A Creativity Assessment Scale for Text-to-Image Prompting: Challenges & Observations

Sophia Lichtenberg, Almila Akdag
Department of Information and Computing Sciences, Utrecht University, Netherlands
{s.lichtenberg, a.a.akdag}@uu.nl

Abstract

Online platforms offering Text-to-Image (TTI) generation services are dynamic environments where the creative process of ideation and verbal skills of the users can be observed over time. User interactions on these platforms produce extensive behavioral data, encompassing patterns of engagement, verbal expression, visual imagination, and prompt engineering, a specialized skill essential for image generation. Consequently, these platforms provide a unique opportunity to examine the verbal creativity of user's prompts. However, the creativity assessment literature is restricted to specific conditions such as strictly controlled user studies, or product-based assessments that are not generalizable. As such, there is an obvious need to create an assessment method that fits the prompt language to generate images. In this study, we provide a framework for prompt creativity assessment in the format of an online tool. The tool is developed based on the Creativity Product Semantic Scale (CPSS). We tested this framework with eight trained annotators, analyzing prompt samples from Midjourney and Stable Diffusion. Our evaluation examines how user engagement, measured by time spent, amount of prompts, and prompt length, affects overall creativity scores.

Introduction

Text-to-Image (TTI) models have emerged as a transformative advancement within the creative industries, facilitating the generation of high-quality visual output from textual descriptions. Models such as DALL-E, Midjourney, Stable Diffusion, FLUX, Adobe Firefly and Ideogram enable rapid image synthesis. This shows a shift from traditional digital to AI-assisted creative tools that has also raised critical questions about the nature of creativity, authorship, and how AI-driven tools enhance or limit human ideation. Instead of just automating image creation, TTI models allow humans to collaborate with AI by transforming their ideas into text prompts that guide image generation. This interaction produces outputs that blur the line between human and machine creativity.

Despite their advantages, TTI models pose significant risks, such as biases embedded in training data that influence the generated images, privacy and copyright concerns, and the potential displacement of jobs within creative industries (Naik and Nushi, 2023; de Almeida and Rafael, 2024; Katirai et al., 2023; Bird, Ungless, and Kasirzadeh, 2023). These models can be misused to generate harmful content, including misinformation, deepfakes, and abusive material, raising ethical concerns about their societal impact (Wei et al., 2024; Shalabi et al., 2023; Hao et al., 2024; Sha et al., 2023; Bird, Ungless, and Kasirzadeh, 2023).

The interconnection between users and TTI systems has opened new opportunities for TTI specific creativity research, but still lacks the fundamental question if TTI prompts have creative potential. Analysis of TTI datasets reveals large-scale user behavior trends, but often fails to capture the subjective experience of creativity (Wang et al., 2022; McCormack et al., 2024; Chen and Zou, 2023; Sun et al., 2023; Palmini, Wagner, and Cetinic, 2024). However, surveys and interviews with TTI artists and users provide qualitative insights into how artists integrate TTI tools into their workflows (Goloujeh, Sullivan, and Magerko, 2024; Sanchez, 2023). Furthermore, controlled user studies have examined prompt formulation as an iterative process, highlighting how users experiment with word choice, specificity, and abstraction to achieve desired results (Chang et al., 2023; Trinh et al., 2024; Oppenlaender, Linder, and Silvennoinen, 2024). These studies contribute to an expanding body of literature documenting TTI's role in new artistic processes. However, despite these contributions, no framework exists for evaluating the creativity of prompts, therefore we propose and evaluate a TTI Prompt Creativity evaluation scale.

Despite recent advances, existing research has not yet established a clear connection between definitions of creativity and the corresponding assessment frameworks in the context of TTI systems. In particular, there is no standardized metric for evaluating the creativity of prompts. Advancing this area of study requires the development of novel assessment methodologies that integrate established creativity constructs into the domain of prompt engineering, thereby enabling a more nuanced understanding of creativity within the context of generative AI.

To address this limitation in the existing literature, our study has three key contributions: 1) We introduce the Prompt Creativity Scale (PCS), an online tool designed to assess prompt creativity in a systematic way. This framework is based on the Creative Product Semantic Scale

(CPSS) (Besemer and O'Quin, 1986), adapting its dimensions to four crucial aspects of prompt engineering: *Style*, *Content*, *Idea Combination* and *Prompt Writing*. 2) We conduct a pilot test with eight annotators. The results are analyzed using qualitative and statistical methods that evaluate consistency, robustness, and annotator agreements. 3) We analyze two prompt datasets (Stable Diffusion, Wang et al. (2022) and Midjourney, McCormack et al. (2024)) to examine user engagement, categorizing users into four distinct groups based on their activity. We curate a balanced sample from these groups and evaluate differences in their creative dimensions, exploring correlations between skill and creativity.

Related Work

Creativity is a subjective, complex, and context-dependent concept, making its evaluation inherently challenging. The generally agreed definition focuses on two fundamental components, as described by Runco and Jaeger (2012): originality and effectiveness. Originality refers to the novelty or uniqueness of an idea, while effectiveness refers to its usefulness, appropriateness, or value within a given context. Beyond these core components, several influential factors have been identified that contribute to the creative process (Amabile, 1983; Urban, 1991; Runco and Chand, 1995). Due to context dependency, various fields, such as psychology, computational creativity, and the arts, are using distinct frameworks for assessment. In the context of TTI prompt engineering, the lack of a domain-specific framework is further complicated by the interplay between human intention and the actual semantic interpretation of the model, raising important questions about how creative input translates into generative output. To clarify these differences, this section reviews key approaches to evaluating creativity and explores their relevance to prompt assessment. It also examines existing research on prompt engineering, highlighting various methodological challenges.

Creativity Assessment Approaches

Traditional Creativity assessment methods generally fall into four categories: performance tests, self-reports, expert evaluations, and creative product evaluations.

Performance-based assessments measure divergent thinking through tests such as the Alternative Uses Task (AUT) (Guilford, 1967), the Torrance Test of Creative Thinking (TTCT) (Torrance, 1966), and the Wallach and Kogan tests (Wallach and Kogan, 1965), which all evaluate originality, flexibility, fluency, and elaboration. Although these tests are simple to administer and quantifiable metrics, they are designed for structured examination with narrow focus and scope and can show cultural bias Kim (2006). Remote Associates Test (RAT) (Mednick, 1962) measures convergent thinking, while problem-solving tasks such as the Candle Problem (Duncker and Lees, 1945) and Maier's Two-String Problem (Maier, 1931) assess functional fixedness. These tasks emphasize the identification of fixed solutions or overcoming constraints, but have limited ability to assess broader, real-world creativity and generalizable problem-solving skills. In the arts, assessments such as the Creative Thinking-Drawing Production (TCT-DP) (Urban and Jellen, 1996) and Clark's Drawing Abilities Test (CDAT) (Clark, 2004) evaluate free drawing from imagination. The Test of Figural Combination (Finke, 2014) involves sketching creative objects based on given images. These assessments are limited to drawing abilities and do not capture creativity expressed through other mediums. The second category, self-reports measure individuals' perceptions of their creativity, with tools such as the Creative Achievement Questionnaire (CAQ) (Carson, Peterson, and Higgins, 2005) and the Runco Ideational Behavior Scale (RIBS) (Runco, Plucker, and Lim, 2001). The CAQ is susceptible to self-report and self-enhancement bias (Carson, Peterson, and Higgins, 2005; Rosenman, Tennekoon, and Hill, 2011), as individuals may deviate from accurate self-assessment in administered settings.

Expert evaluations are often conducted using the Consensual Assessment Technique (CAT) (Amabile, 1983), where independent judges rate creative artifacts without predefined criteria, ensuring high inter-rater reliability and stability (Baer, 1994). Research suggests that even non-experts can use this method to reliably assess creativity (Hennessey, 1994). However, CAT is highly dependent on subjectivity, expert consensus, and expert selection, which can introduce bias. Objective product assessments use explicit criteria to evaluate creativity. Creative product evaluations, such as the Creative Product Inventory (CPI) (Taylor and Sandler, 1972) consider originality and effectiveness, while the Creative Product Analysis Matrix (CPAM) (Besemer, 1998) and the Creative Product Semantic Scale (CPSS) (Besemer and O'Quin, 1986, 1987, 1989) both categorize works by novelty, resolution, and elaboration. The Creative Solution Diagnosis Scale (CSDS) (Cropley and Kaufman, 2012) focuses on functional creativity, assessing relevance, novelty, elegance, and emergence. Although product assessments offer structured, consistent, and quantifiable evaluations based on clear criteria, they can also oversimplify creativity by evaluating solely the final product and therefore lacking the flexibility to capture the full complexity of creative work.

Overall, traditional assessments do not account for the complex interplay between users' ideation and TTI generation. They lack the standardization needed for evaluating the creative potential of human input, the foundational idea, which guides and defines the TTI models' image output. This is because the effectiveness of the image depends not only on the TTI model but also on the creativity and clarity of the prompt provided by the user. Franceschelli and Musolesi (2024) classifies TTI models as an example of exploratory creativity, as they sample from latent space based on input prompts. Although these models generate diverse outputs, value, novelty, and surprise are not guaranteed and typically stem from the conditioning input, not the model itself. Since the prompt itself provides limited information about the user's ideation process or their personal creativity, self-report, and performance-based assessment methods cannot be applied. The CAT relies on expert judgment, but this approach is challenging given that TTI prompt engineering is still a relatively new creative domain and not widely acknowledged. A prompt is not merely an instruction, but

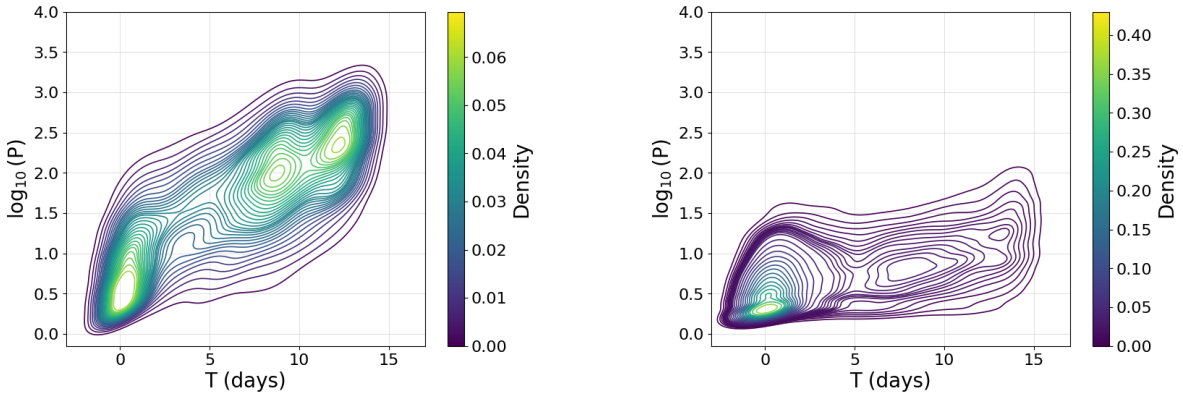

Figure 1: Kernel Density Estimation (KDE) distribution plotted for DiffusionDB (left) and Midjourney2023 (right), where T represents the total timespan in days and P denotes the prompt count.

a carefully constructed tool shaped by human thought, refinement, and skill, guiding the creative process and shaping the final output. Given the real-world prompts collected for analysis, we base our scale on the CPSS, which has proven to be reliable for assessing creative products. It can also be evaluated by non-expert judges and has been widely applied in the field.

TTI Prompt Engineering

Prompt engineering refers to the skill of creating a good TTI prompt, i.e. writing textual descriptions or instructions guiding the generative AI models to create images.

The image quality of the early TTI models was heavily dependent on the clarity, specificity, and user's expertise in formulating the prompts. Although recent TTI models have significantly improved visual language comprehension and prompt interpretation, making high-quality image generation more accessible, the task itself continues to pose considerable challenges.

Popular techniques include: creating and iteratively refining prompts, utilizing prompt templates, experimentation with latent space and modal constraints, use of specific (domain) vocabulary, leveraging prompt modifiers to achieve unique visual styles, oftentimes through a process of trial and error and shared knowledge, and by engaging with a co-creative ecosystem. Prompt communities often collaborate online, sharing prompts, using resources, and advice for effective prompt development. Prompt generation has been emphasized as the critical role in creating high-quality AI-driven art (Chang et al., 2023).

A prompt specifies diverse characteristic features for the desired output, such as themes, styles, or specific visual features. Oppenlaender (2023) introduced a taxonomy of prompt modifiers: technical term, style modifier, note, image prompt, quality improver, repeat term, and the magic term. A magic term adds a semantic deviation from the rest of the prompt, with the intention of surprising the user. Repetitive terms, solidifiers, can be any other type of modifier. Oppenlaender, Linder, and Silvennoinen (2024) discuss the lack of intuitiveness in prompt engineering practice and the potential for misalignment between human-written prompts and the way TTI models interpret prompts. Generative models can assign very different meanings to some keywords in the prompt due to their latent associations. The authors conducted three studies with Amazon Mechanical Turk participants to test their understanding of prompt quality by evaluating the aesthetic appeal of prompts, the ability to write prompts, and finally the ability to revise their own prompts. The results show that the participants are able to understand what makes a good prompt and distinguish it from bad prompts. However, participants were not able to add modifiers, stylistic vocabulary, and keywords, or improve the quality of the artwork through modifications suggesting that prompting is not intuitive but is a skill. Prompt engineering thus offers itself as an interesting concept to understand creativity in action. If building prompting skills enhances creativity, then we expect to see differences between the prompts of most prolific users, and the users rarely engage with these platforms.

Methodology

In this section, we first explain how we prepare a prompt dataset for evaluation. The process is follows three steps: We sample prompts from two datasets from the StableDiffusion and Midjourney Discord Channels. We process these datasets by looking into user activity and prompt features, and describe our processing and sampling strategy. Second, we introduce the PCS by defining its categories and dimensions and outlining its experimental and assessment setup.

Data

DiffusionDB-2M[1] contains 2,000,000 data entries collected in a time span of two weeks in Summer 2022 (Wang et al., 2022). Midjourney2023[2] (McCormack et al., 2024) contains 973,134 data entries from 24,487 unique users, and is

[1] https://huggingface.co/datasets/poloclub/diffusiondb

[2] https://bridges.monash.edu/articles/dataset/Midjourney_2023_Dataset/25038404

Dataset	User Group	# Users	# Prompts	# Days	Prompts:User	Prompts:User:Day
Midjourney23	Infrequent (33%)	5,257	17,425	1.02	3	3
	Moderate (33%)	5,257	26,462	1.95	5	3
	Prolific (33%)	5,257	103,417	3.98	20	5
	Extreme (1%)	160	25,768	9.18	161	18
DiffusionDB	Infrequent (33%)	3,185	36,628	1.94	12	6
	Moderate (33%)	3,185	187,996	5.12	59	12
	Prolific (32%)	3,185	1,107,495	9.62	348	36
	Extreme (1%)	97	182,978	12.97	1886	145

Table 1: User Groups in the DiffusionDB and Midjourney23 dataset. DiffusionDB contains more prompts per user while Midjourney23 has more users. For both, the prolific users contribute the most to the amount of prompts.

collected in a time span of two weeks in Fall 2023. For both datasets, the preprocessing removed deleted users, duplicated prompts, reducing the number of unique users for DB to 10,143 and the unique prompts to 1,528,514 and for the Midjourney23 this number is 24,011 unique users and 187,900 unique prompts. Both datasets contain information about the user, time, prompt, and the generated image. We decided to separate the users into four distinct groups based on their active time spent on Midjourney and StableDiffusion platforms and the amount they have prompted. Building such groups helps with understanding the relationship between becoming experts in prompt engineering and creative expression. The activity ratio A is calculated given the prompt count P and T the total timespan in hours as shown in 1.

$$A = T^{log(P)} \tag{1}$$

The Group G is assigned based on the Quantile Q of A as shown in 2.

$$G(A) = \begin{cases} Infrequent & \text{if } A < Q_{33} \\ Moderate & \text{if } Q_{33} \leq A < Q_{66} \\ Prolific & \text{if } Q_{66} \leq A < Q_{99} \\ Extreme & \text{if } A \geq Q_{99} \end{cases} \tag{2}$$

Table 1 shows the total number of users and prompts, and the average number of unique active days by group for each dataset. Midjourney23 has more extreme users, with fewer prompts and fewer active days compared to DiffusionDB. Figure 1 shows the KDE distribution of the activity value for all users in the dataset. There are peaks at the lower and upper ends for DiffusionDB, while Midjourney2023 concentrates more at the lower en,d since there are also fewer prompts per user.

Creating a sample Dataset for Creativity Evaluation To create a representative and varied set of prompts we first employed a pairwise cosine distance approach for preselection from DiffusionDB and Midjourney, which increased the likelihood of capturing a broader range of prompts compared to random selection. To prepare the dataset, we split the prompts into phrases by punctuation and stopwords. We discarded prompts with only one phrase. The processed prompts were embedded with the pretrained language model SBERT for ranking. From the top 10 highest-scoring results per user, samples were manually curated.

A total of 160 prompts were selected, with 80 prompts drawn from each datasets. Within the 80 prompts per dataset, the distribution was as follows: 14 prompts from Extreme users and 22 prompts each from Infrequent, Moderate, and Prolific users. To further balance the dataset, prompts were categorized into three classes based on the amount of tokens (i.e. word) length (excluding stopwords): Class 1 (up to 10 tokens), Class 2 (from 10 up to 25 tokens), and Class 3 (from 25 up to 50 tokens).
Finally, the entire dataset was divided into 8 sub-splits, each containing an equal number of prompts from both datasets to ensure balanced representation across all subsets.

Prompt Creativity Scale

Our framework is based on the design of the CPSS (Han et al., 2019) which uses a 7-point Likert scale table with 18 bipolar pairs of items referring to novelty and utility. These items are mixed and some are reversed to avoid evaluators' inertia. The creativity score is the sum of the scores of all the individual items. Adaptation of this scheme is done by using a 5-point Likert scale and a "not applicable" option, and altering the dimensions of the CPSS to fit the purpose of the TTI prompting context. The final PCS categories and their bipolar pairs as dimensions are shown in table 2. The arguments in choosing these dimensions are informed by creativity, cognitive sciences, aesthetics, and prompt engineering literature, as detailed in this section.

We started by highlighting the distinction between style and content for prompt writing. While familiarizing ourselves with the prompt dataset, we noticed the complexity and nuances within the prompts to describe the content but not the style, and vice versa. Differences in the perception of characteristics or content and styles of artworks of art are examined in the context of aesthetic perception. In art perception, described by Augustin et al. (2008); Leder and Nadal (2014), art differs not only in the way that subjects are depicted, i.e. the style, but also in the overall choice of subjects, i.e. the content. Style, as well as content are proposed as central variables when processing visual art (Leder et al., 2004; Leder and Nadal, 2014). Content is seen as central for classifying (Augustin, Leder, and others, 2006) and appreciating art, in particular for individuals without an artistic background (Hekkert and van Wieringen, 1996).

Categories	Bipolar Pairs	
Idea Combination	Unexpected few concepts far-related concepts visually imaginable high amount of visual ideas	Predicable high amount of concepts close-related concepts visually impossible few visual ideas
Content (Subjects, Scene, objects)	diverse elaborate original Superficial	homogeneous simple unoriginal Fully developed Depth
Style (Medium, Technique, Genre, Mood/Tone/Lightning, Artistic References, Perception)	diverse elaborate original Superficial	homogeneous simple unoriginal Fully developed Depth
Prompt Writing Style	Generic Vocabulary Easily Readable effective for TTI Model Contradictory	Domain Specific Vocabulary Difficult to Read not suited for TTI Model Coherent

Table 2: PCS: Main Categories and their bipolar pairs, assessed in a 5-point Likert scale

Style is a characteristic specific to art (Leder et al., 2004) essential for the distinctions in the perception of art compared to other forms of perception. According to Augustin et al. (2008), who analyzed the temporal aspects of style- and content-related processing, as well as the relationship between the two sub-processes in the perception of art, both style and content are interdependent and not easily separable categories. They evolve together as part of the same artistic process and both are shaped by historical, cognitive, and cultural factors. However, in TTI prompts, such a temporal dimension is lacking, and the separation between style and content is tangible. This disentanglement of style and content is a common part in training diffusion models (Wang, Zhao, and Xing, 2023; Kotovenko et al., 2019; Chung, Hyun, and Heo, 2024; Wu, Nakashima, and Garcia, 2023). Thus, we dedicate one category to the assessment of content and one to style.

Content refers to concepts, objects, subjects, and scenes described in the prompt (Dehouche and Dehouche, 2023). The dimensions address originality, superficiality, homogeneity, and the elaborateness of the descriptions.

Style addresses the visual descriptions in the prompts. Dehouche and Dehouche (2023) categorized style elements in TTI prompts: 1. Artistic references to use as inspiration, 2. Medium, i.e. "digital illustration", 3. Technique, i.e. traditional tools and software used to create the image i.e. "watercolor", 4. Genre such as "baroque", 5. Mood, i.e. features describing the atmosphere and emotions such as "beautiful", or "eerie", 6. Tone, referring to the color palete ("pastel", "synthwave colors"), 7. Lighting ("dark", "cinematic lighting"), and 8. Resolution, which describes the level of detail ("highly-detailed"). We use all these elements together for the assessment of the **Style** category. However, we address the combination of these elements within another dimension, i.e. the idea combination (see below our reasoning). The dimensions of this category are the same as listed in **Content**.

Idea Combination is based on Boden's (Boden and others, 1994; Boden, 1998, 2005, 2009, 2010) definition of computational creativity. Among the three types of creativity she described (i.e. exploratory, transformative and combinational) the first two only lead to surprise, but the latter leads to shock. This combinational aspect of creativity is also supported by Frigotto and Riccaboni (2011); Henriksen (2014); Han et al. (2018) and Ward and Kolomyts (2010). This category is crucial to evaluate how well idea combinations are done in a prompt, as this is a feature often observed in prompt descriptions. The dimensions are unexpectedness of the combination (surprise), amount of concepts combined (fluency), distance of concepts combined (novelty). To address the combinations of only stylistic elements, we also introduced two more dimensions: amount of visual ideas (fluency) and imagination, i.e. if the generated image is imaginable (originality).

Prompt Writing Style is the evaluation of the formulation of the prompt itself. Even though the prompting language does not follow conventional grammatical rules of natural language, it is still crucial that the ideas are well formulated and coherent. Vocabulary choice to communicate the ideas is judged by looking at the prompts' readability, coherence, (domain specific vs generic) vocabulary and its effectiveness to shape image generation. It is important to note that this dimension is assessing the clarity of the prompt, and not creativity.

PCS: Experiment Setup

The PCS is designed as an online tool[3]. The interface is a web application developed with Gradio and is hosted on Huggingface Spaces. The choice of embedding the PCS in

[3]https://www.projects.science.uu.nl/ics-promptannotations/

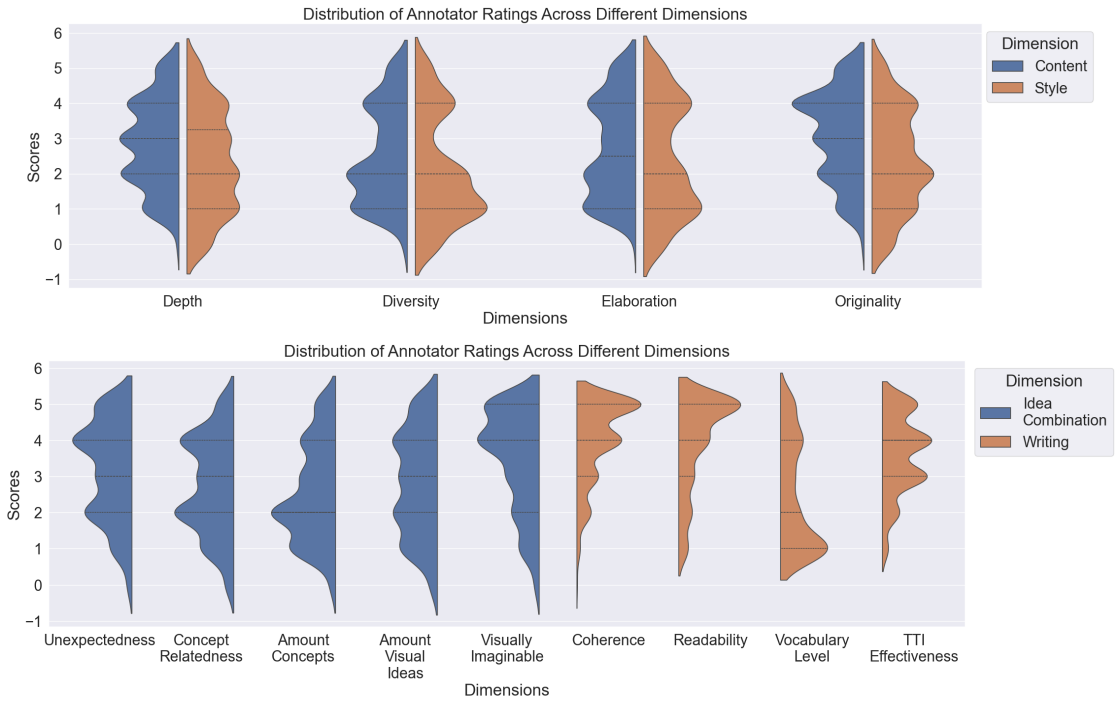

Figure 2: Score Distributions by main dimensions, Content and Style (upper figure), and Idea Combination and Prompt Writing (lower figure)

a web environment was deliberate and based on the reason that the scale is complex and not intuitive to use for the layman who is not familiar with prompt engineering examples. Thus, the interface has a dedicated page with examples and instructions, which includes a link to Stable Diffusion XL (SDXL)[4], allowing participants to gain prompting experience while becoming familiar with the PCS. The assessment page itself includes the same functionality, ie. the participants are given the choice of generating the prompts they assess with SDXL. However, this option is provided only at the very end of the page, after the assessment section. This placement de-emphasizes the function, as the evaluation is intended to focus on the prompt itself rather than the resulting (original) image.

Preliminary test

To test the web environment, the user interface (UI), the clarity of the PCS instructions, the examples, and the overall annotation experience, we conducted a preliminary test with four participants. Among the preliminary test participants (N=4), one had experience in UI design, two were familiar with the literature on creativity, two had some prior prompting experience, and two were laymen with no relevant background. After two rounds of trials, the following changes were applied: For each dimension, we add the option "not-applicable" next to the 1-5 Likert scale. This adjustment facilitated annotating prompts with no style or content elements, or where no idea combination was present. We prepared a dropdown menu with a list of phrases such as "I like this", "I did not like this", "Incomprehensible", "I had to look up artists' names", and "I had to look up words". The list is based on the topics on which participants stressed the need to provide additional feedback. The interface maintained the option to leave other feedback than the provided list.

Process

For the pilot, participants were introduced to the interface to familiarize themselves with the PCS and the annotation task through a training session. They were also given the option to annotate an example set to get used to the process, which is not included in our analysis. The annotation process is divided into three rounds, each consisting of 20 prompts. Participants were asked not to take a break before finishing a round and given a week to finish all three rounds. They were also instructed to look up artist names and unfamiliar words. Afterwards, annotators filled out a short survey for a qualitative analysis with three questions: "Were any items unclear or difficult to answer?", "Did any items seem repetitive or unnecessary?", and "Do you want to share other observations?". Each prompt is annotated by three participants to assess inter-annotator agreement in a robust manner. To ensure this, we ask annotators to register with the website and use their assigned ID every time they annotate. The data set is divided into eight subsets with 20 prompts. Each data split is assigned to three different annotators to compare each split with a different set of annotators.

Participants

For the pilot study, eight participants (N=8) from Utrecht University were recruited, comprising an equal number of

[4] https://huggingface.co/spaces/google/sdxl

Sample Split	Content		Idea Combination	Writing		Style	
	all	selection	all	all	selection	all	selection
1	0.7457	0.8224	0.9548	0.4232	0.2827	0.7248	0.6102
2	0.9382	0.9518	0.9634	0.8961	0.8779	0.9705	0.9796
3	0.7099	0.8359	0.8375	-4.0295	-1.1453	0.6874	0.7071
4	0.9367	0.9462	0.9288	0.8281	-2.8226	0.9612	0.9267
5	-20.4430	-0.7056	0.9024	0.8770	0.5116	0.8877	0.9387
6	-0.4286	-0.4198	0.8368	0.7330	0.7624	0.9548	0.9817
7	0.5898	0.4668	0.8464	-0.1942	-1.4289	0.9583	0.9556
8	0.9473	0.9489	0.2043	-2.3451	0.1324	0.4062	0.4883
mean	-2.0001	0.4808	0.8093	-0.3514	-0.3537	0.8189	0.8235
std	6.5841	0.5887	0.2203	1.6248	1.1562	0.1771	0.1707

Table 3: PCS reliability: Cronbach's α for each data split and each items within the same category across the three annotators within the data split. The mean and standard deviation is across all data splits. Including either all scales or the selection for Content, Style and Writing by removing the scales Depth and TTI Effectiveness

females and males. The sample included five Master's students, two PhD candidates, and one Assistant Professor. The background of the recruited participants varied, including Artificial Intelligence (n=6), Human-Computer Interaction (n=2), Natural Sciences (n=1) and Art History (n=1). All participants had prior prompting experience on Text-to-Text Tasks, six of them were knowledgeable on Text-to-Image Tasks and half of them indicated expertise on Image-to-Image tasks. All participants had experience with ChatGPT, five with Midjourney, half with Stable Diffusion, Gemini, and few on Adobe Firefly, Claude, DALL-E and NightCafe. Half of the participants indicated a regular prompting experience.

Assessing PCS

Annotator bias is assessed by calculating the mean and standard deviation for each annotator across all data splits and comparing these values. To evaluate the reliability of the PCS, Cronbach's α is applied to the categories to determine whether they measure the same underlying construct. Cronbach's α assesses the internal consistency of a set of items measuring the same latent variable. Given the small number of annotators, we also include a qualitative analysis from a brief survey after annotations. Furthermore, annotator agreement on the dimensions is measured using the Intraclass Correlation Coefficient (ICC). Specifically, we use ICC2k, as not all annotators assess every prompt in the dataset, and annotators were selected randomly.

Results

Here, we summarize our results in three sections. First, we provide the qualitative evaluation collected from the survey, as well as from the feedback given during the annotations. The second section examines the descriptive statistics of the scores for each annotator and dimensions. To ensure the robustness of our measurements, we evaluate the reliability of the PCS and assess inter-annotator agreement. Lastly, based on the results of the reliability and agreement assessments, we analyze our dataset to investigate whether specific features influence PCS scores. In particular, we examine score variations between different user groups and prompt lengths.

A	Mean	Std	Q_{25}	Q_{50}	Q_{75}	Data Splits
1	**2.4716**	1.4052	1	2	4	2,4,7
2	2.9137	1.4123	2	3	4	3,6,8
3	2.7078	**1.6759**	1	2	4	1,4,5
4	3.0186	1.4881	2	3	4	1,2,3
5	3.0716	**1.2247**	2	3	4	3,4,8
6	**3.1373**	1.3206	2	3	4	2,5,6
7	2.9078	1.5407	2	3	4	5,7,8
8	2.8235	1.5340	2	3	4	1,6,7

Table 4: Descriptive Statistics of Score distribution by each Annotator

Qualitative Evaluation

The most common feedback given during the annotations were "I am curious about the generated image" (19.55%), followed by "I had to look up artist name(s)" (18.41%) and "like" (18.41%), less common was "I had to look up word(s)" (8.86%), "dislike" (4.77%), "not understand" (2.05%), "incomprehensible" (1.82%), "exceptional" (0.91%).

The survey responses highlighted common struggles: Five Participants find the dimension **TTI Model Effectivity** under the category **Prompt Writing Style** hard to answer. **Content Depth** and **Style Depth** were found by three participants difficult to answer as well. Most of the participants (n=6) find no items redundant, and only two participants indicated **Content Depth** and **Style Depth** as redundant.

Quantitative Evaluation

Annotator Score Distributions Table 4 shows the descriptive statistics for the eight annotators. Most annotators were rather consistent in their score distribution, annotator 1 deviated the most towards lower values while annotator 6 deviated the most to the higher end and showed the highest inconsistency in ratings. However, these deviations are small enough for us to include all annotators for the rest of our analysis.

	mean(std)	Dimension	mean	std
Content	0.5149 (0.1412)	Depth	0.2499	0.3109
		Diversity	0.4844	0.2206
		Elaboration	0.4924	0.2118
		Originality	0.2983	0.2202
Idea	0.5328 (0.1493)	Amount Concepts	0.5223	0.1405
		Amount Visual Ideas	0.5195	0.0863
		Concept Relatedness	0.4830	0.1742
		Unexpectedness	0.4624	0.1691
		Visually Imaginable	0.4660	0.1914
Writing	0.1849 (0.3082)	Coherence	0.2478	0.3317
		Readability	0.5779	0.1610
		TTI Effectiveness	0.0339	0.4111
		Vocabulary Level	0.5535	0.1611
Style	0.7791 (0.0606)	Depth	0.4559	0.3624
		Diversity	0.7299	0.0716
		Elaboration	0.8028	0.0774
		Originality	0.6310	0.0649

Table 5: Annotator Agreement on Scales evaluated with ICC2k. Mean and standard deviation by category and dimension across all data splits

Score Distributions for Dimensions Figure 2 shows the score distributions across all annotators and data splits. During annotation, the scales were reversed for certain items to break the monotony of annotation. This is corrected for the analysis, with the highest score of 5 and the lowest 1, and "not applicable" coded as 0 for all dimensions.
The distribution for **Style** and **Content** is fairly balanced, suggesting that the entire scale is being used effectively. **Style** shows a higher frequency of "not applicable" values and generally yields scores toward the lower end of the scale. **Content:Originality** score distribution shows a tendency toward higher values compared to **Style: Originality**. This could be due to the difficulty of evaluating style elements with a lot of (unknown) artists names. The qualitative evaluation indicated 18% instances in which the annotators needed to look up the name of the artists. The distributions of **Idea Combination: Unexpectedness and Visually imaginable** dimensions had higher values compared to the other dimensions.

Reliability The first assessment of the Cronbach's α shows that **Content** has the lowest value below zero, indicating poor internal consistency. However, this low value is especially due to sample split 5 (α = -20), indicating an outlier. The **Prompt Writing** has a low value below zero. We conclude that this category may not measure the same underlying concept or is too inconsistent to form a reliable scale. Its dimensions are either not well aligned or measure different concepts. The high deviation for these attributes further suggests that their consistency varies greatly between samples. The values for **Idea Combination** and **Style** suggests high internal consistency. The low standard deviations, especially for **Style**, suggest that the ratings are more stable in all data splits.

To understand whether the low α scores are due to some misalignment between dimensions, we use participants' feedback from the survey, where the dimensions **TTI Effectiveness** in the category **Prompt Writing**, and Depth dimensions under **Content** and **Style** were tagged as difficult to assess. Thus, we re-calculate the Cronbach α without these items. The results are in Table 3, showing that the **Content** scores improve toward moderate consistent reliability scores, whereas **Writing**, even with improvement, still has a low α to be accepted as reliable. Note that Table 3 shows both calculations of Cronbach α side by side in the same table, to ease comparison between the first and second calculation.

Agreement The inter-annotator agreements are in general moderate to good (**Content** and **Idea Combination**), and strong (**Style**). However, the agreements for **Prompt Writing** is quite poor, especially for **TTI effectiveness**, supporting the observations of the survey and reliability analysis.

The **Content** category shows moderate agreement among annotators (mean = 0.5149) with moderate variability (std = 0.1412). While there is some consistency, dimensions display moderate to low agreement, particularly **Depth**, which annotators found challenging to assess.

With a mean ICC of 0.5328, the agreement on the **Idea Combination** category indicates that there is moderate agreement between annotators. The standard deviation (0.1493) is similar to that for **Content**. All dimension values show moderate agreement, with lower variability, e.g. 0.0863 for **Amount of Visual Ideas**.

Prompt Writing has the lowest mean ICC value 0.1849, indicating overall a poor agreement among annotators. The deviation of 0.3082 reflects high variability, meaning the agreement across dimensions is inconsistent. While **Readability** and **Vocabulary Level** have moderate agreement, **Coherence** has a low value and **TTI Effectiveness** has the lowest value among all dimensions. This is supported by the

Dataset	Content		Idea Combination		Style	
	group	class	group	class	group	class
DiffusionDB	0.3567	0.1808	0.4326	0.0119	0.8836	5.32e-10
Midjourney23	0.7518	0.0003	0.5426	1.55e-05	0.0411	0.0001

Table 6: ANOVA/Kruskal-Wallis p-values based on the average of annotators scores

participants' feedback and the reliability results.

The **Style** dimensions shows the highest mean ICC of 0.7791 and the strongest agreement, already shown as the most reliable. Annotators seem to rate this dimension fairly consistently with a low deviation of 0.0606. The dimensions show moderate to high agreement overall. The **Depth** dimensions showing the lowest agreement, which was also indicated as a challenge by participants.

Impact of User Engagement and Prompt Length

The prompts evaluated with PCS were selected to represent user groups (extreme, prolific, moderate and infrequent) from both datasets. We also included prompts with different length. We analyzed the effect of these features and their impact on the creativity scores. We excluded the category **Prompt Writing** as well as the dimensions **Style Depth** and **Content Depth**, as these were assessed to be not reliable in both qualitative and quantitative analyses. Hypothesis tests were applied on the data in each category to look if there are significant differences in-between user groups or in-between prompt lengths (class). The scores by the annotators are averaged in each dimension, and the category score is a mean of all the dimensions. For normally distributed scores, we applied ANOVA; for the rest, the Kruskal-Wallis test is used. Table 6 shows the results, which indicate that prompt length consistently has a stronger impact than the user group, particularly in the categories **Idea Combination** and **Style**. While user group effects are mostly non-significant, Midjourney23 shows a minor but significant influence of user group on **Style**, suggesting that user expertise or background may slightly affect stylistic choices and how they are evaluated from the point of creativity.

Conclusion

The TTI services offer millions of users the ability to create images from text descriptions. The literature analyzed these services in relation to creativity and their impact on creative & entertainment industries. However, a creativity scale that can help assess individual prompts is a gap in the literature. In this study, we developed and tested the PCS, based on the CPSS, and our observations from the literature from creativity, prompt engineering, computational creativity and cognitive science. We furthermore embedded the PCS as a web-based tool and piloted it for clarity and user friendliness for the UI, as well as to receive user feedback on the scale itself. Our analysis based on a small participant pool showed that one category of the scale, i.e. **Prompt Writing** is not reliable and robust enough.

To design a solid experiment setup, we created a dataset of prompts extracted from two datasets, Midjourney and Stable Diffusion. The final prompt evaluation dataset has prompts from different types of users that we have categorized based on their engagement with these platforms, as well as the lengths of their prompts. Thus, we also analyzed if there are differences between these groups (of users, length of prompts, TTI services) using PCS scores of reliable dimensions. Only prompt length proved to be a significant factor for receiving higher creativity scores.

To follow up, we suggest 3 further studies: 1) An experiment design with a higher number of annotators to turn PCS into a valid construct for assessing prompt creativity, 2) an extension of PCS where the interaction between users and TTI models are taken into account, and 3) to create a model based on PCS for automatic creativity assessments of prompts.

Limitations

The study relies on older datasets, that do not reflect recent rapid advances in TTI models and prompting. Participant and data sample size, might introduce bias, limiting the generalizability of the findings. A larger scale study with a more diverse dataset could increase robustness and performance across contexts as well as the analysis of temporal stability of the PCS. The study focuses solely on static text input, excluding image analysis and the dynamic generation process, both crucial for understanding TTI and human creative collaboration. The PCS is a human-centered evaluation scale, which may not fully capture how TTI models interpret prompts, especially when these interpretations diverge from typical human understanding or rely on unconventional prompting strategies. Lastly, the stochastic nature of TTI models leads to varied outputs from the same prompt, with even simple inputs potentially resulting in unexpectedly complex outputs, further complicating the evaluation process.

Author Contributions

First author was responsible for the research methodology, data preparation, design and planning of the studies, development and implementation of the UI, data analysis, and manuscript writing. The second author provided essential guidance and expertise throughout the studies, as well as, writing and reviewing the manuscript.

Acknowledgments

A sincere thanks to the participants, including Jelle Koolstra, Jann Müller, Ebony Omodara, Gaby Voorbraak, Joeke Wolterbeek and Berke Yazan. Their time, effort and contributions were essential for this research.

References

Amabile, T. M. 1983. The social psychology of creativity.

Augustin, M. D.; Leder, H.; Hutzler, F.; and Carbon, C.-C. 2008. Style follows content: On the microgenesis of art perception. *Acta psychologica* 128(1):127–138.

Augustin, D.; Leder, H.; et al. 2006. Art expertise: A study of concepts and conceptual spaces. *Psychology Science* 48(2):135.

Baer, J. 1994. Performance assessments of creativity: Do they have long-term stability? *Roeper Review* 17(1):7–11.

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

Besemer, S. P., and O'Quin, K. 1987. Creative product analysis: Testing a model by developing a judging instrument. *Frontiers of creativity research: Beyond the basics* 367–389.

Besemer, S. P., and O'Quin, K. 1989. The development, reliability, and validity of the revised creative product semantic scale. *Creativity Research Journal* 2(4):267–278.

Besemer, S. P. 1998. Creative product analysis matrix: testing the model structure and a comparison among products–three novel chairs. *Creativity Research Journal* 11(4):333–346.

Bird, C.; Ungless, E.; and Kasirzadeh, A. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 396–410. ACM.

Boden, M. A., et al. 1994. Dimensions of creativity.

Boden, M. A. 1998. Creativity and artificial intelligence. *Artificial intelligence* 103(1-2):347–356.

Boden, M. A. 2005. What is creativity? In *Creativity in human evolution and prehistory*. Routledge. 27–55.

Boden, M. A. 2009. Computer models of creativity. *Ai Magazine* 30(3):23–23.

Boden, M. A. 2010. *Creativity and art: Three roads to surprise*. Oxford University Press.

Carson, S. H.; Peterson, J. B.; and Higgins, D. M. 2005. Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity research journal* 17(1):37–50.

Chang, M.; Druga, S.; Fiannaca, A. J.; Vergani, P.; Kulkarni, C.; Cai, C. J.; and Terry, M. 2023. The prompt artists. In *Proceedings of the 15th Conference on Creativity and Cognition*, 75–87.

Chen, Y., and Zou, J. 2023. Twigma: A dataset of ai-generated images with metadata from twitter.

Chung, J.; Hyun, S.; and Heo, J.-P. 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8795–8805.

Clark, G. 2004. *Teaching talented art students: Principles and practices*. Teachers College Press.

Cropley, D. H., and Kaufman, J. C. 2012. Measuring functional creativity: Non-expert raters and the creative solution diagnosis scale. *The journal of creative behavior* 46(2):119–137.

de Almeida, F., and Rafael, S. 2024. Bias by default.: Neocolonial visual vocabularies in ai image generating design practices. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–8. ACM.

Dehouche, N., and Dehouche, K. 2023. What is in a text-to-image prompt: The potential of stable diffusion in visual arts education.

Duncker, K., and Lees, L. S. 1945. On problem-solving. *Psychological monographs* 58(5):i.

Finke, R. A. 2014. *Creative imagery: Discoveries and inventions in visualization*. Psychology press.

Franceschelli, G., and Musolesi, M. 2024. Creativity and Machine Learning: A Survey. *ACM Computing Surveys* 56(11):1–41.

Frigotto, M. L., and Riccaboni, M. 2011. A few special cases: scientific creativity and network dynamics in the field of rare diseases. *Scientometrics* 89(1):397–420.

Goloujeh, A. M.; Sullivan, A.; and Magerko, B. 2024. The social construction of generative ai prompts. In *CHI Extended Abstracts*, 320–1.

Guilford, J. P. 1967. Creativity: Yesterday, Today and Tomorrow. *The Journal of Creative Behavior* 1(1):3–14.

Han, J.; Shi, F.; Park, D.; Chen, L.; Childs, P.; et al. 2018. The conceptual distances between ideas in combinational creativity. In *DS 92: Proceedings of the DESIGN 2018 15th International Design Conference*, 1857–1866.

Han, J.; Hua, M.; Shi, F.; and Childs, P. R. 2019. A further exploration of the three driven approaches to combinational creativity. In *Proceedings of the Design Society: International Conference on Engineering Design*, volume 1, 2735–2744. Cambridge University Press.

Hao, S.; Shelby, R.; Liu, Y.; Srinivasan, H.; Bhutani, M.; Ayan, B. K.; Poplin, R.; Poddar, S.; and Laszlo, S. 2024. Harm amplification in text-to-image models.

Hekkert, P., and van Wieringen, P. C. 1996. The impact of level of expertise on the evaluation of original and altered versions of post-impressionistic paintings. *Acta psychologica* 94(2):117–131.

Hennessey, B. A. 1994. The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal* 7(2):193–208.

Henriksen, D. 2014. Full steam ahead: Creativity in excellent stem teaching practices. *The STEAM journal* 1(2):15.

Katirai, A.; Garcia, N.; Ide, K.; Nakashima, Y.; and Kishimoto, A. 2023. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach.

Kim, K. H. 2006. Can we trust creativity tests? a review of the torrance tests of creative thinking (ttct). *Creativity research journal* 18(1):3–14.

Kotovenko, D.; Sanakoyeu, A.; Lang, S.; and Ommer, B. 2019. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4422–4431.

Leder, H., and Nadal, M. 2014. Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode–developments and challenges in empirical aesthetics. *British journal of psychology* 105(4):443–464.

Leder, H.; Belke, B.; Oeberst, A.; and Augustin, D. 2004. A model of aesthetic appreciation and aesthetic judgments. *British journal of psychology* 95(4):489–508.

Maier, N. R. 1931. Reasoning in humans. ii. the solution of a problem and its appearance in consciousness. *Journal of comparative Psychology* 12(2):181.

McCormack, J.; Llano, M. T.; Krol, S. J.; and Rajcic, N. 2024. No longer trending on artstation: Prompt analysis of generative ai art. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, 279–295. Springer.

Mednick, S. 1962. The associative basis of the creative process. *Psychological Review* 69(3):220–232.

Naik, R., and Nushi, B. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 786–808. ACM.

Oppenlaender, J.; Linder, R.; and Silvennoinen, J. 2024. Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering. arXiv:2303.13534 [cs].

Oppenlaender, J. 2023. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology* 1–14.

Palmini, M.-T. D. R.; Wagner, L.; and Cetinic, E. 2024. Civiverse: A dataset for analyzing user engagement with open-source text-to-image models.

Rosenman, R.; Tennekoon, V.; and Hill, L. G. 2011. Measuring bias in self-reported data. *International Journal of Behavioural and Healthcare Research* 2(4):320–332.

Runco, M. A., and Chand, I. 1995. Cognition and creativity. *Educational psychology review* 7:243–267.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity research journal* 24(1):92–96.

Runco, M. A.; Plucker, J. A.; and Lim, W. 2001. Development and psychometric integrity of a measure of ideational behavior. *Creativity Research Journal* 13(3-4):393–400.

Sanchez, T. 2023. Examining the text-to-image community of practice: Why and how do people prompt generative ais? In *Proceedings of the 15th Conference on Creativity and Cognition*, 43–61.

Sha, Z.; Li, Z.; Yu, N.; and Zhang, Y. 2023. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23. ACM.

Shalabi, F.; Nguyen, H. H.; Felouat, H.; Chang, C.-C.; and Echizen, I. 2023. Image-text out-of-context detection using synthetic multimodal misinformation. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 605–612. IEEE.

Sun, K.; Pan, J.; Ge, Y.; Li, H.; Duan, H.; Wu, X.; Zhang, R.; Zhou, A.; Qin, Z.; Wang, Y.; Dai, J.; Qiao, Y.; Wang, L.; and Li, H. 2023. Journeydb: A benchmark for generative image understanding.

Taylor, I. A., and Sandler, B. E. 1972. Use of a creative product inventory for evaluating products of chemists. In *Proceedings of the Annual Convention of the American Psychological Association*. American Psychological Association.

Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and psychological measurement*.

Trinh, K.; Spracklen, J.; Wijewickrama, R.; Viswanath, B.; Jadliwala, M.; and Maiti, A. 2024. Promptly yours? a human subject study on prompt inference in ai-generated art. *arXiv preprint arXiv:2410.08406*.

Urban, K., and Jellen, H. 1996. *Test for Creative Thinking - Drawing Production (TCT-DP)*. Swets Test Services.

Urban, K. K. 1991. Recent trends in creativity research and theory in western europe. *European Journal of High Ability* 1(1):99–113.

Wallach, M. A., and Kogan, N. 1965. Modes of thinking in young children.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2022. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. Publisher: arXiv Version Number: 4.

Wang, Z.; Zhao, L.; and Xing, W. 2023. Styledifusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7677–7689.

Ward, T. B., and Kolomyts, Y. 2010. Cognition and creativity. *The Cambridge handbook of creativity* 5:93–112.

Wei, Y.; Zhu, Y.; Hui, P.; and Tyson, G. 2024. Exploring the use of abusive generative ai models on civitai.

Wu, Y.; Nakashima, Y.; and Garcia, N. 2023. Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In *Proceedings of the 2023 ACM International conference on multimedia retrieval*, 199–208.