasets, and all the FID scores of the machine group are higher than the human group. The average FID of the machine group in comparison with the five reference groups is 288 ($\sigma = 6.07$), which is higher than the average FID of the human group ($\mu = 233, \sigma = 5.43$) by 23.8%.

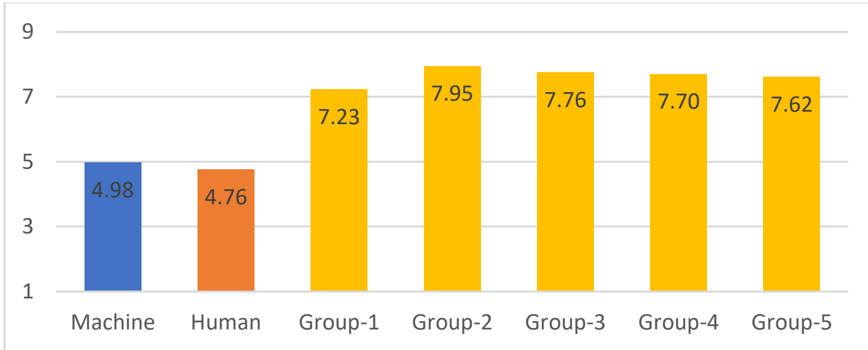


Figure 4. The IS values of different test groups

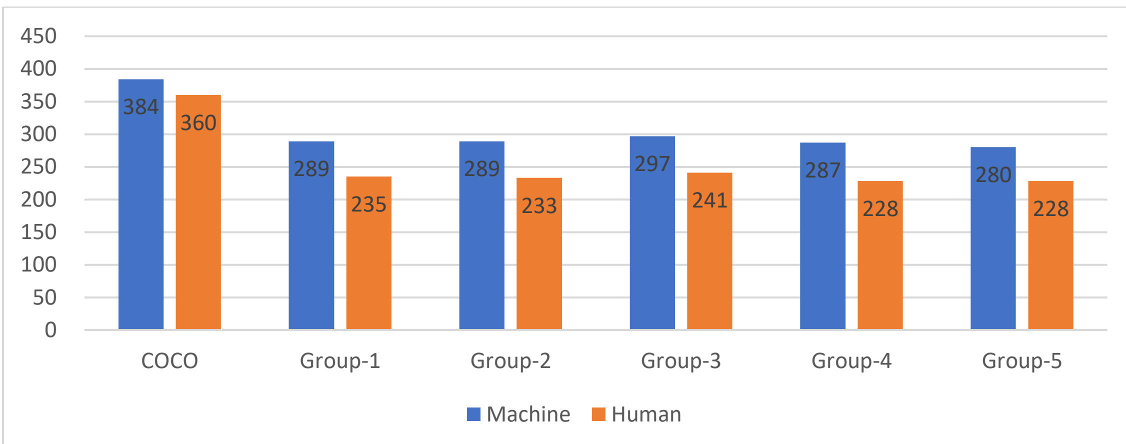


Figure 5. The FID scores of different test groups

In addition to calculating FIDs with the mixed data of bases and additives in five reference groups, we further computed the FIDs comparing with base groups and additive groups respectively, as shown in Figure 6. The FID of the machine group ($\mu = 273, \sigma = 6.02$) is slightly higher than the human group ($\mu = 247, \sigma = 6.54$) by 10.5% in comparison with base groups, while the FID of the machine group ($\mu = 324, \sigma = 4.44$) is significantly higher than the human group ($\mu = 257, \sigma = 2.88$) by 26% in comparison with additive groups. It is useful to investigate the influence of base and additive on the overall FID respectively. The FID scores in comparison with five base or additive groups (called base-FID and additive-FID respectively) are presented in Figure 7. As shown in the Figure 7 (a) (machine dataset), the additive-FIDs are higher than the base-FIDs on average by 18.7%, while Figure 7 (b) (human dataset) shows that the additive-FIDs are slightly higher than the base-FIDs only by 4.0%.

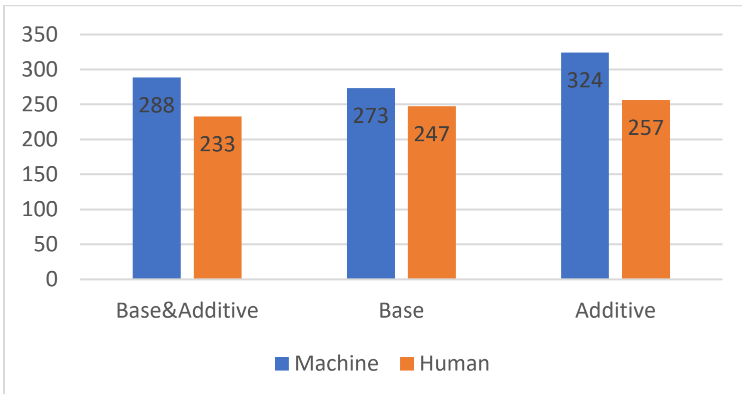
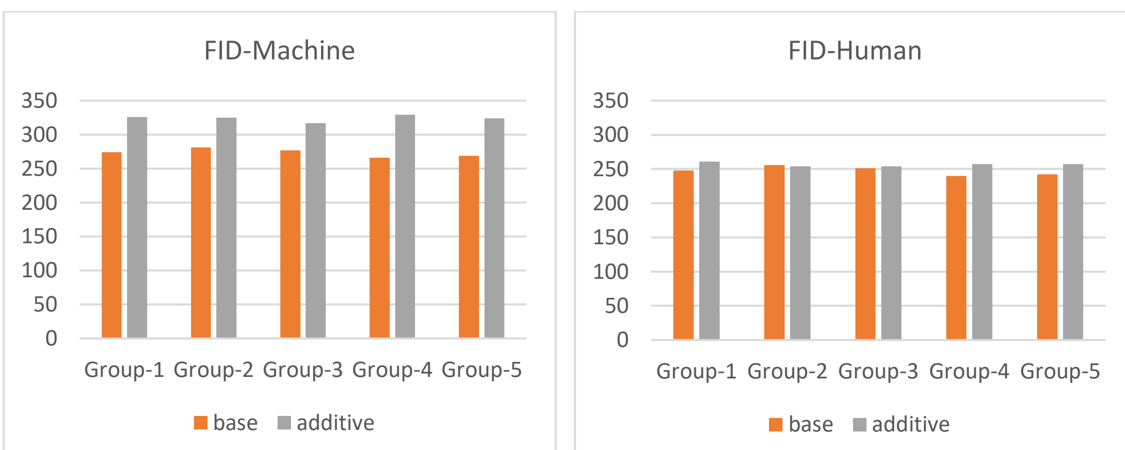


Figure 6. The FID scores in comparison with divided reference groups



(a) (b)

Figure 7. The base-FIDs and additive FIDs comparison within the machine (a) and human (b) datasets

5.3. Expert Test

With consideration of CAT requirements and the burden of evaluation, 19 professional designers with more than three years of working experience participated in the expert test. The four metrics proposed are calculated and presented in Table 5. In terms of novelty, more than half of the groups scored lower than 3, and the maximum value is lower than 3.5. The human dataset achieved higher novelty ($\mu = 2.90, \sigma = 0.14$) than the machine group ($\mu = 2.78, \sigma = 0.39$). There are three groups related to the machine dataset that obtained higher novelty scores than the human dataset. As shown in the table, the human dataset has a higher feasibility score ($\mu = 3.41, \sigma = 0.36$) than the machine dataset ($\mu = 3.23, \sigma = 0.46$). The same groups related to the machine dataset surpass the human dataset regarding feasibility. Similarly, the human dataset achieved higher creativity completeness ($\mu = 3.36, \sigma = 0.25$) than the machine group ($\mu = 3.09, \sigma = 0.49$). Two groups related to the machine dataset obtained higher creativity completeness scores than the human dataset. For variety, the human dataset has a significantly higher score ($\mu = 3.52, \sigma = 0.50$) than the machine dataset ($\mu = 2.95, \sigma = 0.47$), but there are three groups related to the machine dataset that surpass the human dataset. Both the human and machine datasets have higher variance than other metrics.

Table 5. Results of expert test

Metrics	Data Origin	1	2	3	4	5	6	7	8	Mean	Variance
Novelty	Machine	2.38	2.58	2.97	**3.23**	**3.43**	2.48	2.43	**2.76**	2.78	0.39
	Human	**2.89**	**2.83**	**3.15**	3.04	2.86	**2.74**	**2.94**	2.75	**2.90**	0.14
Feasibility	Machine	3.80	2.88	3.03	**3.40**	**3.68**	2.49	2.93	**3.58**	3.23	0.46
	Human	**3.88**	**3.48**	**3.92**	3.09	3.46	**3.01**	**3.04**	3.35	**3.41**	0.36
Completeness	Machine	3.48	2.68	2.98	**3.57**	3.42	2.44	2.54	**3.64**	3.09	0.49
	Human	**3.53**	**3.06**	**3.64**	3.46	**3.48**	**3.40**	**2.89**	3.44	**3.36**	0.25
Variety	Machine	3.00	2.74	2.37	**3.42**	**3.47**	3.16	**3.26**	2.21	2.95	0.47
	Human	**3.84**	**4.37**	**3.84**	2.89	3.05	**3.21**	3.21	**3.74**	**3.52**	0.50

6. Discussion

6.1. Turing Test

The average mathematical expectation of random answers to all the questions in the Turing test is 50%, while the closer of overall accuracy to 50% indicates the more undistinguishable between human and machine generated data. Though the overall accuracy in the Turing test is above 50%, the gap is only 5.9%. The F1 scores of the machine and human datasets are both close to 50%, while the machine's score is slightly higher than the human's score due to high recall in the machine dataset. High variance within every group in both datasets indicates that participants have low certainty to make their judgements. Besides, as indicated in the confusion matrix in Figure 8, TN (predicted machine and actual machine) and FN (predicted machine and actual human) are relatively higher, which corresponds to higher recall and F1 score of the machine dataset. This suggests that the results reveal that DALL·E can deceive participants to a large extent, and the participants could hardly indicate which image is from the human or machine dataset, while the participants subjectively tended to believe that the data in the Turing test were more likely from machines rather than humans.

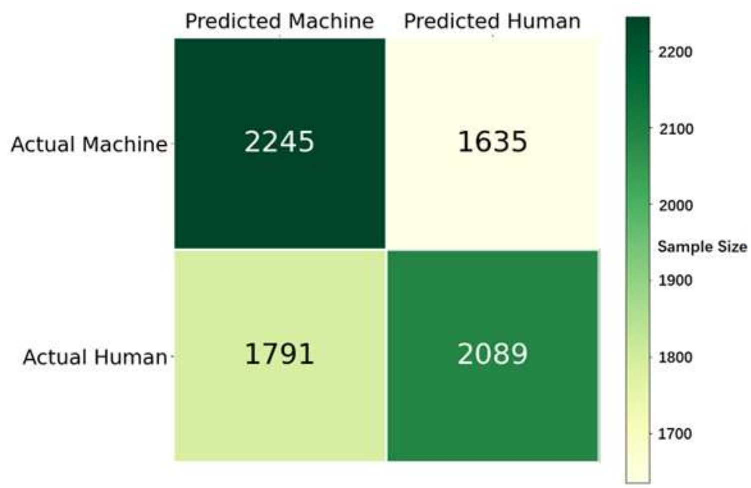


Figure 8. The confusion matrix of Turing test results

From our interview, it is shown that designers tend to use sketch and image processing software (such as Photoshop) to create drawings rather than 3D modelling and rendering, which makes their drawings more distinguishable from machine data. On the other hand, the images generated by DALL·E tend to be blurred, unsmooth, and unreal due to technical limitations, which makes them distinct from normal images. Besides, the logic behind a combination idea in machine data is sometimes different from human data. The ‘cut and paste by Photoshop’ pattern is considered a machine pattern by some participants, since some designers tend to create a collage-style image to express a combination idea while participants believe that machine is good at creating collages.

6.2. Computational Test

We created five reference groups in the computational test, and all the results related to the five groups have low variance, which indicates that there is only little bias brought into the reference groups. Regarding the IS metric, the machine dataset achieved a higher IS score than the human dataset, which means the machine generated data have higher quality than the designers', but the gap is as little as 4.6%. Five reference groups obtained much higher IS, since these reference images contain rich information about bases and additives and they are natural rather than combinational which is more favoured by the Inception model used for calculating IS. On the other hand, the machine dataset obtained a higher FID score than the human dataset when comparing with both COCO data and the five reference groups of data, indicating the machine generated data have a lower quality than the human generated data. All the FIDs in comparison with the five reference groups are significantly lower than in comparison with COCO data, validating that the images in our reference groups are closer to both the machine and human data than the images in COCO. The difference of FIDs between the human and machine datasets in comparison with five reference groups is bigger than the difference of FIDs in comparison with COCO data. This may reflect the difference in combinational design between humans and machines. Since it is required that drawings should be produced based on textual descriptions containing combinational creativity, novice designers tend to keep essential information from both base and additive in a combinational design while DALL·E is not trained to obtain this capability. This indicates that these designers have a better lingual understanding of combinational ideas and are able to transform them into designs than machines.

It is found that the difference of FIDs in comparison with base is less than in comparison with additive, as shown in Figure 6. This might suggest that designers are better at maintaining additive information than DALL·E to some extent. Furthermore, as shown in Figure 7, designers tend to balance base and additive information in a combinational design while DALL·E tends to maintain more information from the base rather than from the additive. However, there is no clear evidence that how much information should be maintained from base and additive respectively in a combinational design.

6.3. Expert Test

The human dataset obtained higher scores than the machine dataset by a small percentage (6.17% on average) when comparing the results regarding novelty, feasibility and creativity completeness, despite that the machine dataset has higher scores in some groups. This indicates that the novice designers performed slightly better than DALL·E in combinational designs in these three metrics. Besides, the designers outperform DALL·E evidently regarding variety by an overall gap of 19.15%, even though the machine dataset outperformed in three groups. This gap could be explained by two reasons. One is that the human data are from seven novice designers while the machine data is from DALL·E exclusively, which is unfair for DALL·E in this test. Another reason is the difference in working mechanism between the DALL·E model and designers, in which DALL·E takes text as input and generates various images based on random noise while designers are skilled in producing various images using divergent thinking. It is noticed that two to three groups in the machine data have higher scores regarding all four metrics, indicating the capability of producing combinational creativity images between novice designers and DALL·E is not significantly different.

6.4. Overall Discussion

There are no clear criteria to determine whether DALL·E passes the Turing test, but it can be concluded that DALL·E's performance is close to novice designers according to the results of our Turing test. In the computational test, DALL·E outperforms designers in terms of IS but loses to designers regarding FID, and the difference in values is both small, indicating that the performance between DALL·E and novice designers is very close. It is noticed that the results of IS and FID are in conflict, which indicates that the effectiveness of the two metrics for evaluating combinational creativity needs to be further investigated. A larger difference in FIDs in comparison with our reference data implies that human designers are better at synthesizing features from base and additive for a combinational design. According to the results of the expert test, designers outperform DALL·E from the perspective of combinational creativity. There is slight advance for designers regarding novelty, feasibility and creativity completeness, but evident advance regarding variety. By summarizing the conclusions from the three tests in this study, DALL·E's performance is no better than novice designers but the gap is small.

There are two key directions for future research. There is little research on evaluating computational creativity. In this study, we applied three common methods from different areas to evaluate the performance of DALL·E and compare it with novice designers, which are labour intensive and lack scalability. How to effectively and systematically evaluate computational algorithms in generating creative ideas or stimuli needs further investigation and research. Another direction is the application of DALL·E or other similar techniques in design, particularly in conceptual design. Design is a process of transforming requirements and ideas into realisation, while DALL·E has the capability of transforming an idea described in texts into a conceptual design solution visualized in images. This would potentially provide a mental leap for designers, particularly novices, facilitating creative idea generation.

There are a few limitations in this study. First, eight sets of data related to combinational creativity, containing forty machine generated images and forty human generated ones, were used in the study for evaluation. The limited amount of data was a result of the restricted access to DALL·E's source code and data, as well as the high cost of human resources. Although the amount of data is sufficient for the purpose of the study, more data will be included in future studies by recruiting more human designers and accessing more DALL·E data to yield further useful insights. This would require the involvement of more human designers and accessing more DALL·E's data.

Second, one hour was provided to the designers to complete one combinational creativity design task to construct the human dataset, but it is still far less to produce a high-quality image. More time will be provided to the participants in future research to improve the quality of the images generated. Third, DALL·E is a deep learning model mainly aiming at transforming texts into images rather than generating combinational creativity, which is less fair to compare with human designers. In future research, more advanced artificial intelligence models, such as ChatGPT and GPT-4, will be included in the comparison.

7. Conclusion

This paper is the first research that has explored the comparison of combinational creativity capability between human beings and computers. It starts with the preparation of two datasets, the machine dataset is created by collecting data from a computational system, DALL·E, and the human dataset is created by inviting novice designers to produce images based on textual combinational ideas. Three tests, including a Turing test, a computational test and an expert test, are designed and implemented on the two datasets. The results of the three tests reveal that DALL·E's performance is very close to novice designers, while human designers are better at synthesizing features from the base and the additive for a combinational design. The results provide some useful insights for supporting the development of next-generation computational systems to aid creative idea generation. The study represents a contribution to the body of knowledge in research on computational methods for design. It leads towards new research directions in evaluating computational creativity and applying advanced computational techniques, particularly in conceptual design.

FUNDING

This is paper is funded by the National Natural Science Foundation of China [62207023] and the Ng Teng Fong Charitable Foundation in the form of ZJU-SUTD (Zhejiang University – Singapore University of Technology and Design) IDEA Grant.

REFERENCES

[1] Childs, P., J. Han, L. Chen, P. Jiang, P. Wang, D. Park, Y. Yin, E. Dieckmann, and I. Vilanova, The Creativity Diamond - A Framework to Aid Creativity. Journal of Intelligence, 2022. **10**(4): p. 73

[2] Amabile, T.M., The Social Psychology of Creativity. 1983, New York: Springer.

[3] Shute, V.J. and S. Rahimi, Stealth assessment of creativity in a physics video game. Computers in Human Behavior, 2021. **116**: p. 106647.DOI: https://doi.org/10.1016/j.chb.2020.106647

[4] De Bono, E., Six Thinking Hats. 1985: Little, Brown.

[5] Eberle, B., Scamper: Games for Imagination Development. 1996: Prufrock Press.

[6] Zwicky, F., Discovery, Invention, Research Through the Morphological Approach. 1969: Macmillan.

[7] Altshuller, G.S., Creativity as an exact science: The theory of the solution of inventive problems. 1984, Amsterdam, Netherlands: Gordon and Breach Publishers.

[8] Linsey, J.S., A.B. Markman, and K.L. Wood, Design by Analogy: A Study of the WordTree Method for Problem Re-Representation. Journal of Mechanical Design, 2012. **134**(4).DOI: 10.1115/1.4006145

[9] Yilmaz, S., S.R. Daly, C.M. Seifert, and R. Gonzalez, Evidence-based design heuristics for idea generation. Design Studies, 2016. **46**(Supplement C): p. 95-124.DOI: https://doi.org/10.1016/j.destud.2016.05.001

[10] Helms, M., S.S. Vattam, and A.K. Goel, Biologically inspired design: process and products. Design Studies, 2009. **30**(5): p. 606-622.DOI: https://doi.org/10.1016/j.destud.2009.04.003

[11] Chakrabarti, A. and L.H. Shu, Biologically inspired design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 2010. **24**(4): p. 453-454.DOI: 10.1017/S0890060410000326

[12] Oman, S.K., I.Y. Tumer, K. Wood, and C. Seepersad, A comparison of creativity and innovation metrics and sample validation through in-class design projects. Research in Engineering Design, 2013. **24**(1): p. 65-92.DOI: 10.1007/s00163-012-0138-9

[13] Han, J., F. Shi, L. Chen, and P.R.N. Childs, A computational tool for creative idea generation based on analogical reasoning and ontology. Artificial Intelligence for

Engineering Design, Analysis and Manufacturing, 2018. **32**(4): p. 462-477.DOI: 10.1017/S0890060418000082

[14] Sarica, S., J. Luo, and K.L. Wood, TechNet: Technology semantic network based on patent data. Expert Systems with Applications, 2020. **142**: p. 112995.DOI: https://doi.org/10.1016/j.eswa.2019.112995

[15] Siddharth, L., L.T.M. Blessing, K.L. Wood, and J. Luo, Engineering Knowledge Graph From Patent Database. Journal of Computing and Information Science in Engineering, 2021. **22**(2).DOI: 10.1115/1.4052293

[16] Obieke, C.C., J. Milisavljevic-Syed, A. Silva, and J. Han, A Computational Approach to Identifying Engineering Design Problems. Journal of Mechanical Design, 2023. **145**(4).DOI: 10.1115/1.4056496

[17] Boden, M.A., The creative mind: Myths and mechanisms. 2 ed. 2004, London, UK: Routledge.

[18] Simonton, D.K., Domain-general creativity: On Generating Original, Useful, and Surprising Combinations, in *The Cambridge Handbook Of Creativity across Domains*, Kaufman J.C., Glaveanu V.P., and B. J., Editors. 2017, The Cambridge University Press.: Cambridge, UK. p. 18-40.

[19] Han, J., F. Shi, L. Chen, and P.R.N. Childs, The Combinator – a computer-based tool for creative idea generation based on a simulation approach. Design Science, 2018. **4**: p. e11.DOI: 10.1017/dsj.2018.7

[20] Garvey, B., L. Chen, F. Shi, J. Han, and P.R. Childs, New directions in computational, combinational and structural creativity. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2019. **233**(2): p. 425-431.DOI: 10.1177/0954406218769919

[21] Beaty, R.E. and D.R. Johnson, Automating creativity assessment with SemDis: An open platform for computing semantic distance. Behavior Research Methods, 2021. **53**(2): p. 757-780.DOI: 10.3758/s13428-020-01453-w

[22] Besemer, S.P. and K. O'quin, Analyzing Creative Products: Refinement and Test of a Judging Instrument. Journal of Creative Behavior, 1986. **20**: p. 115-126

[23] Horn, D. and G. Salvendy, Product creativity: conceptual model, measurement and characteristics. Theoretical Issues in Ergonomics Science, 2006. **7**(4): p. 395-412.DOI: 10.1080/14639220500078195

[24] Cropley, D. and A. Cropley, Engineering Creativity: A Systems Concept of Functional Creativity, in *Creativity across domains: Faces of the muse.* 2005, Lawrence Erlbaum Associates Publishers: Mahwah, NJ, US. p. 169-185.

[25] Shah, J.J., S.M. Smith, and N. Vargas-Hernandez, Metrics for measuring ideation effectiveness. Design Studies, 2003. **24**(2): p. 111-134.DOI: https://doi.org/10.1016/S0142-694X(02)00034-0

[26] Han, J., H. Forbes, and D. Schaefer, An exploration of how creativity, functionality, and aesthetics are related in design. Research in Engineering Design, 2021. **32**(3): p. 289-307.DOI: 10.1007/s00163-021-00366-9

[27] Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, Improved training of wasserstein gans. In Advances in neural information processing systems. 2017

[28] Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 2017. **30**

[29] Ward, T.B. and Y. Kolomyts, Cognition and creativity, in *The Cambridge handbook of creativity*, J.C. Kaufman and R.J. Sternberg, Editors. 2010, The Cambridge University Press: Cambridge, UK. p. 93-112.

[30] Yang, H. and L. Zhang, Promoting Creative Computing: origin, scope, research and applications. Digital Communications and Networks, 2016. **2**(2): p. 84-91.DOI: https://doi.org/10.1016/j.dcan.2016.02.001

[31] Nagai, Y., T. Taura, and F. Mukai, Concept blending and dissimilarity: Factors for creative concept generation process. Design Studies, 2009. **30**(6): p. 648-675.DOI: 10.1016/j.destud.2009.05.004

[32] Han, J., F. Shi, D. Park, L. Chen, and P. Childs. The conceptual distances between ideas in combinational creativity. in DS92: Proceedings of the DESIGN 2018 15th International Design Conference. 2018.

[33] Han, J., D. Park, F. Shi, L. Chen, M. Hua, and P.R. Childs, Three driven approaches to combinational creativity: Problem-, similarity- and inspiration-driven. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2019. **233**(2): p. 373-384.DOI: 10.1177/0954406217750189

[34] Chen, L., P. Wang, H. Dong, F. Shi, J. Han, Y. Guo, P.R.N. Childs, J. Xiao, and C. Wu, An artificial intelligence based data-driven approach for design ideation. Journal of Visual Communication and Image Representation, 2019. **61**: p. 10-22.DOI: https://doi.org/10.1016/j.jvcir.2019.02.009

[35] Chen, L., P. Wang, F. Shi, J. Han, and P. Childs. A computational approach for combinational creativity in design. in DS 92: Proceedings of the DESIGN 2018 15th International Design Conference. 2018.

[36] Qiao, T., J. Zhang, D. Xu, and D. Tao, Learn, imagine and create: Text-to-image generation from prior knowledge. Advances in Neural Information Processing Systems, 2019. **32**: p. 887-897

[37] Hu, K., W. Liao, M.Y. Yang, and B. Rosenhahn, Text to Image Generation with Semantic-Spatial Aware GAN. arXiv preprint arXiv:2104.00567, 2021

[38] Ramesh, A., M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092, 2021

[39] Brown, T.B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020

[40] Ramesh, A., M. Pavlov, G. Goh, and S. Gray, *DALL·E: Creating Images from Text*. 2021, OpenAI.

[41] Turing, A.M., Computing Machinery and Intelligence, in *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, R. Epstein, G. Roberts, and G. Beber, Editors. 2009, Springer Netherlands: Dordrecht. p. 23-65.

[42] Boden, M.A., The Turing test and artistic creativity. Kybernetes, 2010. **39**(3): p. 409-413.DOI: 10.1108/03684921011036132

[43] Pease, A. and S. Colton. On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal. Citeseer.

[44] Peter Berrar, D. and A. Schuster, Computing machinery and creativity: lessons learned from the Turing test. Kybernetes, 2014. **43**(1): p. 82-91.DOI: 10.1108/K-08-2013-0175

[45] Doersch, C., Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908, 2016

[46] Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets. Advances in neural information processing systems, 2014. **27**

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

[48] Zhu, M., P. Pan, W. Chen, and Y. Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[49] Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick. Microsoft coco: Common objects in context. in European conference on computer vision. 2014. Springer.

[50] Sarica, S., J. Han, and J. Luo, Design representation as semantic networks. Computers in Industry, 2023. **144**: p. 103791.DOI: https://doi.org/10.1016/j.compind.2022.103791

[51] Zhu, Q., X. Zhang, and J. Luo, Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers. Journal of Mechanical Design, 2023. **145**(4).DOI: 10.1115/1.4056598

Figure Captions List

Figure 1	The workflow of the proposed research approach
Figure 2	A question webpage in the Turing test
Figure 3	Webpages of two question examples in the expert test
Figure 4	The IS values of different test groups
Figure 5	The FID scores of different test groups
Figure 6	The FID scores in comparison with divided reference groups
Figure 7	The base-FIDs and additive FIDs comparison within the machine and human datasets
Figure 8	The confusion matrix of Turing test results

Table Caption List

Table 1 An overview of the machine and human data

Table 2 An overview of the reference data

Table 3 Results of the Turing test

Table 4 Variance of accuracy in different groups

Table 5 Results of expert test

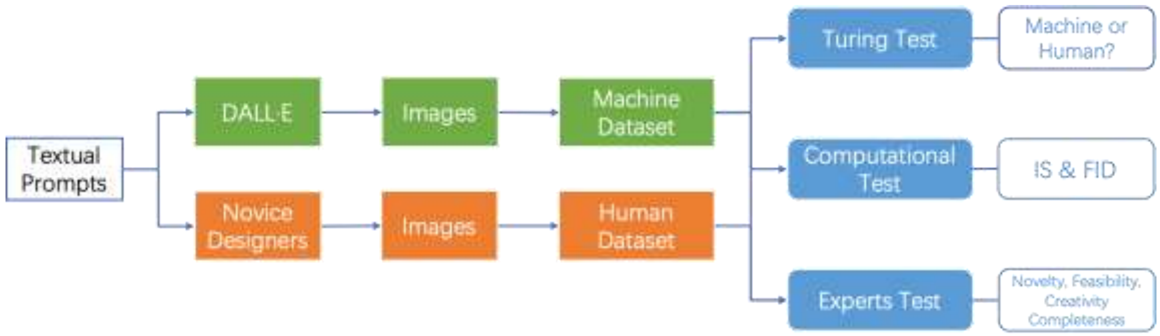
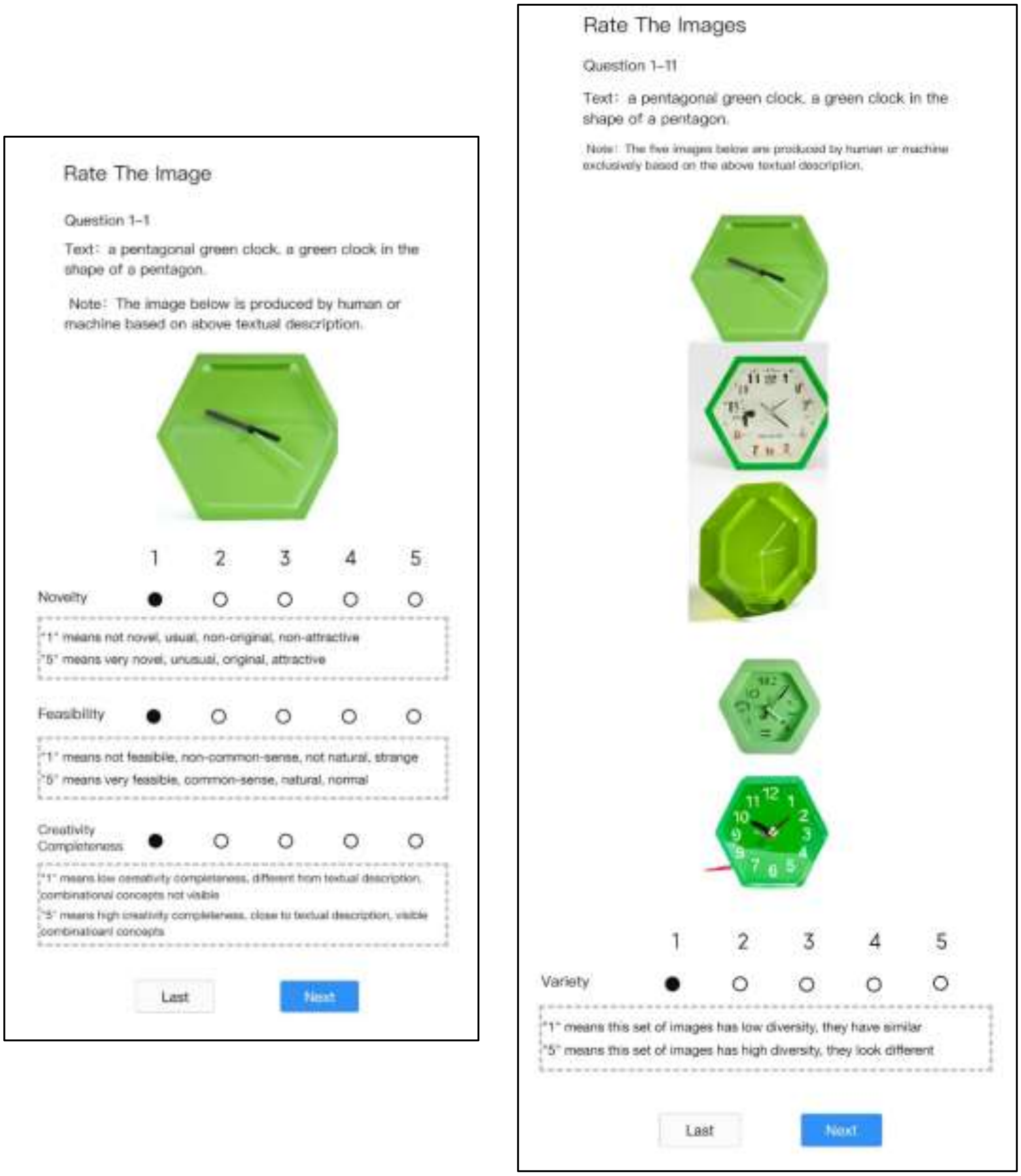


Figure 1. The workflow of the proposed research approach

Figure 2. A question webpage in the Turing test



(a) (b)

861 Figure 3. Webpages of two question examples in the expert test

862

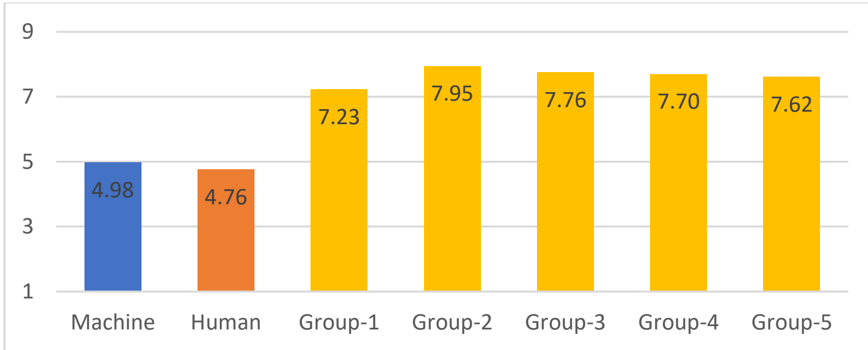


Figure 4. The IS values of different test groups

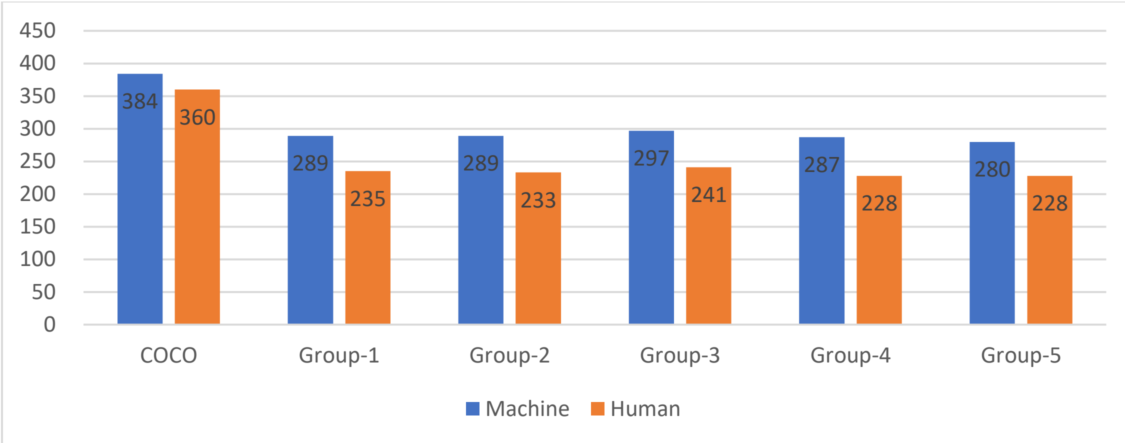


Figure 5. The FID scores of different test groups

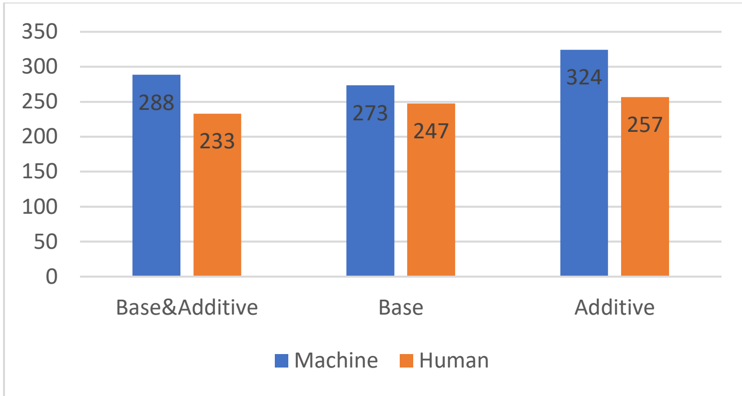
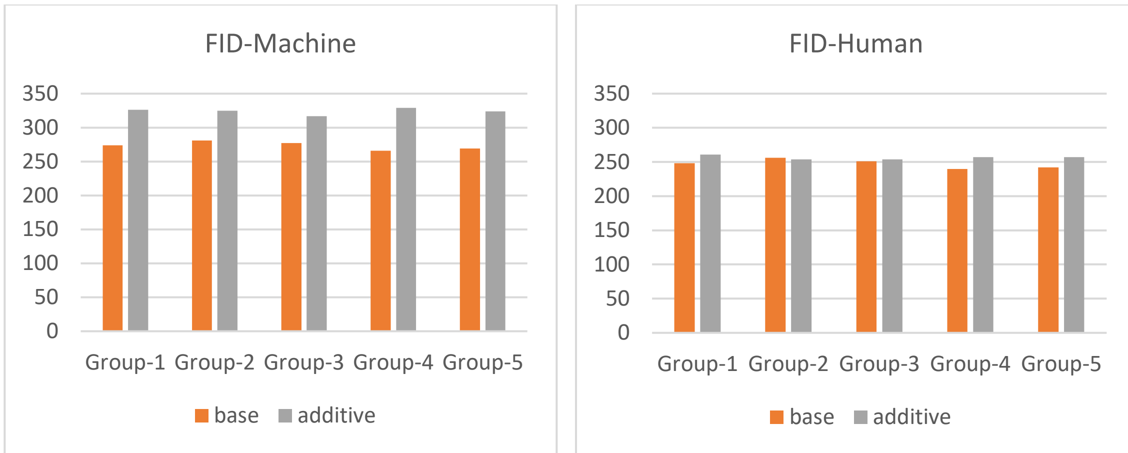


Figure 6. The FID scores in comparison with divided reference groups



(a) (b)

Figure 7. The base-FIDs and additive FIDs comparison within the machine (a) and human (b) datasets

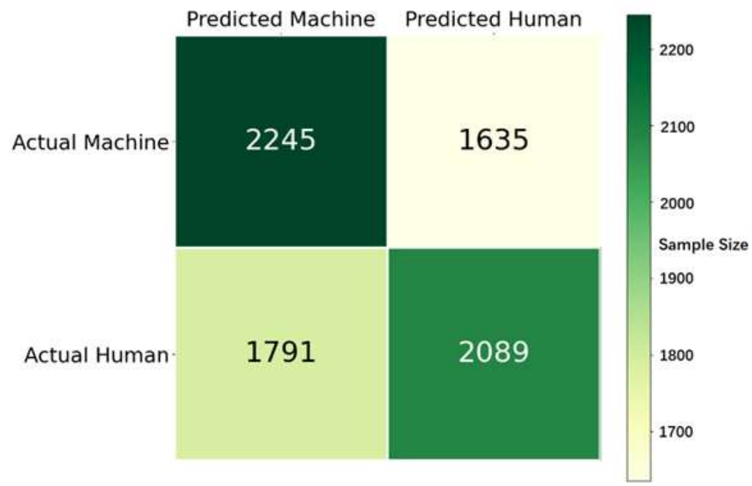


Figure 8. The confusion matrix of Turing test results

Table 1. An overview of the machine and human data

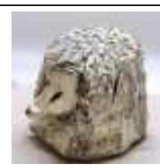
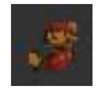
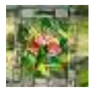
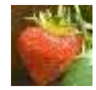
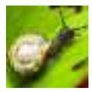
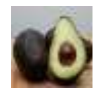
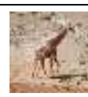
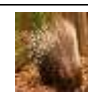
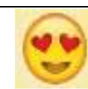
Group No.	Input	Machine Output	Human Output
1	a pentagonal green clock. a green clock in the shape of a pentagon		
2	a capybara made of voxels sitting in the field		
3	a stained-glass window with an image of a blue strawberry		
4	a snail made of harp. A snail with the texture of a harp		
5	an armchair in the shape of an avocado. an armchair imitating an avocado		
6	a giraffe imitating a turtle. a giraffe made of turtle		
7	a cube made of porcupine. a cube with the texture of a porcupine		
8	a professional high-quality emoji of a lovestruck cup of boba		

Table 2. An overview of the reference data

Group	Base	Sample-Base	Additive	Sample-Additive
1	Clock		Pentagonal	
2	Capybara		Voxels	
3	Glass		Strawberry	
4	Snail		Harp	
5	Armchair		Avocado	
6	Giraffe		Turtle	
7	Cube		Porcupine	
8	Cup		Emoji of lovestruck	

891 Table 3. Results of the Turing test

		Mean	1	2	3	4	5	6	7	8
Accuracy		55.9%	60.8%	55.2%	61.9%	60.3%	54.2%	51.1%	61.1%	42.2%
Machine	Precision	55.6%	60.6%	54.4%	63.5%	60.8%	54.2%	51.1%	59.0%	42.2%
	Recall	57.9%	61.9%	64.1%	55.7%	57.9%	54.2%	53.4%	73.0%	42.7%
	F1	56.7%	61.2%	58.8%	59.3%	59.3%	54.2%	52.2%	65.3%	42.5%
Human	Precision	56.1%	61.1%	56.3%	60.6%	59.8%	54.2%	51.2%	64.6%	42.1%
	Recall	53.8%	59.8%	46.2%	68.0%	62.7%	54.2%	48.9%	49.3%	41.6%
	F1	54.9%	60.4%	50.7%	64.1%	61.2%	54.2%	50.0%	55.9%	41.9%

892

893

894

Table 4. Variance of accuracy in different groups

		Min	Max	Max-Min	Difference
1	Human	37.1%	90.7%	53.6%	23.7%
	Machine	51.5%	81.4%	29.9%	
2	Human	19.6%	74.2%	54.6%	18.6%
	Machine	40.2%	76.3%	36.1%	
3	Human	43.3%	79.4%	36.1%	-2.1%
	Machine	38.1%	76.3%	38.1%	
4	Human	45.4%	93.8%	48.5%	26.8%
	Machine	46.4%	68.0%	21.6%	
5	Human	16.5%	89.7%	73.2%	45.4%
	Machine	43.3%	71.1%	27.8%	
6	Human	27.8%	77.3%	49.5%	7.2%
	Machine	32.0%	74.2%	42.3%	
7	Human	40.2%	63.9%	23.7%	-5.2%
	Machine	57.7%	86.6%	28.9%	
8	Human	25.8%	69.1%	43.3%	21.6%
	Machine	26.8%	48.5%	21.6%	
Overall	Human	16.5%	93.8%	77.3%	17.5%
	Machine	26.8%	86.6%	59.8%	
Mean	Human	32.0%	79.8%	47.8%	17.0%
	Machine	42.0%	72.8%	30.8%	

Table 5. Results of expert test

Metrics	**Data Origin**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Mean**	**Variance**
Novelty	Machine	2.38	2.58	2.97	**3.23**	**3.43**	2.48	2.43	**2.76**	2.78	0.39
	Human	**2.89**	**2.83**	**3.15**	3.04	2.86	**2.74**	**2.94**	2.75	**2.90**	0.14
Feasibility	Machine	3.80	2.88	3.03	**3.40**	**3.68**	2.49	2.93	**3.58**	3.23	0.46
	Human	**3.88**	**3.48**	**3.92**	3.09	3.46	**3.01**	**3.04**	3.35	**3.41**	0.36
Completeness	Machine	3.48	2.68	2.98	**3.57**	3.42	2.44	2.54	**3.64**	3.09	0.49
	Human	**3.53**	**3.06**	**3.64**	3.46	**3.48**	**3.40**	**2.89**	3.44	**3.36**	0.25
Variety	Machine	3.00	2.74	2.37	**3.42**	**3.47**	3.16	**3.26**	2.21	2.95	0.47
	Human	**3.84**	**4.37**	**3.84**	2.89	3.05	**3.21**	3.21	**3.74**	**3.52**	0.50