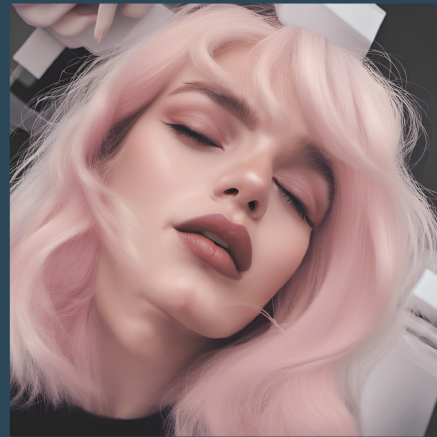


Aesthetics in Visual Training Datasets

Master Thesis at the Faculty of Industrial Design Engineering

Céline Offerman



Images generated with Stable Diffusion, using the prompt: "aesthetic"

Aesthetics in Visual Training Datasets

by

Céline Offerman

4663616

Chair: Prof. dr. ir. Bozzon, A.
Mentor: Drs. Maden, W.L.A. van der
Project Duration: February, 2023 - August, 2023
Faculty: Faculty of Industrial Design Engineering, Delft

LaTeX template: TU Delft Report Style, modifications by D. Zwaneveld

Acknowledgement

I had a lot of fun working on my thesis, and I want to thank a number of people who made this period so enjoyable.

First of all, I would like to thank my supervisory team, prof. dr. ir. Alessandro Bozzon and drs. Willem van der Maden. Without your expertise, clever perspectives, critical feedback, and overall support this thesis would never have been what it is today. Alessandro, through our conversations you sharpened my scientific viewpoint, I appreciate the honest feedback and am happy to have a supervisor who wants the work to be meticulous and of proper quality. Willem, thanks to our weekly conversations my enthusiasm for the subject has turned into a real passion. I admire your knowledge, skill, enthusiasm and opinions. I want to thank you very much for the inspiration, guidance, insights and for sustaining my caffeine addiction.

In addition to the crowdworkers, I would like to thank my 60 friends who participated in one of the experiments to develop the stimulus set. Of these 60, there are a few I want to give special thanks to, which are my colleagues from the Graduation Hole. Without the camaraderie, stolen sparkling water from Studiolabs and *roze koeken*, the vitamin D deficiency we probably contracted would not have been nearly as enjoyable.

Who are not among the previous group but also deserve a special thanks are my roommates: lanthe, Jules, Vita, Odile, Rosa and Kiri. It must have been exhausting to hear yet another excited story about generative models and unexpected experimental outcomes every night at dinner but you remained very polite.

Last but definitely not least, I would like to thank my parents, Theo and Gerti. Being your child, writing theses must be in my DNA. I also want to thank my brother, Coen, who only obstructed this thesis by persuading me to party with him. Besides being excellent proofreaders and feedback providers extraordinaire, we always have a great time together. Although I have not been able to use the non-negotiable eight-hour sleep opportunity window throughout for the entirety of my thesis, many wisdoms still come in handy. I won't waste words on thanking you further, because a certain voice is still echoing through my brain with *"Ik zeg maar zo, ik zeg maar niets, dan zeg ik al veel teveel."*

Summary

Correctly processing accumulated information is beneficial for our survival. Berghman and Hekkert (2017) argue that this is why we humans derive pleasure from having a sense of aesthetics. These aesthetic experiences can be seen as our brain's reward system for correctly perceiving and interpreting the world around us. While our senses have evolved to perceive and organise the physical world, these very mechanisms also come into play when we interact with the digital realm. Aesthetics in visual training datasets are of importance as it allows us to derive a sense of aesthetic pleasure from digital media. Integrating aesthetics into artificial intelligence, especially in text-to-image generators, becomes important to cater to humans psychological reward systems and to engage them at a deeper level.

This thesis is focused on investigating the annotation method used in the development of the LAION-Aesthetics V2 datasets and comparing it to other annotation methods for measuring aesthetics. The purpose is to explore whether there are more suitable alternatives to the current annotation method (where people are asked to annotate images with the instruction *"how much do you like this image on a scale from 1 to 10?"*, (Schuhmann, 2022) which is not backed by literature to actually measure aesthetics), and to evaluate the alignment between the LAION Aesthetics Predictor scores and human ratings.

This thesis explores different distinct levels of inquiry: one focuses on the design of instructions for image annotation tasks (alternative task design), while the other centers around measuring aesthetics during the annotation process (alternative metrics). Both lines of inquiry are supported by relevant literature, indicating their potential capacity to capture aesthetics. In addition to comparing alternative annotation methods, this thesis investigates three hypotheses related to the annotation of aesthetics within the project's context.

Four experiments are conducted using crowdsourcing to compare alternative task design and alternative metrics. The experiments include semantic concept activation, different phrasing of the annotation instruction, and alternative modalities (such as ranking and two-alternative forced choice). Next to these four experiments, a separate fifth experiment is deployed which looks into the evaluation of image content versus overall image liking. Two post hoc analyses are performed, one which compares scores that the LAION Aesthetics predictor assigns to the stimulus set to human image liking ratings, and one examining the influence of region on image liking ratings.

The LAION aesthetics approach performed equal to the alternatives with scientific backing. The ranking treatment even performed worse. For this data, region did not impact image liking ratings. No significant difference was found between participants' overall image liking and content liking. The LAION Aesthetics predictor scores partially aligned with human liking ratings but showed some disparities, particularly in extreme ratings. Qualitative analysis suggests that more research is necessary to make a judgement on whether *"liking"* is a relevant and appropriate approach for capturing aesthetics.

The limitations of the experiments include small sample sizes and the focus on a specific image class (buildings). Recommendations for future research include exploring different image classes, investigating other ranking modalities, and considering n-alternative forced choice experiments. It is also suggested to examine the influence of regions on aesthetic experiences in more detail, explore Gibbs Sampling with People for measuring image aesthetics, and explore different demographic groups and contexts.

Contents

Acknowledgement	i
Summary	ii
1 Introduction	1
1.1 More information on the project context	2
1.2 Relevance of this thesis for design	3
1.3 Research question and hypotheses	4
1.3.1 H1: Semantic Concept Activation	5
1.3.2 H2: Region	5
1.3.3 H3: Aesthetic Value	6
1.3.4 H4: Ranking	6
1.3.5 H5: 2AFC	6
1.3.6 H6: Content VS Overall Image	6
1.3.7 H7: Predictor Comparison	7
1.4 Overview experiments	7
1.5 Overview report	9
2 Background	10
2.1 The importance of aesthetics in visual datasets	10
2.1.1 Aesthetic experiences	10
2.1.2 The functionality of aesthetics	12
2.1.3 The value of aesthetics in visual datasets	12
2.1.4 Conclusions from the importance of aesthetics in visual datasets	13
2.2 Existing visual datasets claiming to reflect aesthetics	13
2.2.1 Objective and subjective aesthetic datasets	13
2.2.2 Participant avoidance	13
2.2.3 Mean opinion score (MOS)	14
2.2.4 Experts	14
2.2.5 Conclusions of existing visual datasets claiming to reflect aesthetics	14
2.3 Literature leading to H1: Semantic Concept Activation and H2: Region	15
2.3.1 Empirical aesthetics	15
2.3.2 Objective and subjective aesthetics metrics	15
2.3.3 Challenges in generalising empirical aesthetics models	15
2.3.4 Semantic concept activation	17
2.3.5 Conclusions of the literature	17
2.4 Literature leading to H3: Aesthetic Value	18
2.4.1 Automatic linguistic behaviour	18
2.4.2 Existing studies where participants are asked about aesthetic value	18
2.4.3 Conclusions from the literature	18
2.5 Literature leading to H4: Ranking	19
2.5.1 Ranking modality	19
2.5.2 Conclusions from the literature	19
2.6 Literature leading to H5: 2AFC	19
2.6.1 2AFC modality	19
2.6.2 Conclusions from the literature	19
2.7 H6: Content VS Overall Image	20
2.7.1 High internal consistency between subjects for a subjective experience	20
2.7.2 Conclusions from the literature	20
2.8 Ethical considerations	20

2.8.1	(Lack of) consent	20
2.8.2	Copyright	21
2.8.3	Crowdsourcing	21
3	Experimental design	23
3.1	Introduction	23
3.2	Stimulus sets	24
3.2.1	Extracting images from the LAION dataset	24
3.2.2	Considerations for image class selection	24
3.2.3	Ethical considerations for the stimulus set	24
3.2.4	Two iterations of the stimulus set	25
3.2.5	Stimulus set 1 (1.1 and 1.2)	25
3.2.6	Stimulus set 2 (2.1 and 2.2) - a controlled distribution of aesthetic value	25
3.3	Participants	27
3.3.1	Overview participants stimulus set 1	27
3.3.2	Overview participants stimulus set 2	28
3.4	Task planning	28
3.5	Control treatment <i>“how much do you like this image on a scale from 1 to 10?”</i>	28
3.5.1	Structure of the control treatment implemented in this thesis	28
3.5.2	Analysis	29
4	Experiment 1: semantic concept activation with the Unified Model of Aesthetics	31
4.1	Introduction	31
4.2	Experiment 1 - hypotheses and expected outcomes	32
4.3	Confounding variables	32
4.4	Method	33
4.4.1	Materials	33
4.4.2	Procedure	33
4.4.3	Analysis	34
4.5	Conclusions	36
5	Experiment 2: rating images on aesthetic value	38
5.1	Introduction	38
5.2	Experiment 2 - hypothesis and expected outcome	38
5.3	Method	39
5.3.1	Materials	39
5.3.2	Procedure	39
5.3.3	Analysis	40
5.4	Conclusions	41
6	Experiment 3: ranking images on aesthetic value	42
6.1	Introduction	42
6.2	Experiment 3 - hypothesis and expected outcome	42
6.3	Method	43
6.3.1	Materials	43
6.3.2	Procedure	43
6.3.3	Analysis	44
6.4	Conclusions	45
7	Experiment 4: image preference with two alternative forced choice	46
7.1	Introduction	46
7.2	Experiment 4 - hypothesis and expected outcome	46
7.3	Method	47
7.3.1	Materials	47
7.3.2	Procedure	47
7.3.3	Analysis	48
7.4	Conclusions	49

8	Experiment 5: content vs overall image liking	50
8.1	Introduction	50
8.2	Experiment 5 - hypothesis and expected outcome	50
8.3	Method	51
8.3.1	Materials	51
8.3.2	Procedure	51
8.3.3	Analysis	51
8.4	Conclusions	52
9	Post hoc analysis: comparison between participant ratings and predictor scores	53
9.1	Introduction	53
9.2	Post hoc analysis - hypothesis and expected outcome	53
9.3	Analysis	54
9.3.1	Qualitative results	54
9.4	Conclusions	56
10	Discussion	57
10.1	Summary of the findings	57
10.1.1	Summary of the results	58
10.2	Interpretations of the findings and implications for the problem context	59
10.3	Limitations of the experiments conducted in this thesis	60
10.4	Suggestions for future research	61
10.4.1	Semantic concept activation with the original semantic concepts	61
10.4.2	Semantic concept activation with the Unified Model of Aesthetics with the original UMA stimulus set	61
10.4.3	Different image classes	61
10.4.4	More iterations on ranking modalities	62
10.4.5	More iterations on alternative-forced choice	62
10.4.6	An additional study examining the impact of regions on aesthetics.	62
10.4.7	Gibbs sampling with people for image aesthetics	62
10.4.8	Different contexts and demographical groups	62
10.5	Personal reflection	63
11	Conclusion	64
A	Bibliography	65
B	Exploring aesthetic experiences from a philosophical perspective	75
B.0.1	A brief overview of philosophical viewpoints	75
B.0.2	Everyday aesthetics	76
B.0.3	Exploring aesthetic attitude, aesthetic sensibility, and aesthetic fluency	76
B.0.4	Controversy around aesthetic emotions	77
C	Exploring aesthetic experiences from a neuroscientific perspective	78
C.0.1	The neuroscientific basis of the aforementioned philosophical theories	78
D	An overview of empirical aesthetics literature	80
E	Existing visual datasets claiming to reflect aesthetics	82
F	HREC approved consent	85
F.1	HREC submission	85
F.2	HREC Revisions	85
G	Stimulus set	96
G.1	Rationale for the approach	96
G.1.1	Goal	96
G.1.2	Participants	96
G.1.3	Summary of the two experiments	96
G.1.4	Analysis	96

H	Confounding variables	99
H.1	Confounding variables included in Experiment 1	99
H.1.1	Demographical information provided by Prolific (Chamorro-Premuzic, Furnham and Reimers, 2007):	99
H.1.2	What is measured by Prolific:	100
H.1.3	Self-efficacy	100
H.1.4	Variable: Socioeconomic status and education level (Mcmanus and Furnham, 2006)	100
H.1.5	Variable: noise levels	101
H.1.6	Variable: working environment	101
H.1.7	Variable: conformity pressure and sense of being watched (Hesslinger et al., 2017)	102
H.1.8	Variable: colourblindness (Kang et al., 2020)	102
H.1.9	Variable: aesthetic fluency (Cotter et al., 2023)	103
H.1.10	Variable: aesthetic attitude Mcmanus and Furnham (2006)	103
H.2	Confounding Variables Excluded in Experiment 1	105
H.2.1	Personal history	105
H.2.2	Personality traits	105
H.2.3	Aesthetic sensibility (Berleant, 2015)	105
H.2.4	The device used for crowd working (Hettiachchi et al., 2020)	105
I	Experiment 1 analysis	106
I.1	Results difference per stimulus	106
I.2	Results Variance	107
I.3	Results Region	107
J	Experiment 2 analysis	108
J.1	Results difference per stimulus	108
K	Experiment 3 analysis	109
K.1	Overview of the two treatments	109
K.2	Internal consistency per participant for the Experiment 3 treatment	112
L	Comparison between aesthetic scores assigned by participants and the LAION aesthetic predictor analysis and overview	113
L.1	The two linear regressions	113
L.2	Overview of the images, their scores assigned by humans and their scores assigned by the predictor	114
L.3	Qualitative results - stimuli where humans values stimuli noticeably lower/higher than the predictor	116
M	Experiment 4 analysis	117
M.1	Results difference per stimulus	117
M.2	Internal consistency 2AFC treatment	117
M.3	Qualitative results	119
N	Earlier version of the stimulus set	120
O	Analysis content vs overall image liking	122
P	Approved design brief	123

1

Introduction

Aesthetics, as an essential aspect of human cognition and perception, plays a universal role in our species. Our ability to appreciate aesthetics is deeply ingrained in our biology, and it has evolved over time to support our survival and adaptability. The *by-product* hypothesis, as described by Hekkert and Leder (2008) and Johnston (2003), provides insights into the origins of our aesthetic appreciation. According to this theory, aesthetics can be seen as our brain's reward system for accurately perceiving and interpreting the world around us.

As humans, we rely on our senses to fulfill our needs and survive in our environment. Our senses enable us to gather information about our surroundings, recognise potential threats, and identify opportunities. This processing of accumulated information is crucial for our survival, and as a result, our brains have developed mechanisms to reward us for this correct perception. Berghman and Hekkert (2017) argue that this is why we derive pleasure from having a sense of aesthetics. Impressions that support our survival and the development of our senses evoke aesthetic appreciation, and aesthetics become a mechanism through which we find pleasure in perceiving and organising the world.

The impact of aesthetics extends beyond our physical environment to the digital realm. In today's world, our interactions with technology and artificial intelligence increasingly shape our daily experiences. Although the digital environment may not directly impact our physical survival, our brains still respond to digital information as they do with the physical world. Our visual system, being the most prominent sensory system, plays a crucial role in our engagement with the digital world. The visual system seeks patterns that facilitate perceptual organisation, and even with digital media, we can experience a sense of aesthetic pleasure (Hekkert, 2006).

The integration of aesthetics into the realm of artificial intelligence has significant implications for text-to-image generators, AI models that employ a machine learning method which involves training on very large visual datasets which consist of image-text pairs. Through this process, the AI model learns to make connections between natural language and visual features (Brisco, Hay and Dhami, 2023). This enables the generator to essentially turn words into images. Neglecting these aesthetic experiences in AI development can lead to missed opportunities to engage users at a deeper level.

In this thesis, I delve into the relationship between aesthetics and AI, particularly in the context of the development of these visual training datasets for generative models. I focus on the LAION Aesthetics V2 datasets, which are used to train Stable Diffusion, a text-to-image generator with up to 10 million daily users (Cai and Martin, 2023). In order to build these datasets, the LAION researchers asked participants "*how much do you like this image on a scale from 1-10?*" (Schuhmann, 2022). This is a remarkable approach because it is not backed by literature that liking measures aesthetics. By examining this approach to image annotation and comparing it with alternative methods supported by aesthetic theory, I want to explore the most appropriate way to integrate aesthetics into AI-driven image generation. I do this through the following research question:

How does the annotation method used in the development of the LAION-Aesthetics V2 datasets compare to other annotation methods for measuring aesthetics, considering different approaches and annotating aesthetics in this context?

By exploring alternative annotation methods, this thesis seeks to answer the question of whether there are more appropriate ways to annotate images to capture their aesthetic qualities accurately. To achieve this, a series of experiments will be conducted, comparing the current annotation method with alternative methods supported by existing literature, namely if semantic concept activation, different question phrasing, alternative modalities such as ranking and two-alternative forced choice (2AFC) result in different evaluations than the current setup. Next to this, it is also studied if participants who are instructed to annotate images specifically on their content yield different results than participants who are instructed to annotate images on their overall image liking. Additionally, the scores that the LAION Aesthetics Predictor assigns to the stimulus set are compared to human liking ratings, and the influence of region on liking ratings is examined.

By focusing on aesthetics in visual training datasets, we can create AI systems that not only excel in their functional capabilities, but also resonate with users on a deeper level. The results of this thesis could be a small step in the right direction in the way AI interacts with humans and pave the way for more human-centric and aesthetically enriching experiences.

1.1. More information on the project context

Schuhmann (2022) describes that to create the LAION Aesthetics Predictor V2, several models were trained that predict the rating people gave images when they were asked *"how much do you like this image on a scale from 1 to 10?"*. The researchers examined the outputs of the different models and picked the model which, in their subjective view, produced the visually most appealing results, even though other models performed better on more objective metrics. Next to this, the annotation instruction used is not validated in the literature. When looking at literature on aesthetics, this seems to be a complex sensory experience (Baumgarten, 1750: 2007; Merleau-Ponty, 1960: 2011; Mandoki, 2007; Saito, 2010), which induces both appraisal (Simpson, 1975) and contemplation (Schopenhauer, 1818: 2010; Chatterjee, 2002, 2003; Vartanian and Skov, 2014) in beholders. With this theory in mind, the procedure of asking participants about image liking does not seem to cover the load of this complex human experience.

When zooming in on the content of the LAION Aesthetics V2 subsets, it becomes clear that the images with the highest 'aesthetic' scores are mainly landscapes and portraits of women (Baio, 2022), which look suspiciously like generated images (Figure 1.1). Click [here](#) to view the images that receive the highest score from the LAION Aesthetics Predictor.



Figure 1.1: Sample of images from the LAION dataset and their respective aesthetics scores ascribed by the LAION-Aesthetics V2 Predictor.

Well-known and established aesthetic principles emphasize the importance of diversity for aesthetic experiences (Berghman and Hekkert, 2017). The researchers describe that our senses have evolved to collect information about the complex environment in which we find ourselves. With overly unified stimuli this could lead to sensory dullness. Thus, we appreciate having variety in our stimuli to counterbalance unity. Variety is the amount and intensity of perceived differences between different perceptual characteristics and elements (Berlyne, 1972). With this aspect of aesthetic experiences in mind, we can comment on the homogeneity exhibited by the highest-scoring groups of images (with an aesthetic score of 6.5+). As the scores increase, the diversity noticeably decreases. This can be considered an indication that the current LAION Aesthetics V2 datasets may not align with literature on aesthetics.

In addition, homogeneity in a training dataset is also not beneficial. Text-to-image models like Stable Diffusion aim to create images from all text prompts. However, if the training dataset consists of only a limited number of subjects and visual styles with high aesthetic scores, it hinders the model's ability to generate diverse images across different topics and styles of high aesthetic quality. With a generative model, you want the images to be beautiful as such, rather than the model predominantly being able to generate images with high aesthetic quality of stereotypical aesthetic subjects (e.g. women, flowers). To achieve this, a highly diverse training dataset is of importance, including images from the highest-scoring aesthetic categories depicting for instance trash cans, toilets, shoes, roads, as well as people of diverse ethnicities and genders.

1.2. Relevance of this thesis for design

Rapid changes are taking place in the design field, driven by openly accessible generative models such as Stable Diffusion. These text-to-image generators can inspire designers during the ideation phase. Examples of their implementation include the development of contextual user collages (Brunns, 2007; Muller, 2001) or to serve as a communication tool between intended users and designers for abstract concepts that may be difficult to articulate for users. They can also be utilised as an inspirational tool during brainstorming (Roozenburg and Eekels, 1995; Higgins, 1994) and bridge the gap between brainwriting and brain drawing explorations (Roozenburg and Eekels, 1995; Boeijen et al., 2014). Additionally, in the embodiment phase, they can serve as an inspirational resource for the visual appearance of the intended product. These are just a few examples of the countless instances where models like Stable Diffusion can potentially support design processes.

The current dataset's observed homogeneity may restrict the diversity of input for designers. The perception of text-to-image generators, as highlighted in the work of Brisco, Hay, and Dhimi (2023),

mainly producing otherworldly and dreamy images, can hinder the consideration of crucial structural and behavioural elements in industrial design. These elements are essential for meeting functional requirements of product design (Gero and Kannengiesser, 2014; Umeda and Tomiyama, 1997). While these dreamy images might provide abstract inspiration to designers, current models are not (yet) practical for generating functional concepts. Additionally, generative models tend to favour the development of everyday objects that occur more frequently in their training dataset (Brisco, Hay, and Dhimi, 2023). Overall, due to the lack of diversity, it can be expected that implementing current text-to-image generators in design processes carries a risk of design fixation. This fixation could result in derivative and non-novel outcomes, which is detrimental in a creative field like design, where diversity is of high importance. Especially when generative models are implemented in the early stages, where design explorations can considerably shape designers' perceptions of users and the context, uniform output from a model can potentially heavily influence the design process.

This thesis examines the comparison of the annotation method used in the development of the LAION-Aesthetics V2 datasets with other approaches for measuring aesthetics. The primary goal is to explore the integration of aesthetics into visual datasets to effectively train generative models. Aesthetic experiences are at the core of this investigation. By examining different annotation methods for measuring aesthetics and their effectiveness in annotating images, this thesis bridges the gap between aesthetics, generative models and design practice. The results could contribute to a better understanding of aesthetics' role in generative models for design inspire the creation of designs that resonate with users' aesthetic preferences.

1.3. Research question and hypotheses

To address the research question at hand, within comparing alternative annotation methods to the current situation, this thesis encompasses two different levels of inquiry: one focused on the design of instructions associated with image annotation tasks (alternative task design), and the other centered around the measurement of aesthetics within the annotation process (alternative metrics) (Figure 1.2). Both experimental lines of inquiry draw support from relevant literature suggesting their potential suitability for capturing aesthetics.

This first line of inquiry explores alternative task design, specifically incorporating semantic concept activation. This approach aims to examine the effect of exposing individuals to aesthetic-based semantic concepts, considering the complex nature of aesthetics and the lack of support for using mere "liking" as a valid measure of aesthetics in existing literature. By activating specific concepts in participants, the objective is to heighten their attention to underlying aesthetic principles when evaluating images, potentially leading to different ratings compared to direct inquiries about image liking. However, Experiment 1 yielded no significant effect on liking ratings from exposing participants to semantic concept activation (see Section 4.4.3). Consequently, additional investigations are conducted to examine whether the measurement method itself influences image evaluation.

Next to the comparison of alternative annotation methods, this thesis also deploys three hypotheses looking into the annotation of aesthetics within the context of the project. The full hypotheses and their corresponding motivations are described below.

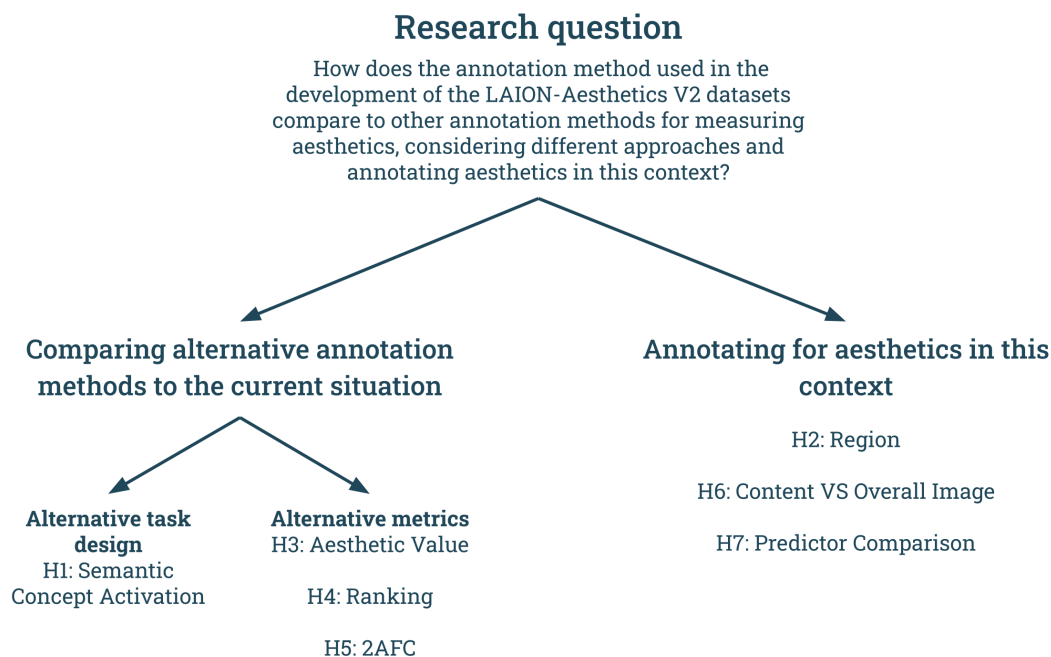


Figure 1.2: Different lines of inquiry in this thesis.

After this, all hypotheses will be discussed, in order of appearance in this thesis.

1.3.1. H1: Semantic Concept Activation

In Experiment 1 (Chapter 4), the control treatment, as described in Section 3.5, is compared to an alternative task design, namely participants who are first exposed to semantic concept activation (the process of activating specific concepts in the mind based on prior knowledge) with the Unified Model of Aesthetics, an empirical aesthetics model tested by Berghman and Hekkert (2017). This is conducted because a promising previous study by Faerber et al. (2010) found significant results in a similar set-up. The motivation behind this is that by first drawing participants' attention to universal principles of aesthetics, this will cause them to pay more attention to these and perceive the stimulus differently. The premise of the method is that participants are subconsciously nudged towards an aesthetic attitude. For the full rationale, please refer to Section 2.3. Experiment 1 was designed to investigate the following hypothesis:

- **H1: Semantic Concept Activation: exposing participants to semantic concept activation through questions about typicality and novelty, unity and variety, and relatedness, will influence their ratings of how much they like images.**

1.3.2. H2: Region

In Experiment 1, the influence of participants' region will be analysed post hoc, as some researchers describe that participants' region can potentially influence their aesthetic experience (Hekkert and Leder, 2008; Berghman and Hekkert, 2017). Therefore, I will investigate the effect of a participants' region on aesthetic ratings. For the full rationale, please refer to Section 2.3. The hypothesis which will be analysed:

- **H2: Region: participants from different regions experience different images as aesthetic.**

1.3.3. H3: Aesthetic Value

In Experiment 2 (Chapter 5), the impact of using an alternative metric, specifically a different question phrasing "how aesthetic do you find this image?" is compared to "how much do you like this image on a scale from 1-10?". This is examined because this sentence is frequently cited in existing literature for measuring aesthetics. Semin and De Poot (1997-a, 1997-b) found that word choice in question phrasing can achieve significantly different answers among participants. For a more comprehensive explanation and description of the relevant literature, I would like to refer you to Section 2.4. This experiment was formulated to study the following hypothesis:

- **H3: Aesthetic Value: the manner in which participants are asked a question on aesthetics significantly impacts their responses.**

1.3.4. H4: Ranking

The purpose of the 3rd experiment (Chapter 6) is to examine whether there are significant differences in results between rating vs ranking as an alternative metric. The ranking modality is already used to develop a dataset on attractiveness, an aesthetics adjacent concept (Nguyen et al., 2012). In addition, some literature has used human preference ranking to train a language model (Yuan et al., 2023). This can be taken as an indication that it may be an appropriate modality for training text-to-image models. It is expected that the ranking treatment will significantly differ from the control treatment, because the ranking modality obliges participants to assign an image for every value between 1-10. More details can be found in Section 2.5. This experiment is conducted to look into the following hypothesis:

- **H4: Ranking: participants' rankings of image aesthetics will show significant differences when compared to their subjective ratings of image liking on a scale of 1-10.**

1.3.5. H5: 2AFC

In Experiment 4 (Chapter 7) rating is compared with the alternative metric two-alternative forced choice (2AFC). This comparison is motivated by the extensive literature supporting the use of this approach in image annotation for aesthetics or related concepts (Palmer, Schoss and Sammartino, 2013; Wu et al., 2023; Swanson et al., 2012; Bara et al., 2021; Bıyık et al., 2020; Sadigh et al., 2017). The expectation is that this comparison may lead to significantly different results compared to rating, as 2AFC requires participants to choose between two stimuli. For further rationale and literature, please see Section 2.6. The following hypothesis is the reason to deploy experiment 4:

- **H5: 2AFC: the alternative forced choice annotations of image aesthetics will result in significantly different outcome scores compared to participants' subjective ratings of image liking on a scale of 1-10.**

1.3.6. H6: Content VS Overall Image

This hypothesis was formulated after reflecting on the results from the first four experiments conducted in this thesis, and previously conducted studies with similar setups in combination with aesthetic theory.

Previous research on empirical aesthetic metrics and the experiments conducted in this thesis demonstrate a high level of agreement among participants. This agreement is desirable from a dataset perspective because when there is a general consensus among people regarding their aesthetic preferences, the dataset can be used with a degree of confidence for training Stable Diffusion. Yet the theory indicates that aesthetic experiences are very subjective, which makes high internal consistency between subjects an interesting finding. A common characteristic of these experiments is the use of functional stimuli such as bicycles, coffee makers, and in this thesis, buildings. It is possible that these functional stimuli primarily evoke functional contemplation in participants rather than aesthetic contemplation. To investigate whether the participants' liking judgements are influenced by the affordances of the buildings, a hypothesis was formulated. Further details can be found in Section 2.7.

- **H6: Content VS Overall Image: when participants rate their liking of the image content, there will be significant difference with when they indicate the image liking of the overall image.**

1.3.7. H7: Predictor Comparison

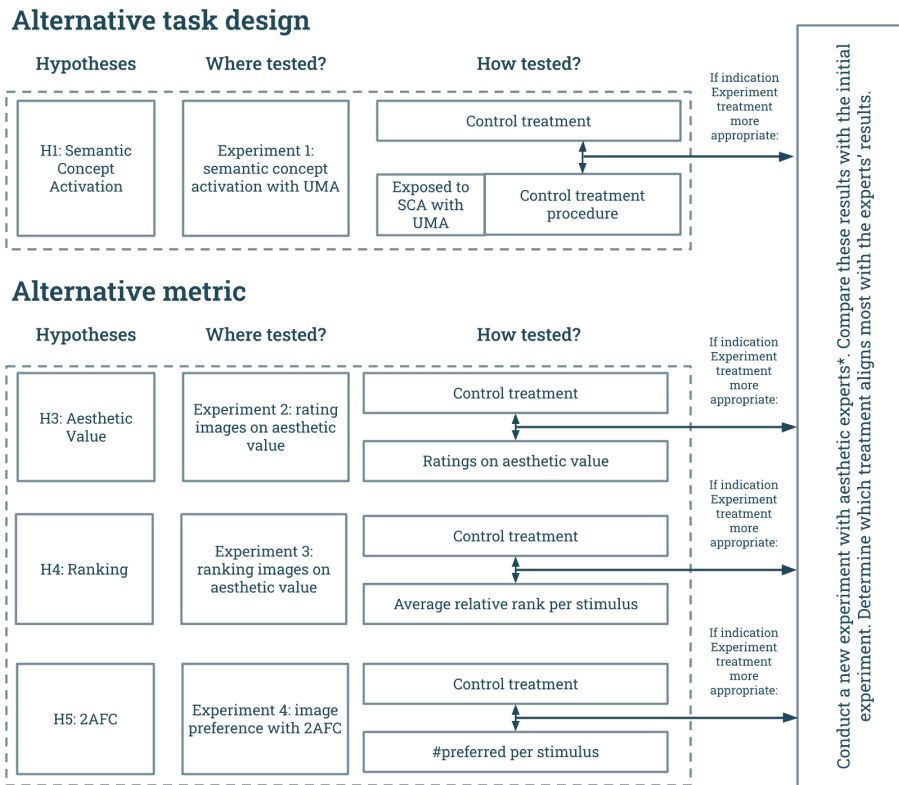
To examine the alignment between the LAION Aesthetics Predictor scores and human ratings, participants' ratings on image liking are compared to the scores provided by the predictor for the stimulus set, which will be introduced in Section 3.2.6. This comparison is looked into due to the observation discussed in the previous section, which highlighted that the higher-scoring image segments in the LAION-Aesthetics v2 datasets mainly consist of landscapes and portraits of women in a generated image style. By analysing if there is a significant difference between the scores and the ratings, it can be investigated whether the homogeneity in the higher image segments indicates a significant misalignment between the predictor and human liking ratings. For more explanation and the analysis of this hypothesis, please refer to Section 9.

- **H7: Predictor Comparison: the predicted aesthetic scores assigned by the LAION Aesthetic Predictor are significantly different from the average image liking scores assigned by participants.**

1.4. Overview experiments

Figure 1.3 shows all hypotheses tested in this thesis, where they are tested, and how they are tested. Four hypotheses are compared directly to the control treatment (H1: Semantic Concept Activation, H3: Aesthetic Value, H4: Ranking and H5: 2AFC). These hypotheses are deployed with the purpose to compare alternative annotation methods to the current situation. Next to this, H2: Region, H6: Content VS Overall Image and H7: Predictor Comparison are deployed to learn more about annotating for aesthetics in this context.

Comparing alternative annotation methods to the current situation



Annotating for aesthetics in this context

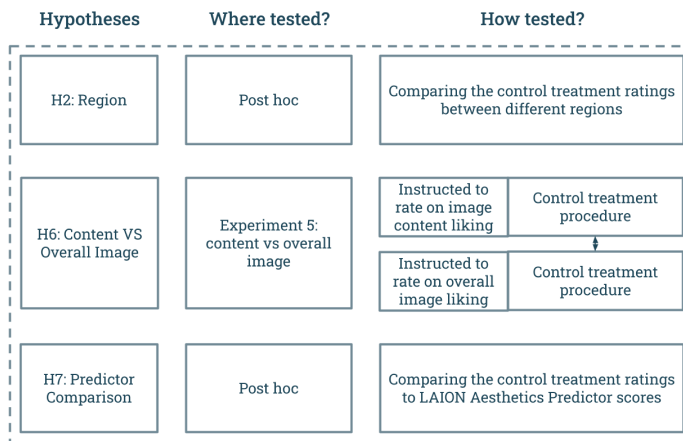


Figure 1.3: The hypotheses tested in this thesis grouped on their different levels of inquiry, where they are tested, and how they are tested. *What is understood under the term 'aesthetic expert' is described in Section 2.2.4 and 3.1.

The exact set-up of the control treatment will be discussed in more detail in Chapter 3. When an experiment indicates that it might be a more appropriate approach than the control treatment, a follow-up experiment will be conducted, with the purpose of checking in which treatment the results of the crowdworkers align more with the results of aesthetic experts.

1.5. Overview report

The objective of this thesis is to compare the current approach, which involves asking participants to rate images on a scale from 1 to 10 based on how much they like them, despite lacking scientific evidence of its effectiveness in measuring aesthetics, with alternative approaches that are supported by literature in this regard. Image annotation, in this context, refers to the process of creating descriptive metadata for images, moving beyond their pixel-level representation.

I will look at whether there are more appropriate alternatives to the current annotation method for creating the training dataset for the LAION-aesthetics V2 predictor. This will be done by running several crowdsourcing experiments, a method already widely used for image annotation (Daniel et al., 2019; Sun and Stolee, 2016; Chittilappilly et al., 2016). In these experiments, the current context is compared with alternative methods that have found support in the literature.

Chapter 2 will delve deeper into aesthetic theory. Existing datasets that claim to reflect aesthetics will be examined. The relevant literature is discussed, which is used as a basis for methodology decisions, and as backing for the experiments deployed. Chapter 2 concludes with ethical considerations. Chapter 3 outlines the experimental design. Although all experiments test different features, there are some common denominators (e.g. stimulus set). The common elements are the focus of this chapter. In Chapter 4, 5, 6, 7, and 8 the experiments are outlined one by one. In Chapter 9 a post hoc analysis describes the comparison of aesthetic prediction scores by the LAION Aesthetics Predictor to human participant ratings. An overview of which hypothesis is discussed in detail in which chapter is depicted in Table 1.1.

Chapter	Which hypothesis
Chapter 4	H1: Semantic Concept Activation and H2: Region
Chapter 5	H3: Aesthetic Value
Chapter 6	H4: Ranking
Chapter 7	H5: 2AFC
Chapter 8	H6: Content VS Overall Image
Chapter 9	H7: Predictor Comparison

Table 1.1: Which chapter describes which hypothesis (with hyperlinks).

The hypotheses that will be looked into in these chapters have been introduced earlier in this chapter. Each experimental chapter is structured as follows:

- Introduction
- The hypothesis that will be tested with the expected outcome
- Method
 - Materials
 - Procedure
 - Analysis
- Conclusion

The experimental chapters will lead to the Discussion (Chapter 10) and Conclusion (Chapter 11).

2

Background

2.1. The importance of aesthetics in visual datasets

This section explores aesthetic experiences. I define aesthetic experiences through a personal encounter with a painting by Mark Rothko, detailing this disinterested encounter which includes sensory perception, appraisal, and a contemplative state. The functionality of aesthetics is discussed, highlighting its universal role in human perception and survival. Additionally, the value of aesthetics in visual datasets is addressed, noting their importance in digital environments. The section concludes with insights on how aesthetic experiences can be considered as the brain's reward system for accurate perception of the world.

2.1.1. Aesthetic experiences

Markovic (2012) states that aesthetic experiences are the most ill-defined and vague concepts from the psychology of art and experimental aesthetics. To appropriately measure aesthetic experiences in this thesis, it is necessary to have an unambiguous definition of the concept. The subject has been examined from different angles. In Chapter 1, a start was made to describe aesthetics. This subsection will continue this attempt from a philosophical as well as neuropsychological perspective, on the basis of an example from my own experience.

During my research for this thesis, I visited *Untitled (White, Pink and Mustard)* by Mark Rothko in a museum, as also shown in Figure 2.1. The moment I entered the room, my perception engaged with the painting in an active and embodied manner, in line with Merleau-Ponty's (1960: 2012) perspective. My gaze was absorbed into the canvas. Partly because of the sheer size of the work, which occupies a large part of the wall, and partly because of the richly pigmented colour fields, I feel an intense and immersive experience.

My aesthetic experience is disinterested. What is meant by this is that, as a beholder, I can derive satisfaction from the stimulus itself (Kant, 1790: 2000). I don't want anything from the Rothko; I appreciate the painting for its own sake.

During this aesthetic encounter, appraisal comes into play. As I evaluate the aesthetic qualities of the painting, I subconsciously take in both my previous experiences and the current context (Simpson, 1975). Since I was little, my parents have been bringing me to modern art museums. This leads me to appreciate the abstract qualities of Rothko's work. The large painting has its own space in the museum, allowing me to really become absorbed in it. Had the space been cluttered with other work, my experience would certainly have been different. The fact that the painting has received the quality seal of a reputable museum also undoubtedly plays a role in my experience.

As my contemplative state deepens, my thoughts move from the surprising balance of the shapes and the soothing harmony of the use of colour into a deeper state, of introspection and reflection (Schopenhauer, 1818:2010; Chatterjee, 2002, 2003; Vartanian and Skov, 2014). The soothing mental floating

amongst Rothko's colour fields, for instance, makes me think about my own fast-paced life, and even leads to daily habitual change where I start my day with five minutes of stretching and clearing my mind.



Figure 2.1: Mark Rothko - Untitled (White, Pink and Mustard) from 1954, Museum Folkwang. Photograph taken by me.

Aesthetics has its own discipline in philosophy, and any description I could include in this thesis on aesthetic experiences would never do this body of literature justice. To elaborate a little more on the text snippet included above, I have enclosed a further description in Appendix B. Next to the philosophical viewpoint Appendix C provides further elaboration on neuroscientific findings that substantiate parts of these philosophical concepts.

The literature reviewed for this thesis, described in Chapter 1 and this section, led to the following definition of an aesthetic experience (schematically presented in Figure 2.2):

An aesthetic experience is an embodied sensory experience, which leads to a disinterested encounter which involves appraisal and induces a contemplative state in the beholder.

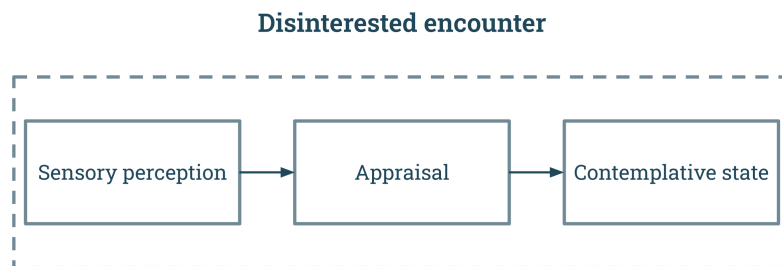


Figure 2.2: A schematic overview of an aesthetic experience.

2.1.2. The functionality of aesthetics

How have aesthetics come to play such an important role in how we interact with our immediate environment?

There are general principles for aesthetic pleasure that are grounded in human nature (Hekkert, 2006). Although our perception of aesthetics operates according to universal laws, it does not directly imply that we are in absolute agreement. Albeit the foundation is the same, individual differences are a result of interpretation differences. This means that although aesthetic responses may differ from each other, they do behave in a lawful manner (Hekkert, 2006). This phenomenon can easily be explained by one of the attributes of aesthetics that will be introduced later in this chapter. A widely accepted principle of aesthetics is a balance between typicality and novelty (Berghman and Hekkert, 2017). Although this aesthetic law lies in the foundation of our human nature, people with different mental models may deal with it differently.

Why do aesthetics play a universal role in our species? This can be explained by the *by-product* hypothesis, which is described by Hekkert and Leder (2008) and Johnston (2003). Various research is based on this theory and is presented below.

Our ability to appreciate aesthetics is part of our biology. Like the rest of our human qualities, this has the consequence that it has been subject to the evolution of our species (Berghman and Hekkert, 2017). Hekkert (2006) describes that all evolutionary reasoning centralises around our ability to adapt as a species, so we can survive in order to reproduce. Because of natural selection, and plenty of time to evolve, our psychological mechanisms have grown to support us in these goals (Barkow, Cosmides and Tooby, 1992).

We rely on our senses to fulfil our needs in order to survive (e.g. find shelter and food, stay safe). These senses enable us to get an understanding of our immediate environment, allowing us to recognise threats (e.g. a poisonous plant) and opportunities (e.g. a trail in the forest). Because the correct processing of collected information is useful for our survival, Berghman and Hekkert (2017) argue that this is why we humans derive pleasure from having a sense of aesthetics. Aesthetic experiences can thus be seen as our brain's reward system for correctly perceiving and interpreting the world around us. This is in line with the theory of senses and perception, formulated by Goldstein (2002). He reasons that the goal of our senses is to inform us about aspects of the environment that are necessary for our survival.

To conclude, impressions that support our survival and the development of our senses evoke aesthetic appreciation.

2.1.3. The value of aesthetics in visual datasets

Today, our environment is not only physical. Although our senses have evolved to efficiently perceive and organise the material world around us, the mechanisms that support this are also active when we engage in the digital world.

Hekkert (2006) describes that our visual system is our most prominent sensory system by miles. The main goal of the visual system is to see what our environment affords us (e.g. spot a passage, estimate distance, note obstacles so as to not bump into things). We also rely on our vision to identify things (is an object dangerous? Is it edible?). Hekkert concludes that we like to look at patterns in the world around us that facilitate our perceptual organisation. And while our chances of survival are not directly increased by accurately perceiving digital information, this visual input does appeal to the biological mechanisms that reward us for correctly perceiving the world around us. We can still acquire a sense of aesthetic pleasure through digital media.

2.1.4. Conclusions from the importance of aesthetics in visual datasets

After reviewing literature on the viewpoints various philosophers and neuroscientists, this section concludes the following definition of an aesthetic experience which will be used in this thesis: an aesthetic experience is an embodied sensory experience, which leads to a disinterested encounter which involves appraisal and induces a contemplative state in the beholder. From the by-product hypothesis, it is concluded how aesthetic experiences can be seen as our brain's reward system for correctly perceiving and interpreting the world around us.

With one to a few prompts, deep learning text-to-image models such as Stable Diffusion can generate small digital visual worlds. It is important that in the development of a product that is being deployed on such a large scale as Stable Diffusion, it is taken into account that it aligns with psychological reward systems such as aesthetic experiences.

2.2. Existing visual datasets claiming to reflect aesthetics

In addition to empirical aesthetic metrics, I also looked at alternative datasets that claim to reflect aesthetics. The datasets I included here are all from scientific papers, and thus unlike the LAION Aesthetics V2 datasets have been subject to peer review, indicating that the quality of their methodology is up to par. The fact that these datasets have been published does not directly imply that the researchers are backing their datasets with appropriate aesthetic theory. Therefore, the approaches will be examined for their applicability in the context of this thesis. For an entire overview of the existing datasets analysed in this thesis, please refer to Appendix E.

2.2.1. Objective and subjective aesthetic datasets

A number of datasets are based on objective aesthetic scales (e.g., attributes like colour, lighting, and depth of field). As previously described, these scales are outside the scope of the project. Examples of subjective scales are the TAD66K and the FLICKR-AES dataset, as proposed by He et al. (2022) and Ren et al. (2017), where participants had to rank images on a scale of *"lowest to highest aesthetics"*. In the development of the EVA dataset by Kang et al. (2020), the researchers used *"beautiful"* as a subjective aesthetic measure. The SPAQ dataset by Fang et al. (2020) and the CUHK-PQ dataset by Luo et al. (2011) look at the *"quality"* of images to evaluate their aesthetics.

2.2.2. Participant avoidance

In addition, there are several datasets based on likes or upvotes from online websites such as photography platforms. Examples include AVA by Murray et al. (2012) on which the LAION-aesthetics V2 predictor is also partially trained, DP Challenge by Datta et al. (2008), PCCD by Chang et al. (2017) and Photo.net by Joshi et al. (2011). There are several points of critique to be made about this type of approach. An image can have upvotes for many different reasons, not necessarily related to aesthetics. For example, it may be that the algorithm of the website presents some images to its users more often, which might result in more votes. It could also be that people like what is depicted, rather than the image as such. An example would be an image depicting a baby turtle. This image might get a lot more upvotes than another image with a less cute subject, which is of higher aesthetic quality.

Wu et al. (2023) aim to improve the aesthetic quality of images generated with text-to-image models. To achieve this goal the researchers trained a human preference classifier to derive a human preference score. This is done by creating a dataset using information from Dreambot, a chatbot feature of the Stable Foundation Discord. The chatbot data exported by Wu et al. (2023) describes how users generate four images using text prompts. Via preference indication, the user selects which image they find aligns most closely with their prompt. Although preference can be an interesting approach to measure aesthetics, in this instance it is not clear that the chatbot's preference data actually reflects aesthetics, or which generated image best matches the text prompt.

2.2.3. Mean opinion score (MOS)

The Mean Opinion Score (MOS) was used to develop the datasets KonIQ-10k by Hosu et al. (2019), EVA by Kang et al. (2020) and SPAQ by Fang et al. (2020). This means that they average all annotations for an image of all participants. For DP Challenge, Datta et al. (2008) reason that the average score an image receives can be seen as the estimator for its intrinsic aesthetic quality.

To develop AADB, Kong et al. (2016) had each image annotated by 5 participants, and used their average score as the ground-truth aesthetic score. Similarly, Ren et al. (2017) and Hosu et al. (2019) average the ratings of 5 crowd workers to determine the aesthetic score for each image.

2.2.4. Experts

Hosu et al. (2019) employed freelance photographers with on average more than three years of professional experience as experts in the context of image annotation on the basis of aesthetics. These experts were employed to annotate a subset of images to generate a MOS expert. The average ratings of these experts were used by the researchers as ground-truth ratings and used to compare with the MOS participants. The purpose of this approach was to validate the quality of the output of their dataset.

Kang et al. (2020) ask participants to complete a self-assessment of their experience in photography. Presumably with the purpose of seeing how this influences further ratings. From this, it could be concluded that the researchers consider photography skills as a form of expertise regarding the aesthetics of images.

Various datasets are based on ratings from users of photography websites. Murray et al. (2012), from the AVA dataset, used ratings from beginner to expert photographers, reasoning that these have a practiced eye, and these annotations are of value because they catch the way hobbyists and experts consider visual aesthetics. Chang et al. (2017), Datta et al. (2008) and Joshi et al. (2011) all use images with ratings derived from photography websites. All three studies use similar reasoning to Murray et al. (2012).

2.2.5. Conclusions of existing visual datasets claiming to reflect aesthetics

Objective aesthetic metrics do not include the human experience, which is precisely what is so central in a topic like aesthetics. Another segment of the literature centres around ratings people have provided on websites or even data derived from human-chatbot contact. In this project, it is argued that metrics and preferences provided by people *"in-the-wild"* are not sufficient to meaningfully annotate a dataset on aesthetics, due to lack of information about how the algorithms of these photography websites present images to its visitors. There is a lack of evidence that upvotes and likes actually reflect the aesthetics of images. With the chatbot data approach it is not clear whether the indicated preference reflects aesthetics or which generated image best suits the text prompt.

The existing subjective aesthetic datasets ask participants to rate the overall quality of an image or the level of aesthetics. While these are certainly steps in the right direction, it is not clear whether the *overall quality* of an image directly indicates its aesthetics. The approach of asking participants directly to rate images based on aesthetics seems more promising.

The precedent set by Kong et al. (2016), Ren et al. (2017) and Hosu et al. (2019) has five crowd workers annotate an image, and use the average of their scores.

As described earlier, several studies use photographers with similar reasoning: because they are considered to be experts in evaluating images by the researchers. For this study, this definition of experts of aesthetics in the context of annotating images is also adopted.

2.3. Literature leading to H1: Semantic Concept Activation and H2: Region

This section describes the literature which backs the hypotheses for Experiment 1, which will be elaborated in Chapter 4:

- H1: Semantic Concept Activation: exposing participants to semantic concept activation through questions about typicality and novelty, unity and variety, and relatedness, will influence their ratings of how much they like images.
- H2: Region: participants from different regions experience different images as aesthetic.

2.3.1. Empirical aesthetics

Empirical aesthetics is the academic approach to studying aesthetic experiences. It comprises how people perceive and respond to various stimuli, and tries to describe this with underlying psychological processes. Extensive research has been conducted on measuring these aesthetic experiences, and their various embodiments. The study of aesthetics has its roots in philosophy, which makes empirical aesthetics an interesting construct. Philosophy emphasises conceptual thought and excludes empirical observations and measurements. This means that empirical aesthetics dances on a tightrope of philosophical inquiry and empirical research. This interesting friction also results in the fact that some metrics may seem arbitrary. Thus, when selecting appropriate empirical aesthetic metrics, it is important to be careful that the metrics do reflect the various facets of aesthetics as defined in philosophy.

2.3.2. Objective and subjective aesthetics metrics

From a design perspective, objective features like colour, texture, form, and tone are key in aesthetics perceptions (Postrel, 2003; Sonderegger and Sauer, 2010). Subjective aesthetics refers to the reaction someone has to these objective features. It shows *"the degree to which a person believes that the (product) is aesthetically pleasing to the eye"* (van der Heijden, 2003; p544). Sonderegger and Sauer (2010) further reason that this reaction is not only influenced by the design features. Aspects of the individual like their age, gender, cultural background and personality (Crilly et al., 2004), and their mental model, also influence their subjective aesthetic experience.

This is supported by Simpson (1975). He divides aesthetic appraisal into anthropocentric (objective) and egocentric (subjective) aesthetic qualities. Anthropocentric/objective qualities are qualities that could be identified the same by everyone (for example: *'blue', 'round'*). Egocentric/subjective qualities are different across people. An example of this is my experience with the Rothko painting I described earlier in this chapter. I find this painting very intricate and moving, where my father (who prefers Italian barok painters like Caravaggio) told me that he could paint a similar work for me if I liked it so much.

This project will focus on subjective aesthetics. All aesthetics metrics that rely on objective aesthetics are beyond the scope of the project.

2.3.3. Challenges in generalising empirical aesthetics models

There are multiple aspects to consider when looking at applying empirical aesthetics models. This stems from the fact that aesthetics is originally rooted in philosophy, a field where no empirical research is conducted. The empirical studies that have been conducted are very context dependent (e.g., measuring people's attractiveness (Langlois et al., 2000; Carbon et al., 2010)). Another aspect that hampers generalisability is that research often focuses on just one flavour of the aesthetic experience (e.g. aesthetic pleasure (Blijlevens et al., 2014), aesthetic impressions (Augustin et al., 2011), beauty (Hassenzahl and Monk, 2010)).

Another consideration is that in the literature there is often no direct evidence that the scales used actually measure what they claim to measure. These scales are devised by the researchers (Blijlevens et al., 2017). It was found that many researchers use *'attractiveness'* as a scale, referring back to Page and Herr (2002), or *'beauty'* (Hassenzahl and Monk, 2010). Hung and Chen (2012) brought the dimensions of *'trendiness', 'complexity',* and *'emotion'* in contact with novelty and aesthetic preference. To measure the aesthetic pleasure of objects, Blijlevens and colleagues (2017) tested a scale on three

different continents. Their research provides the five items that capture aesthetic pleasure as *'beautiful,' 'attractive,' 'pleasing to see,' 'nice to see,'* and *'like to look at'*. As can be seen, there is no uniform consensus in the literature as to what scales can measure aesthetics.

An attempt to bridge this has been made by Berghman and Hekkert (2017) in their Unified Model of Aesthetics (UMA). The principles from which the UMA is constructed have been validated individually from each other in previous studies. However, because these separate principles describe different aspects of how we process stimuli, it is important that they are studied together. Berghman and Hekkert (2017) found that aesthetic appreciation is influenced by a combination of perceptual, cognitive and social factors operating simultaneously. The UMA describes a conceptual framework comprised of two opposing dimensions: safety and accomplishment, as depicted in Figure 2.3.

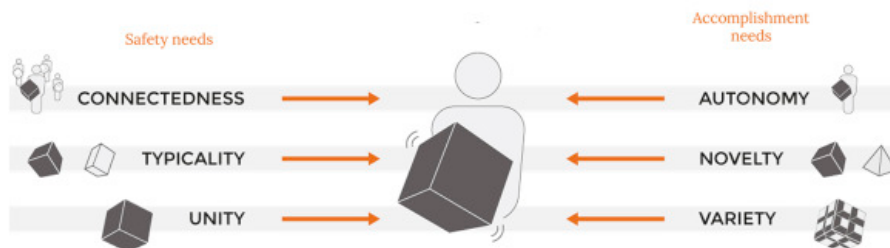


Figure 2.3: The Unified Model of Aesthetics. Safety and accomplishment needs on the perceptual, cognitive and social dimension of stimulus processing (Berghman and Hekkert, 2017).

As previously described, our underlying motivations for having aesthetic experiences have an evolutionary foundation. Berghman and Hekkert describe the duality of our safety and accomplishment needs as follows; we survive by staying safe. Yet sometimes we also have to take risks, to seek refuge and food. Survival relies on balancing these two extremes of our nature (Berghman and Hekkert, 2017). Within these dimensions are three levels of stimulus processing. A perceptual level; unity-in-variety which encaptures that humans prefer coherent and organised stimuli, but also appreciate variety to avoid sensory dullness (Berghman and Hekkert, 2017; Berlyne, 1972; Post et al., 2016). The cognitive level of the model ascribes meaning to these impressions, and to do this we rely on previous experiences. If the stimulus shows more similarity to what is known, it is smoother to process, which is why a degree of typicality is appreciated (Berghman and Hekkert, 2017). Bornstein (1989) reports that novelty is also important because it provides an opportunity for learning. The third social level of the model describes that products can be associated with groups of people, and that social value can be attributed to products as they symbolise a group identity (Kleine et al., 1993). Being part of a social group can provide certain security, but we also have a need to diversify within this group, for example, to attract partners, which is why we appreciate products that characterise us as belonging to a particular group, but also emphasise our uniqueness, ergo, connectedness and autonomy (Blijlevens and Hekkert, 2019).

As can be concluded from this description, unlike the other two levels, connectedness and autonomy are inherently intertwined with product design, and cannot be directly applied to other contexts. However, it is important to incorporate a social level, as Hekkert and Berghman (2017) argue that although the ability to appreciate aesthetics seems to manifest universally, it varies greatly over time and between different regions. Therefore, this thesis sought a context-relevant alternative to measure the social factor of aesthetics. I opted for the notion of relatedness, as described in the research of Deci and Ryan (2000). This was decided because their theory was named as the foundation for autonomous yet connected by Blijlevens and Hekkert (2019).

For a complete overview of all metrics examined in this thesis, I would like to refer you to Appendix D. This overview is not exhaustive, but can be seen as an excerpt from the literature.

2.3.4. Semantic concept activation

The construct of semantic concept activation, studied in psychology, refers to our brain's automatic spreading of semantic activations. These activations move in a network-like structure when a node is activated in your brain. An example of this is the word "tree". The moment you read this word in this report it activates a node in your mind, which activates other words related to the word tree as well (for example leaves, earth, apple, Newton). This construct describes that the moment you are exposed to a stimulus, associated concepts of this stimulus are also activated, leading to the retrieval of other associated concepts (Collins and Loftus, 1975; Chwilla et al., 1998).

Faerber et al. (2010) investigated the influence of semantic concept activation of aesthetic metrics on how participants rate images of car interiors. The researchers found that exposing participants to semantic concepts associated with aesthetic appreciation significantly impacts the dynamics of their aesthetic experience, which they rated after exposure. Participants' ability to perceive and appreciate the provided samples can be influenced by their expectations, which can lead them to focus on certain aspects of what they are evaluating.

The original experiment where semantic concept activation to affect the dynamics of aesthetic appreciation was investigated, Faerber et al. (2010) exposed participants to the following semantic concepts: *attractiveness, arousal, interestingness, valence, boredom and innovativeness*.

Arousal, valence and boredom are concepts linked to emotions. The topic of emotions is a controversial one in the world of aesthetics. Several papers argue that aesthetic pleasure is not an emotion, since an aesthetic experience is derived only from the sensory perception of the stimulus, and does not involve our personal concerns or emotions. An aesthetic response is disinterested, since it only focuses on perceiving the object itself without any ulterior motives. However, this does not mean that an aesthetic experience cannot evoke emotions. These only arise after the aesthetic experience (Hekkert and Leder, 2008; Blijlevens, Thurgood and Hekkert, 2017). Further rationale can be found in B. Due to this controversy, instead of the concepts proposed by Faerber et al. (2010), it was decided to expose the participants to the concepts of the Unified Model of Aesthetics by Berghman and Hekkert (2017).

2.3.5. Conclusions of the literature

Several empirical aesthetics metrics have been highlighted. Because this thesis focuses on aesthetic experiences, only subjective aesthetics metrics are examined. It is also important that the metrics actually measure aesthetics, and are grounded in aesthetic theory. With these criteria, several empirical aesthetics metrics were considered, and it was chosen to adhere to the Unified Model of Aesthetics (Berghman and Hekkert, 2017).

I opted for different aesthetic metrics than those used by Faerber et al. (2010) because the different levels of the Unified Model of Aesthetics were examined together by Berghman and Hekkert (2017), which was not done by Faerber et al. (2010). As described in Section 2.3.3, Berghman and Hekkert (2017) found that aesthetic appreciation is influenced by a combination of perceptual, cognitive and social factors operating simultaneously. The Unified Model of Aesthetics describes a conceptual framework comprised of two opposing dimensions: safety and accomplishment. Within these dimensions are three levels of stimulus processing: a perceptual level; unity-in-variety, a cognitive level; typicality and novelty, and a social level: connectedness and autonomy. As further described in the background of this thesis the model is slightly adjusted for relevance in the context of the annotation of images on the internet. I opted for the notion of relatedness, as described in the research of Deci and Ryan (2000). Their theory of relatedness and autonomy was referred to as a rationale for the social factors of the UMA.

Because Faerber et al. (2010) found a significant difference between participants exposed to semantic concept activation and those who were not, the following hypothesis was formulated:

H1: Semantic Concept Activation: exposing participants to semantic concept activation through questions about typicality and novelty, unity and variety, and relatedness, will influence their ratings of how much they like images.

Multiple papers state that aesthetic appreciation varies greatly over time and between different regions (Hekkert and Leder, 2008; Berghman and Hekkert, 2017). This leads to the following hypothesis:

H2: Region: participants from different regions experience different images as aesthetic.

2.4. Literature leading to H3: Aesthetic Value

This section describes the literature which backs the hypothesis of Experiment 2, which will be elaborated in Chapter 5:

- H3: Aesthetic Value: the manner in which participants are asked a question on aesthetics significantly impacts their responses.

2.4.1. Automatic linguistic behaviour

Semin and De Poot (1997-a) have conducted research into automatic linguistic behavior. This refers to the unconscious use of language that influences communication. In research, this can shape how participants are influenced by specific linguistic choices without conscious intention in answering questions. Although this research was performed in the context of interviews, it is plausible that the wording of questions in survey format also affects how participants answer questions. In a separate study, Semin and De Poot (1997-b) describe how word choice in question phrasing can achieve significantly different answers among participants.

2.4.2. Existing studies where participants are asked about aesthetic value

There are various studies which have explored the aesthetic value of visual stimuli. Bhattacharya et al. (2010) conducted a lab-controlled experiment where 15 participants rated images on a 5-point scale for aesthetic appeal, creating a dataset of natural images. Datta et al. (2006) utilized ratings from Photo.net, where images were rated on a scale of 1-7 for aesthetics. They employed machine learning to investigate the correlation between visual properties and aesthetic ratings. Zhang et al. (2014) examined the impact of depth of field (DOF) on aesthetic appeal. They asked one group of participants to rate image sets based on DOF and another group to rate them for aesthetic appeal using a continuous rating scale from 1 to 7, with 1 representing "ugly" and 7 representing "beautiful". Redi et al. (2013) conducted both lab-based and crowdsourced experiments to assess aesthetic appeal. The crowdsourcing experiment asked participants to rate images on a scale of 1-5, labeled as "bad aesthetic appeal" and "excellent aesthetic appeal". In the creation of the Simulacra Aesthetic Captions dataset, which consists of synthetic images labeled based on aesthetics, the authors utilised a similar question, asking annotators to rate the images based on aesthetic appeal (JD-P via Github, 2022).

2.4.3. Conclusions from the literature

Given the widespread use of asking participants about the aesthetic value of stimuli, it is concluded that this might be a suitable approach to measure aesthetics in the context of this thesis. Next to this, considering the impact of automatic linguistic behaviour on participant responses, the following hypothesis was formulated:

H3: Aesthetic Value: the manner in which participants are asked a question on aesthetics significantly impacts their responses.

2.5. Literature leading to H4: Ranking

This section describes the literature which backs the hypotheses for Experiment 3, which will be elaborated in Chapter 6:

- H4: Ranking: participants' rankings of image aesthetics will show significant differences when compared to their subjective ratings of image liking on a scale of 1-10.

2.5.1. Ranking modality

Nguyen et al. (2012) developed a dataset to predict attractiveness, an aesthetics adjacent concept. They investigated different aspects (face, attire, voice) for their predictions. In their crowdsourcing study, participants are presented with eight stimuli (of eight different women) at a time. The participants are asked to rank the stimuli. Yuan et al. (2023) use human preference ranking to train a language model. The language model generates responses, which are scored on their probability. The researchers have a preference reward model which presents how people rank different responses. The goal of the RRHF (Rank Responses to align Human Feedback) is to align scores assigned by the language model with reward scores from the human preference model, so the model is trained to assign higher probabilities to responses that receive higher rankings from human preference.

2.5.2. Conclusions from the literature

The literature highlights the potential of measuring aesthetics via ranking. Yuan and colleagues (2023) indicate the possible relevance of this modality to the context of this thesis. As discussed above, an extensive body of literature asks participants about aesthetic value. Therefore, participants in Experiment 3 will be asked to rank the stimuli from lowest aesthetic value to highest aesthetic value. For these reasons, the following hypothesis is formulated:

H4: Ranking: participants' rankings of image aesthetics will show significant differences when compared to their subjective ratings of image liking on a scale of 1-10.

2.6. Literature leading to H5: 2AFC

This section describes the literature which backs the hypotheses for Experiment 4, which will be elaborated in Chapter 7:

- H5: 2AFC: the alternative forced choice annotations of image aesthetics will result in significantly different outcome scores compared to participants' subjective ratings of image liking on a scale of 1-10.

2.6.1. 2AFC modality

Several studies have measured aesthetic experiences as aesthetic preference and human preference in contexts related to this thesis. Palmer, Schoss and Sammartino (2013) describe in a review on empirical aesthetics that (2) alternative forced choice is a good way to measure aesthetic preference. This approach provides a global indication of relative preference. Wu et al. (2023) connect aesthetics of a dataset of images with human preference. Swanson et al. (2012) apply 2AFC in their study of the composition preference of visual stimuli, obtaining a high-quality model capturing the preference of a large amount of people. Bara et al. (2021) study aesthetic judgments of visual stimuli using a dual-task paradigm, combining a working memory task with 2AFC for aesthetic evaluation of paintings. Bıyık et al. (2020) propose a preference-based learning approach, using human feedback to compare robot arm trajectories. Sadigh et al. (2017) present an algorithm efficiently learning reward functions based on human preferences (inferred via 2AFC).

2.6.2. Conclusions from the literature

The research described above indicates that alternative forced choice could be an appropriate way to measure the aesthetics of images. This is why the following hypothesis is formulated:

H5: 2AFC: the alternative forced choice annotations of image aesthetics will result in significantly different outcome scores compared to participants' subjective ratings of image liking on a scale of 1-10.

2.7. H6: Content VS Overall Image

This section describes the literature which backs the hypotheses for Experiment 5, which will be elaborated in Chapter 8:

- H6: Content VS Overall Image: when participants rate their liking of the image content, there will be significant difference with when they indicate the image liking of the overall image.

2.7.1. High internal consistency between subjects for a subjective experience

From a dataset point of view, it makes sense that a high Cronbach's alpha (inter rater reliability) is desirable. When different people agree with each other about which images they consider aesthetic, this can potentially say something about the generalisability of the dataset. When there is a general consensus among people regarding their aesthetic preferences, it indicates that the dataset can be deployed with a certain level of confidence for training Stable Diffusion, a widely used product. Empirical aesthetic literature also shows that a very high inter rater reliability is often found in such experiments (Berghman and Hekkert, 2017; Blijlevens et al., 2017).

Literature shows that aesthetic experiences are highly subjective (Simpson, 1975; Hekkert and Leder, 2008). From this aspect, it is fascinating that in the initial experiments in chapters (4, 5, 6, 7) and the studies of Berghman and Hekkert (2017) and Blijlevens et al. (2017) find such high internal consistencies between subjects. The stimuli of these experiments that find high Cronbach's alpha values are products (e.g. coffee machines, bicycles). For this thesis buildings are used as an image class, which is also a functional topic. Kant describes two types of beauty, free and dependent (2000; 1790). Here, free beauty is seen as a disinterested pleasure derived from the contemplation of the stimulus itself, without reference to its utility. Dependent beauty refers to the beauty we ascribe to a stimulus that is judged by how well it corresponds to its intended purpose. This can be seen as a state of functional contemplation. It may be that these functional stimuli mainly trigger functional contemplation in the participants, a state in which the affordances of the stimulus are primarily appreciated, rather than the aesthetic characteristics as occurs in aesthetic contemplation. This rationale is supported by the qualitative analysis presented in Section 5.3.3, which indicates that the content of images may influence the ratings participants ascribe to it.

2.7.2. Conclusions from the literature

The rationale described before indicates that it is of interest to investigate more about whether the exact topic of images possibly influences participant ratings. For this reason, this hypothesis is formulated:

H6: Content VS Overall Image: when participants rate their liking of the image content, there will be significant difference with when they indicate the image liking of the overall image.

2.8. Ethical considerations

This section will go into more detail on the use of crawled images, which raises ethical issues, such as the lack of explicit consent from the individuals whose images are included. To maintain ethical standards, only images without identifiable individuals were retrieved from the LAION dataset for this thesis. Copyright considerations also play a role. In addition, the ethics of crowdsourcing is considered in this project by fair compensation, transparent communication of acceptance/rejection terms and explicit consent from workers, ensuring ethical treatment during the studies.

2.8.1. (Lack of) consent

The internet is an interesting space, giving access to a wide variety of people and providing them with the tools that allow them to shape this digital world however they see fit. Yet, because of this unrestricted freedom, users frequently fail to evaluate the consequences of sharing images without taking accessibility into account. Despite the fact that users of social media platforms typically agree to terms and conditions that allow third parties access to the data they upload, a review of the pictures gathered by LAION reveals an obvious lack of foresight on the part of some users who uploaded pictures taken in private settings. In addition, a lot of websites—including blogs—serve as places for people to share their own images aside from social media. While it can be argued that by uploading images to the inter-

net you automatically give consent for the images you post to be used by third parties, it is imaginable that these users did not foresee that their images would be included in large-scale datasets via crawling.

General Data Protection Regulation (GDPR) describes European privacy legislation that restricts the use of its citizens' personal data. The implications of this legislation for organisations such as LAION, which use crawled images to create datasets, are that EU citizens should be able to have images that contain personally identifying information removed from the set. As a result, LAION has a takedown option if personally identifiable images are in the set (Schumann, 2022). However, this is something you, as a user, must pursue yourself.

In order to be as ethically conscious as possible within the context of this thesis, only URLs of images not related to people were retrieved from the existing dataset. In files available for downloading from the LAION website, a script can be used to retrieve images by searching for words from their alt-text. For this thesis, only inanimate objects were searched (e.g. roads, coffee machines, buildings). This does not mean that there are no people depicted in the images, but that they were not specifically searched for (e.g. child).

The LAION 5b dataset includes URLs to images of people who did not give explicit consent for their likeness to be included in said dataset. While LAION claims to have created the dataset for scientific purposes (Schumann, 2022), this may raise ethical questions. Hayes and Kuyumdzhieva (2021) prepared a report for the European Commission on ethics and data protection in research. This report is also held up by the TU Delft Ethics committee (TU Delft, n.d.). They expand on the use of "open source" data in research. The committee indicates that even though data is publicly available, it does not mean that it may be used without limit. The moment open source data contains personal data about identifiable persons, new records are created, and personal information about these persons is processed.

To ensure that this thesis has no ethical breach in terms of deploying the crawled images, it was decided to exclude stimuli that depict recognisable people.

2.8.2. Copyright

Researchers and companies concerned with AI reason that crawled images that may be under copyright fall under the *fair use doctrine* (Vincent, 2022). This fair use doctrine stems from U.S. law, and describes that copyrighted images can be used for limited use without owning the license. Purposes that may fall under this include news reporting, teaching and research (U.S. Copyright Office, 2023).

The term "*AI data laundering*" refers to the practice where a commercial party withdraws itself from crawling data and outsources it to a non-profit organisation, for example, with a research purpose. This party can do more under the guise of research with regards to copyright law (Vincent, 2022). In this instance, these researchers launder the images, which are then used by the commercial party to train their models. It could be argued that the training in Stable Diffusion involves AI data laundering practices.

This thesis uses the LAION dataset and training of Stable Diffusion as a case study, and therefore is not directly involved in potentially questionable legal issues. However, it is important to consider copyright in the development of the stimulus set used to conduct the research here. Watermarked images obviously protected by copyright will not be involved in the stimulus set.

2.8.3. Crowdsourcing

Crowdworkers are dependent on requesters, often working for small amounts of money ("*nickels and dimes for tasks between 5-10 minutes*") as stated by Buhrmester et al., (2011). On platforms like Amazon's Mechanical Turk (AMT), after a worker has performed, requesters can choose to accept the work, or reject it, which allows them to keep the work completed for free. Platforms such as AMT track how often crowd workers' work is rejected, and communicate this to potential requesters. So by rejecting work, not only do workers not get paid, but it can also increase the likelihood that they will have fewer work opportunities in the future (Silberman et al., 2010; Chittilappilly et al., 2016).

This thesis will strive to provide fair compensation to the crowdworkers by rewarding them according to the recommended pay by Prolific, the crowdsourcing platform which will be used to conduct the experiment. The accept/reject terms will be communicated to the workers in advance, for transparency.

Explicit consent will also be requested from the workers prior to the experiment. First, the purpose of the experiment will be clarified. Next, it will be communicated what is expected of them. It will also be emphasised that they can leave the experiment at any time. Finally, it will be described how exactly the generated data will be handled, and the steps that will be taken to anonymise it. For the full consent form, see Appendix F.

3

Experimental design

The chapters 4, 5, 6, 7 and 8 will each highlight one experiment, an overview of which can be found in Section 1.4. Although the five experiments conducted in this thesis are all distinctive, they also have common denominators. To avoid repetition, these common denominators will all be described in this chapter. This implicates that the chapters describing the experiments do not always cover all standard topics (e.g. explanation on the participants).

3.1. Introduction

The experiments conducted for this thesis are exploratory in nature. This entails that no conclusive scientific value can be prescribed to the outcomes. They serve merely to indicate whether one of the directions could possibly be more appropriate than the current method or shed light on the annotation for aesthetics in this context.

As described in Section 1.4, to attempt to answer the research question of this thesis, there are two levels of inquiry within comparing the current situation with alternative annotation methods. Alternative task design, which looks at instructions associated with image annotation, and alternative metrics, which focuses on the measurement of aesthetics within the annotation process. The alternative task design aims to determine if participants' image evaluations based on liking change when their attention is directed towards underlying aesthetic principles of the image. However, experiment 1 reveals that these semantic concepts, when activated, do not significantly impact participants' liking ratings. Consequently, the influence of changing the metric directly itself is investigated. By examining alternative metrics, liking evaluations are compared with metrics that are backed by literature as measures of aesthetics. Although the comparison of different modalities may seem unconventional, it has been done in existing literature, e.g. different question formulations (Semin and de Poot, 1997-b), rating vs. ranking (Alwin and Krosnick, 1985), rating vs. 2 alternative forced choice (Yannakakis and Hallam, 2011). In addition to these four experiments, a fifth experiment is conducted to examine how participants evaluate image content compared to their overall liking of the images.

As depicted in Figure 1.3, if an experiment suggests that an alternative approach may be more suitable than the control treatment, a subsequent experiment will be carried out to determine which treatment aligns the results of crowdworkers more closely with those of aesthetic experts. The following points are considered here:

- The results of the experiment produce significantly different outputs from the control treatment.
- The internal consistency is adequate.

When comparing different annotation methods, an adequate Cronbach's alpha is an indication that the metric measures consistently between subjects. Appropriate inter-rater reliability is desirable because it is an indication that a method generates consistent and dependable results. This is useful when we want to create a dataset that reflects the aesthetic experiences of a wide range of people, so that it can be deployed to train Stable Diffusion. Although aesthetic experiences are subjective, high Cronbach's

alphas are not uncommon. Multiple papers that study various aspects of aesthetic experiences exhibit very high Cronbach's alphas, e.g. 0.76 - 0.94 (Berghman and Hekkert, 2017), 0.83 - 0.98 (Blijlevens et al., 2017). The rule of thumb that an alpha value > 0.7 for an acceptable alpha value is adhered to in this thesis.

No treatment is significantly differed to the control treatment and had adequate internal consistency. Therefore, the follow-up experiment was not conducted. Were this to be the case, the follow-up experiment would be as follows: both the control treatment and the experiment treatment are administered again, but this time not with crowdworkers but with aesthetic experts. For the aesthetic experts, the study would use photographers with experience, as several other studies have done, (described in Section 2.2.4). Next, I would look for which treatment the crowdworkers' results align more with expert results. This benchmark aims to gain additional insight into the effectiveness of both annotation methods in capturing aesthetic experiences.

It is important to note that these experiments barely scratch the surface of everything there is to be investigated in the context of this subject. These approaches were chosen because they are the most promising alternatives suggested in the literature. Recommendations for future research are highlighted in Section 10.4.

3.2. Stimulus sets

The following section describes the stimulus sets used, focusing on the rationale for choosing the image class "*buildings*". Ethical considerations include excluding NSFW content and images with recognisable people. Moreover, two iterations of the stimulus are used, with participant surveys to ensure a controlled distribution of aesthetic values.

3.2.1. Extracting images from the LAION dataset

In this thesis, careful consideration is made regarding the stimulus set. From the context of this thesis, it makes sense that the images are derived from the LAION dataset before the aesthetics predictor has divided them into buckets. When retrieving the images from the Parquet files provided by LAION, all NSFW images are excluded for the sake of ethics. Also, only images with height > 1024 and width > 1024 were retrieved to ensure that all images are of decent quality. After selection, all images are resized to the same dimensions to avoid bias. This is done manually to ensure that the main topic of the image is still in focus.

3.2.2. Considerations for image class selection

It is important that the participants are as capable as possible of assessing the aesthetics of the images as such. This influences the selection of an appropriate image class by ensuring that the images do not contain too distracting topics that might interfere with this. An example of this is an image class I explored: roads. This class could possibly interfere with aesthetic ratings on the overall stimulus because participants might be distracted by an image of a road with a woman in the forefront.

What is also important is that the image class is known worldwide. Next to this it should be plausible that images from around the world are available of the class. This is important because there may be a novelty/typicality (Berghman and Hekkert, 2017) effect, which may affect the participants' aesthetic experience.

After exploring various image classes (roads, coffee machines, tables, among others), *buildings* emerges as the image class that meets the requirements mentioned above.

3.2.3. Ethical considerations for the stimulus set

As discussed in the Section 2.8, certain decisions were made regarding the stimulus set for ethical reasons. As stated above, I excluded all NSFW content. Additionally, I only retrieved alt-text from inanimate objects, and thus did not look for images depicting people. I decided to do this because not everyone who is depicted in an image on the internet has actively given consent for said image to be published. Additionally, people do not actively consent for Common Crawl to include images with their

likeness in large-scale datasets. This does not mean that people were never depicted in the images retrieved from the LAION Parquet files, but that they were not actively searched for. There are no images included in the stimulus set where people are recognisable. Also, no images were included which were watermarked for copyright violation.

3.2.4. Two iterations of the stimulus set

When creating a stimulus set, there is a trade-off between specificity and diversity. For this thesis, we need to be specific enough to not compare apples to oranges, but if we get too specific, we risk losing diversity in the stimuli, which can negatively impact the ability to measure aesthetics and move us away from the thesis' context - the LAION dataset. Figure 3.1 shows how each version of the stimulus set is deployed in this thesis.

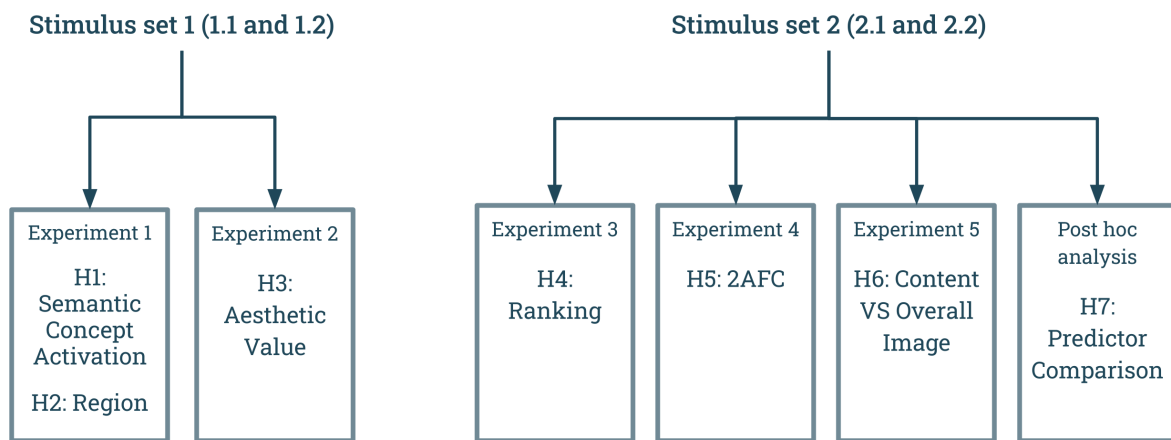


Figure 3.1: Which hypothesis is tested with which stimulus set.

For both versions of the stimulus set, the control treatment was administered anew. The trade off discussed above has been shifted after the first two experiments, to a more clean set. As can be seen, the stimulus set 1 consists of the sub-stimulus sets 1.1 and 1.2 and stimulus set 2 consists of the sub-stimulus sets 2.1 and 2.2. This was done because each stimulus set is split into two groups of ten. Each participant will annotate ten images. Both versions will be discussed below.

3.2.5. Stimulus set 1 (1.1 and 1.2)

As described above, stimulus set 1 contains 20 images of buildings. This group is diverse, there are both images of real buildings and images of things that depict buildings (e.g. a Lego building). Because the experiments in this thesis will use a small amount of images, it has been decided to make sure that the set contains both images that score low and images that score high on aesthetic value. For this stimulus set, we as researchers ourselves have controlled for a stimulus set that is distributed by aesthetic value. However, during the analysis of Experiment 1, I found that our expectations of the aesthetic value distribution did not match the participants' evaluations. Therefore, I decided to develop a second stimulus set. An overview of the images can be found in Appendix N.

3.2.6. Stimulus set 2 (2.1 and 2.2) - a controlled distribution of aesthetic value

As described above, because of the small number of stimuli, I want to control for the aesthetic value distribution. For stimulus set 2 this was established by conducting two surveys (n=60 and n=12). In these surveys, I identified images that received very low, average, and very high aesthetic scores.

For this I recruited participants from my personal network, specifically second-year master's design students from TU Delft, both Dutch and international.

Summary of the prior surveys

In the first survey (n=60), participants are asked to categorise the images into low, medium, and high aesthetic categories. They are also provided with the option to indicate if an image should not be included in the stimulus set.

The second survey (n=12) is conducted to identify images that receive the most extreme ratings. Participants are asked to rate the images on "how much do you like this image on a scale from 1-10?".

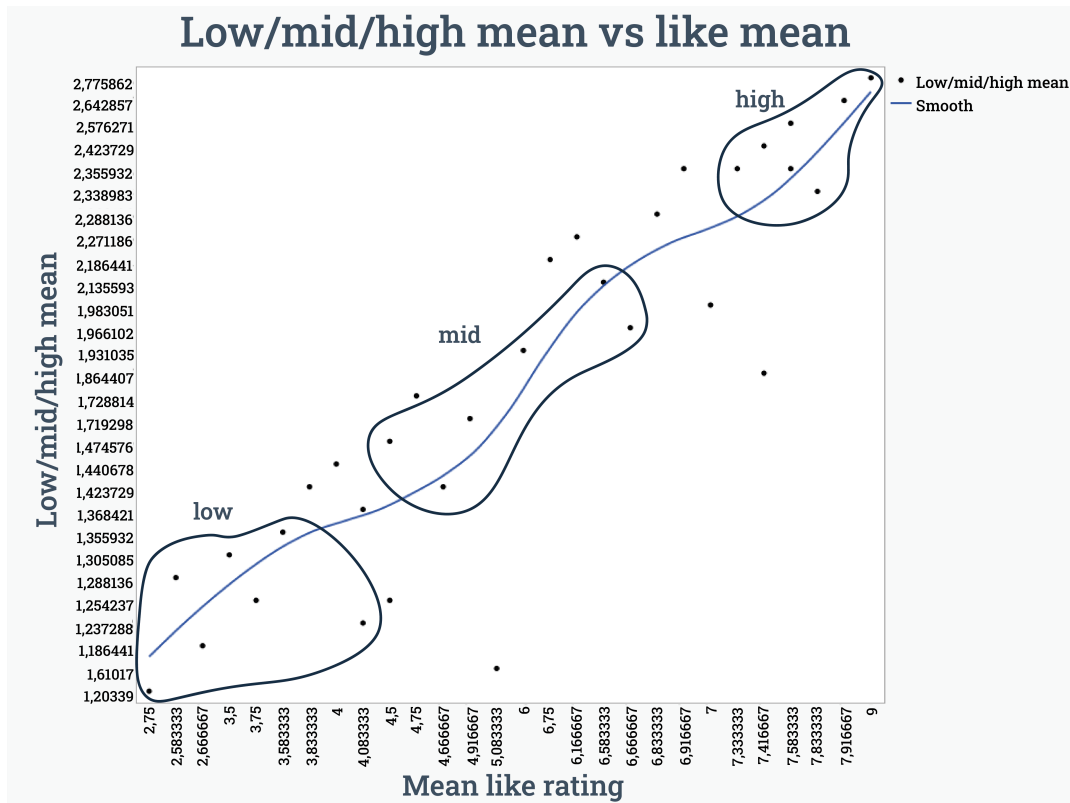


Figure 3.2: Means of both surveys plotted, where each dot represents an image, and the blue line shows a smooth curve through the data. X-axis: mean like ratings. Y-axis: mean score of low/mid/high.

The results for each image are plotted with the mean of survey 1 on the y-axis and the mean of survey 2 on the x-axis, as depicted in Figure 3.2. Based on this scatter plot, images are placed in three aesthetic buckets (low, mid, high aesthetics). For the complete setup and analysis of both surveys, see Appendix G. Figure 3.3 shows an overview of the resulting two subsets of stimulus set 2.

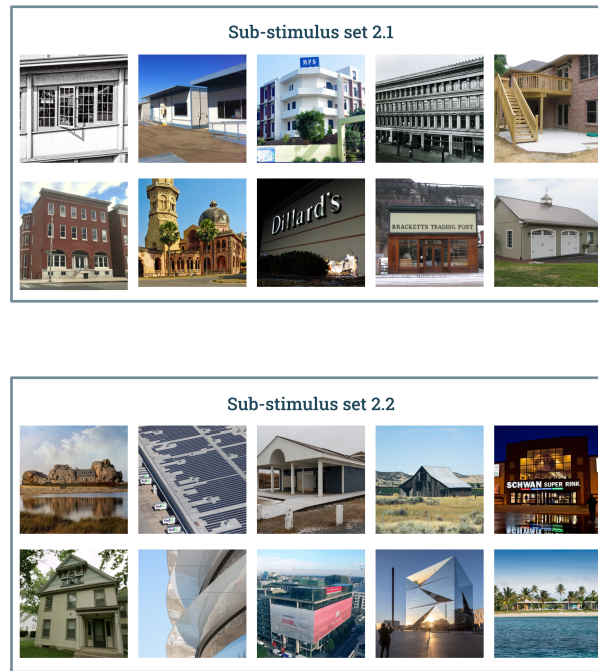


Figure 3.3: Overview of the developed sub-stimulus sets.

3.3. Participants

As described in Chapter 2, the precedent set by Kong et al. (2016), Ren et al. (2017) and Hosu et al. (2019) has five crowd workers annotate one image. To account for crowdworkers who may need to be excluded from the study and to be on the safe side, I aimed at $n=10$ for each group from the experiments, as depicted in the tables included below. The participants excluded from analysis were all rejected for not completing the survey. The control treatment was administered once for each stimulus set.

3.3.1. Overview participants stimulus set 1

Experiment 1: semantic concept activation with the Unified Model of Aesthetics

SCA treatment	Group 1	Sub-stimulus set 1.1	$n = 10$
SCA treatment	Group 2	Sub-stimulus set 1.2	$n = 10$

Table 3.1: Overview participants Experiment 1.

Experiment 2: rating images on aesthetic value

Aesthetic value treatment	Group 1	Stimulus set 1.1	$n = 8$
Aesthetic value treatment	Group 2	Stimulus set 1.2	$n = 9$

Table 3.2: Overview participants Experiment 2.

Control treatment

Control treatment	Group 3	Stimulus set 1.1	$n = 10$
Control treatment	Group 4	Stimulus set 1.2	$n = 10$

Table 3.3: Overview participants control treatment.

3.3.2. Overview participants stimulus set 2

Experiment 3: Ranking images on aesthetic value

Ranking treatment	Group 1	Stimulus set 2.1	n = 10
Ranking treatment	Group 2	Stimulus set 2.2	n = 9

Table 3.4: Overview participants Experiment 3.

Experiment 4: Image Preference with Two Alternative Forced Choice

2AFC treatment	Group 1	Stimulus set 2.1	n = 10
2AFC treatment	Group 2	Stimulus set 2.2	n = 10

Table 3.5: Overview participants Experiment 4.

Control treatment

Control treatment	Group 3	Stimulus set 2.1	n = 10
Control treatment	Group 4	Stimulus set 2.2	n = 10

Table 3.6: Overview participants control treatment.

To account for language barriers in crowdsourcing experiments, Feitosa et al. (2015) find that it is important to limit research written in English to crowdworkers with an IP address in an English-speaking country. Since it is important to have participants with diverse cultural backgrounds when examining relatedness, the experiments performed in this thesis limit themselves to participants who are fluent in English. All participants are recruited through Prolific, a crowdsourcing platform.

3.4. Task planning

Several researchers indicate that clear task description can affect worker outcomes (Daniel et al., 2019; Chittilappilly et al., 2016; Silberman et al., 2010). Thus, it is important to ensure that the tasks are clear and that the survey is not buggy. To make sure the experiments are up to par, pilot studies will be performed where it is checked that the tasks are straightforward and all the interactions are working accordingly.

3.5. Control treatment “how much do you like this image on a scale from 1 to 10?”

The first four experiments are compared to the current aesthetics predictor procedure described on the LAION website (Schuhmann, 2022). All that is known of this is that Schuhmann (2022) noted that to create LAION-Aesthetics, different models were trained to predict the rating that people gave when asked “*how much do you like this image on a scale from 1 to 10?*”. This section will describe how the control treatment replicated this as closely as possible. The details of how the different treatments compare to the control treatment will be discussed in subsequent chapters.

3.5.1. Structure of the control treatment implemented in this thesis

1. **Training questions:** as suggested by Daniel et al. (2019), it may improve the performance of crowd workers if they receive specific training in the tasks to be performed. This can help them become familiar with the required terminology and skills. Prior to annotating the images from the stimulus set, participants practice with the format using two unrelated stimuli.
2. **Rating:** after the training questions, the participants are split into two groups. Both groups will annotate one part of the stimulus set (1.1/1.2 and 2.1/2.2). The images are presented to each participant in randomised order. Both groups are asked to rate ten stimuli using the following question, “*how much do you like this image on a scale from 1 to 10?*”. See figure 3.4 for the rating interface.

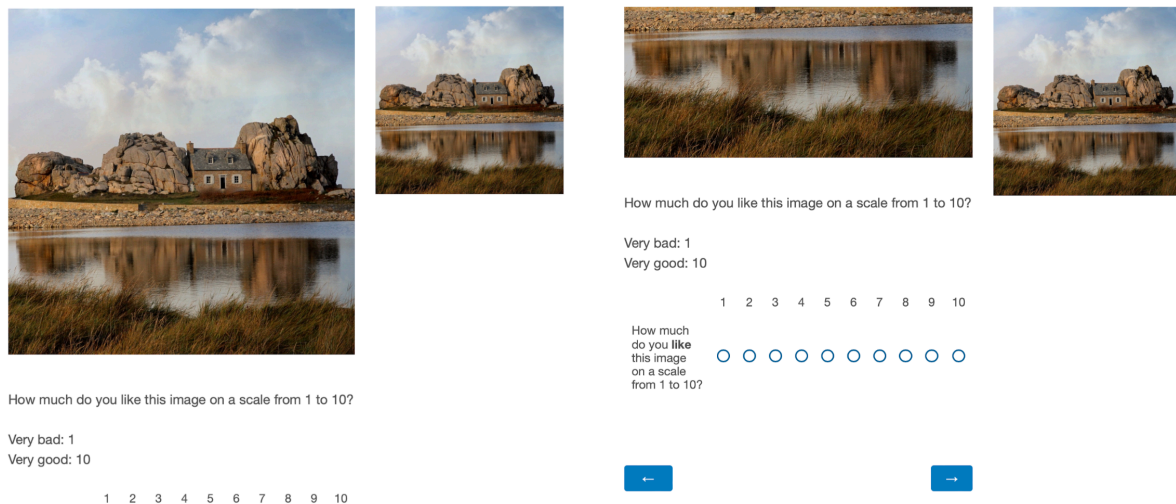


Figure 3.4: Left: the interface when the participants open the Qualtrics survey. Right: the interface when the participants scroll to the rating question.

- 3. Internal consistency:** to measure consistency, participants were asked to rate one image twice (through the randomised sequence of the 10 images from the stimulus set). The internal consistency within the responses of a participant refers to the degree of consistency of the answers the participant provided over the two questions for this image. This way the reliability and coherence of responses can be evaluated. The first time the control treatment was administered (for stimulus set 1), no internal consistency was collected at participant level. This was added for the second stimulus set.

3.5.2. Analysis

Internal consistency between subjects

To assess how consistently the raters are in agreement, I calculate the Cronbach’s alpha, as shown in table 4.2. I follow the commonly used rule of thumb for interpreting alpha values, where a value > 0.7 is considered acceptable for reliable results. In other words, if the test demonstrates 70 percent reliability or higher, it indicates appropriate internal consistency among the participants (Brown, 2002). To measure if participants from the 4 experiment treatments become more reliable raters, the inter-rater reliability measured by Cronbach’s alpha should increase. To analyse the internal consistency of the treatments on group level, the Cronbach’s alpha of all treatments and for both stimulus set groups should be examined. The alpha values for the control treatment are displayed in table 3.7.

Which sub-stimulus set	alpha-value
Stimulus set 1.1	$\alpha = 0,9431$
Stimulus set 1.2	$\alpha = 0,7582$
Stimulus set 2.1	$\alpha = 0,7822$
Stimulus set 2.2	$\alpha = 0,8494$

Table 3.7: Control treatment Cronbach’s alpha results.

Internal consistency within subjects

In the control treatment (2), for stimulus sets 2.1 and 2.2, participants are exposed to one image twice in a randomised loop order. I examined the correlation between participants’ ratings for this image and its corresponding internal consistency twin image in a linear regression, for both stimulus set groups, 2.1 and 2.2. The 2.1 stimulus set shows a strong positive (slope = 0.9032258) significant correlation (p-value = 0.0415). The 2.2 stimulus set also shows a strong positive (slope = 1.011194) significant correlation (p value = 0.0002).

Qualitative results

Ex post, I examined the distribution between sub-stimulus set 2.1 and 2.2. Figure 3.5 shows this distribution of assigned ratings per sub-stimulus set. Looking at this from a qualitative point of view, it can be concluded that the ratings of sub-stimulus set 2.1 are more closely related than sub-stimulus set 2.2. This is an indication that the controlled aesthetic distribution as discussed in Section 3.2.6 may not be equally balanced between the two groups of participants from each experiment.

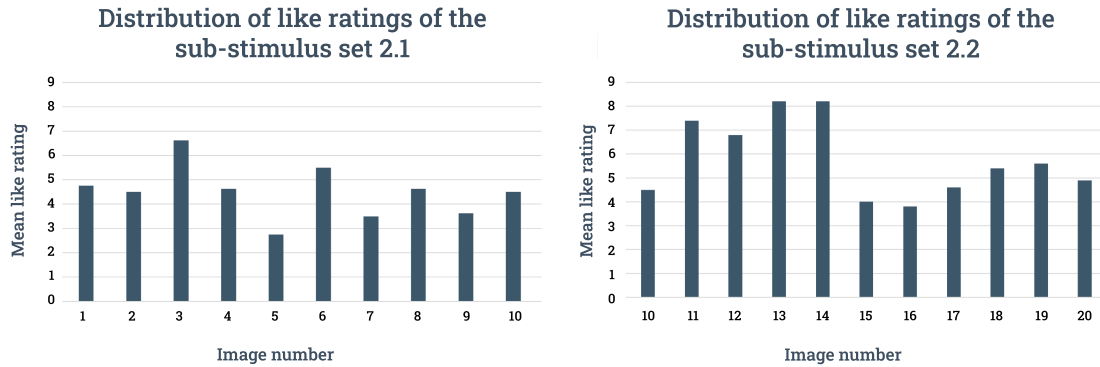


Figure 3.5: Distribution of assigned ratings per sub-stimulus set. X-axis: image number. Y-axis: mean like rating per image.

It is expected that an equal distribution of aesthetic value of the stimuli would not produce completely different results. Nevertheless, it is good to reflect on this, and it has also been incorporated as a limitation of the experiments conducted in this thesis (Section 10.3).

4

Experiment 1: semantic concept activation with the Unified Model of Aesthetics

4.1. Introduction

In experiment 1 I compare participants who have been exposed to semantic concept activation (SCA) (Faerber et al., 2010) with the Unified Model of Aesthetics (UMA) (Berghman and Hekkert, 2017) with ratings of image liking on a scale from 1-10 (Schuhmann 2022).

The motivation for this experiment arises from existing literature, which describes that exposing participants to semantic concepts related to aesthetic appreciation can significantly influence aesthetic perception and evaluation. As also described in Section 2.3.4, the original research on semantic concept activation for aesthetic evaluations suggests some concepts to expose participants to, but in this thesis it was opted to implement other concepts. This was decided because half of these concepts are related to emotions. As researchers have argued in various papers, aesthetic pleasure is not considered an emotion due to the disinterested nature of aesthetic experiences. These experiences focus solely on the perception of stimuli (Hekkert and Leder, 2008; Blijlevens, Thurgood, and Hekkert, 2017). It is important to note that while aesthetics may evoke emotions, these emotions are not intrinsic to the experience itself. Due to the the controversies around emotions in aesthetics, the concepts introduced by Faerber et al. (2010) are disregarded, and the decision was made to present participants with the concepts from Berghman and Hekkert's Unified Model of Aesthetics (2017). Literature on the UMA suggests that aesthetic appreciation is influenced by three dimensions: typicality and novelty, unity and variety, and a social dimension, made context appropriate by interpreting it as relatedness, as described in Section 2.3.3.

Continuing on these insights, the H1: Semantic Concept Activation posits that exposing participants to semantic concept activation through questions related to typicality and novelty, unity and variety, and relatedness will influence their likability ratings of images. H2: Region posits that participants from different geographical regions will have different results for image liking.

In the method section it is outlined how the participants are recruited, what stimulus set is used, and it describes the survey that used Qualtrics and was run on Prolific. It also describes the control treatment and the SCA treatment which uses semantic concept activation with the UMA and a question on image liking. In the analysis section, both treatments are used to address the three hypotheses. In addition, the internal consistency is examined.

The purpose of this study is for exploration rather than validation. Therefore, the number of stimuli and participants is intentionally small.

4.2. Experiment 1 - hypotheses and expected outcomes

These hypotheses is based on literature described in Section 2.3.

H1: Semantic Concept Activation

Exposing participants to semantic concept activation through questions about typicality and novelty, unity and variety, and relatedness, will influence their ratings of how much they like images.

The null hypothesis posits that there will be no significant difference in outcome scores when participants are exposed to semantic concept activation with questions that describe dimensions of aesthetics.

Expected outcome

I expect that the outcome scores will exhibit significant differences in response to exposure to semantic concept activation with questions that describe aesthetic dimensions. This is expected because Faerber et al. (2010) found significant differences in a similar situation.

H2: Region

Participants from different regions experience different images as aesthetic.

It is important to note that region will be measured as a confounding variable, which implicates that I will not control for it.

The null hypothesis suggests that there will be no significant difference in the perception of image-relatedness among participants from different regions.

Expected outcome

The hypothesis posits that participants from different regions may have distinct perceptions of image-relatedness. With the statements made by Hekkert and Berghman (2017) as indication, it is expected that participants from different regions will exhibit differences in their judgements of image aesthetics, indicating that region may influence how individuals experience image' aesthetics.

4.3. Confounding variables

The literature indicates that several aspects can potentially influence participants' aesthetic experiences. These aspects are tested in Experiment 1 as independent confounding variables. To ensure that the confounding variables do not influence the dependent variable, the questions regarding the confounding variables are asked after the treatment.

The following confounding variables are measured:

1. Demographical information (Chamorro-Premuzic, Furnham and Reimers, 2007)
2. Self-efficacy (Urdan and Pajares, 2006; Bandura, 1977)
3. Socioeconomic status (Mcmanus and Furnham, 2006)
4. Education level (Mcmanus and Furnham, 2006; Schneider, 2013)
5. Noise levels (Wang et al., 2020; Nemecek and Grandjean, 1973)
6. Working environment (Szubielska et al., 2021; Wang et al., 2020)
7. Social context (Hesslinger et al., 2017)
8. Colourblindness (Kang et al., 2020)
9. Aesthetic fluency (Cotter et al., 2023; Smith and Smith, 2006))
10. Aesthetic attitude (Mcmanus and Furnham, 2006)

In Appendix H an overview of all confounding variables, the literature that provides an indication that the variable could potentially affect the experiment outcomes, their corresponding hypotheses and the exact way of how they are measured. It also discusses confounding variables for which it was decided to exclude them from Experiment 1, and the motivation for this.

4.4. Method

4.4.1. Materials

Stimulus Set

For this experiment, stimulus set 1 was used, consisting of sub-stimulus sets 1.1 and 1.2. For further description, please refer to Section 3.2.5.

Survey

Experiment 1: semantic concept activation with the Unified Model of Aesthetics

The interface of the survey of the SCA treatment is shown in Figure 4.1. The interface of the control treatment can be found in Figure 3.4.

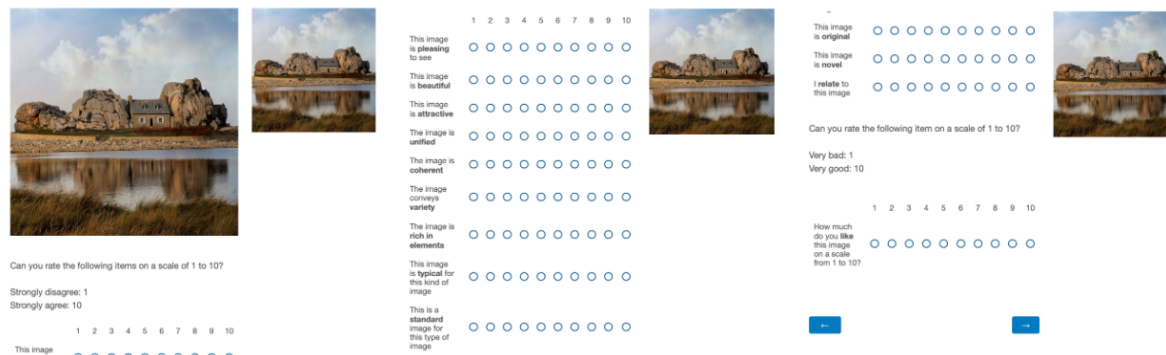


Figure 4.1: SCA treatment questions. Left: the interface when the participants open the Qualtrics survey. Middle: the interface when the participant scrolls down 1. Right: the interface when the participants scrolls down 2.

4.4.2. Procedure

Experiment 1 included two treatments, the SCA treatment and the control treatment, that were varied between participants. The key variable that differed between the treatments was the exposure to semantic concept activation (SCA) using the Unified Model of Aesthetics. In the SCA treatment, participants were presented with 12 items directly derived from Berghman and Hekkert's (2017) research. For rationale why the original semantic concepts of Faerber et al. (2010) were adapted, please visit Section 2.3.4. The connected yet autonomous level of the UMA, based on Deci and Ryan's relatedness and autonomy, was adapted to "I relate to this image" to make it relevant to this context (Berghman and Hekkert, 2017; Deci and Ryan, 2000). Before the participants took the actual treatment, they were presented with two training tasks where everything was the same as in the experiment question, but different stimuli were used.

The items rated in the SCA treatment include:

1. This image is **pleasing** to see
2. This image is **beautiful**
3. This image is **attractive**
4. This image is **unified**
5. This image is **coherent**
6. This image conveys **variety**
7. This image is **rich in elements**
8. This image is **typical** for this kind of image
9. This is a **standard** image for this type of image
10. This image is **original**
11. This image is **novel**
12. I **relate** to this image
13. How much do you **like** this image on a scale of 1-10?

In the research conducted by Berghman and Hekkert (2017), the items of the UMA were rated on a scale of 1-7. To make the model comparable to the current context (control treatment), this scale has been adapted to a range of 1-10. The extreme values of the Likert scale (1 and 10) have been labelled as "*strongly disagree*" and "*strongly agree*," respectively. The ratings are presented in a matrix format, as depicted in Figure 4.1.

The stimulus set was divided into two sets of 10 images. For both treatments, the two sub-stimulus sets were rated by two groups of participants. Participants were randomly assigned to either treatment and either stimulus set.

With this procedure, the influence of semantic concept activation on participants' image liking was examined in this experiment.

Below, Figure 4.1 shows a table summarising the data that will be collected for each participant. The second block explains more about the experiment variables. After this, the participants proceeded to the third block, where the confounding variables were evaluated.

Questions in order of appearance for participants	SCA treatment	Control treatment
12 items asking participants to evaluate the stimulus on aesthetic dimensions	YES	NO
<i>How much do you like this image on a scale from 1-10?</i>	YES	YES
Demographical information provided by Prolific (age, sex, first language, current country of residence, nationality, country of birth, student status, employment status)	YES	YES
Education level	YES	YES
Disturbed by noise during crowd work	YES	YES
Physical working environment	YES	YES
Social environment	YES	YES
Colour blindness	YES	YES
Aesthetic fluency	YES	YES
Aesthetic attitude	YES	YES

Table 4.1: Table with the data to be collected per participant for Experiment 1.

4.4.3. Analysis

The analysis for this experiment was performed using JMP software.

H1: Semantic Concept Activation

Exposing participants to semantic concept activation through questions about typicality and novelty, unity and variety, and relatedness, will influence their ratings of how much they like images. To analyse whether there was a significant difference between the treatments per image, a T-test is conducted. The complete results can be found in Appendix I. The T-tests indicate that for this data the p-value was not significant for any image, for all p-values > 0,1213.

To evaluate the variability among the participants' ratings, an F-test was conducted to examine the variation in variance across different images. This analysis compared groups that are exposed to the same image set. For each individual image, an F-test was performed.

The standard deviations of both treatments were calculated, resulting in means of 1.96691715 (SCA treatment) and 1.90576205 (control treatment), respectively. To determine if there was a significant difference in variance between the treatments, a T-test was performed. The results indicate that the difference in standard deviation between the treatments was not statistically significant (slope = 0.1349461, p-value = 0.4495). Please refer to Appendix I for a table depicting the complete results.

Internal consistency between subjects

To examine the inter-rater reliability, the Cronbach's alpha was calculated, as depicted in table 4.2. The commonly used rule of thumb of interpreting alpha values will be adhered to, where acceptable reliability = >0.7 . This means that if the test is of 70 percent reliability it is interpreted as appropriate internal reliability between subjects (Brown, 2002). When comparing the alpha values of the control treatment to the SCA treatment, it is not the case that this reliability was systematically increased in the SCA treatment. Both treatments showed satisfactory alpha values.

SCA treatment	Sub-stimulus set 1.1	$\alpha = 0,8871$
SCA treatment	Sub-stimulus set 1.2	$\alpha = 0,8209$

Table 4.2: Experiment 1 Cronbach's alpha results.

H2: Region

Participants from different regions experience different images as aesthetic.

To examine H2: Region, a comparison was made between participants from different regions, to investigate if they experienced different images as related. The region was measured as a confounding variable, so it is not controlled for beforehand. Given the study's small sample size, regions were included in the analysis only if they had >5 participants from that region. It will be examined by checking whether region makes a significant difference in participants' liking ratings.

All regions with <5 participants were excluded from this analysis, leaving Poland, Portugal and the UK. In a T-test the results were analysed factored on region, taking into account that different participants have viewed different stimulus sets. See Appendix I for the exact results. The results indicate that the region did not significantly influence the participants' ratings.

It is important to note that the data analysis per region is considered unreliable due to several limitations. There was an unequal number of participants across the regions. There were differing numbers of participants per group (stimulus set), and generally numbers per region are small. The results should be interpreted with caution.

Confounding variables

The confounding variables; demographical information, self efficacy, socio economic status, education level, noise levels, working environment, social context, colourblindness, aesthetic fluency and aesthetic attitude were examined in a linear regression. The results indicated that only noise disturbance had a statistically significant (p-value = 0,0266) influence on participants' ratings.

Qualitative results

Figure 4.2 depicts the mean like ratings assigned by participants to the stimuli plotted against the various levels of the UMA. The procedure involved testing the three levels of UMA using multiple items, which were also presented in Berghman and Hekkert's (2017) paper. Similarly to their study, the 12 individual items were converted into scores for unity-and-variety, typicality and novelty, and relatedness. The original study employed aesthetic appreciation items as anchor data to validate the UMA.

The graphs suggest that participants' like ratings align primarily with aesthetic appreciation, while the individual levels of the Unified Model of Aesthetics (UMA) show less correlation. This implies that the connection between the three UMA levels and image liking is only moderately strong compared to aesthetic appreciation.

Future research could investigate the applicability of the three levels of the UMA beyond product stimuli, which were originally studied. In this thesis, the plan was to replicate the experiment using Berghman and Hekkert's original stimulus set. Unfortunately, they could not provide it. This is recommended as a future research direction (Section 10.4.2).

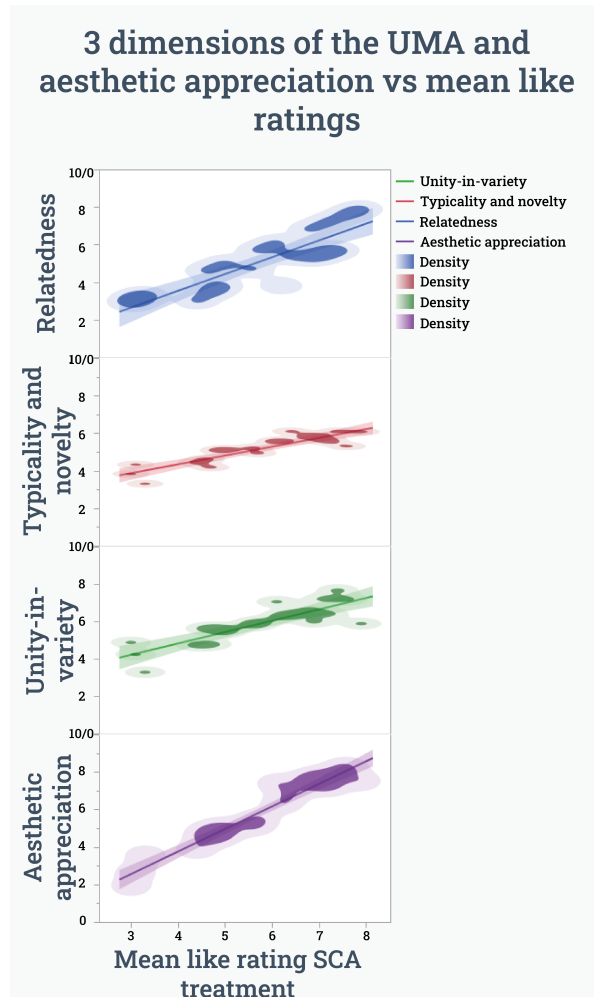


Figure 4.2: Four contour plots and their line of fit. From this figure can be concluded that aesthetic appreciation shows a strong positive correlation between its mean ratings and mean image liking ratings. Relatedness shows a less strong positive correlation between its mean ratings and mean image liking ratings. Typicality and novelty, and unity-in-variety both show a weak positive correlation with their mean ratings and mean image liking ratings. X-axis: mean image liking ratings, measured as the dependent variable. Y-axis: 3 levels of UMA and aesthetic appreciation ratings.

4.5. Conclusions

As for H1: Semantic Concept Activation, this data indicates that there is no significant difference between the ratings of image liking from participants who have been exposed to semantic concept activation with the UMA and participants who have not been exposed. Next to this, this data showed no significant difference in variance for the treatments.

When looking at the internal consistency between subjects, for this data both the control treatment and the SCA treatment show high internal consistencies.

The analysis of H2: Region shows that for this data the participants' regions do not significantly influence their ratings on image liking. These results are to be taken with a grain of salt, because of the small numbers of participants per region.

Of all the confounding variables measured in Experiment 1, only noise disturbance has a significant influence on the participants' ratings.

Qualitative results revealed that aesthetic appreciation is more aligned with image liking than the various levels of the UMA. More research is needed to see if the UMA is applicable beyond product aesthetics.

Overall, this data suggests that exposing participants to semantic concept activation does not significantly influence their liking ratings, or the variability in their answers. Both treatments showed satisfactory internal consistencies. Exposing participants to semantic concept activation is viewed as equivalently appropriate as rating images on liking. Due to the more cumbersome annotation procedure, the current situation is still considered preferred.

5

Experiment 2: rating images on aesthetic value

5.1. Introduction

Chapter 5 presents the details of Experiment 2, which aims to compare aesthetic value ratings to the liking ratings of stimuli on a Likert scale from 1-10. The motivation for this study arises from research which demonstrates that specific question phrasing can influence participants' answers (Semin and De Poot, 1997-a, 1997-b). Existing literature includes quite some studies that have employed aesthetic value to measure the aesthetics of images, underlining its contextual relevance. For this experiment, the approach taken by the Simulacra Aesthetic Captions dataset, among others, is adopted (JD-P via Github, 2022). Continuing on these insights, the H3: Aesthetic Value posits that the manner in which participants are asked a question significantly impacts their responses. By comparing the participants' ratings on image liking (the control treatment) to the ratings on the aesthetic value of images (aesthetic value treatment).

In the method section it is outlined how the participants are recruited, what stimulus set is used, and it describes the survey that uses Qualtrics and is run on Prolific. It also describes the experimental procedure, which involves two treatments: the control treatment (as described in Section 3.5) and the aesthetic value treatment which uses rating on aesthetic value of the stimuli. In the analysis section the results of the two treatments are compared to each other, and the internal consistency is examined.

The purpose of this study is for exploration rather than validation. Therefore, the number of stimuli and participants is intentionally small.

5.2. Experiment 2 - hypothesis and expected outcome

This hypothesis is based on literature described in Section 2.4.

H3: Aesthetic Value

The manner in which participants are asked a question on aesthetics significantly impacts their responses.

The null hypothesis posits that there will be no significant difference in participants' answers based on the manner in which the question is asked.

Expected outcome

The control treatment asks participants about image liking. The aesthetic value treatment asks participants on the aesthetic value of the images. Because of the literature on automatic linguistic behaviour, it is expected that the ratings of participants will significantly differ depending on the manner in which the question is asked, by automatic linguistic behaviour of participants.

5.3. Method

5.3.1. Materials

Stimulus set

For Experiment 2, stimulus set 1 was deployed, consisting of sub-stimulus sets 1.1 and 1.2. For more information, please refer to Section 3.2.5.

Survey

Experiment 2: rating images on aesthetic value

The interface of the survey of the aesthetic value treatment is shown in Figure 5.1. The control treatment interface is depicted in Figure 3.4.

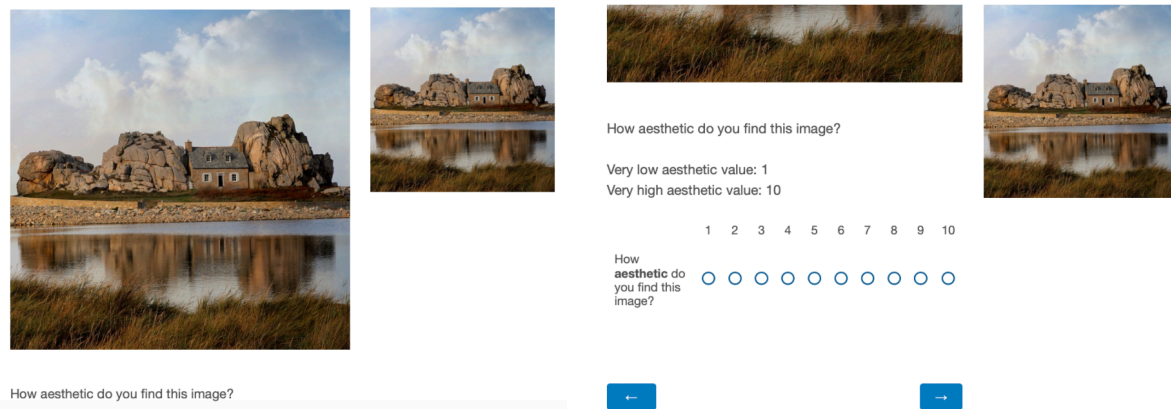


Figure 5.1: Aesthetic value treatment question. Left: the interface when the participants open the Qualtrics survey. Right: the interface when the participants scroll to the rating question.

5.3.2. Procedure

Experiment 2 includes 2 treatments, the aesthetic value treatment, and the control treatment that was already conducted in Experiment 1. The treatments were varied between subjects. The stimulus set consisted of two groups. Participants were randomly allocated to either group of the stimulus set. Before the actual treatment was administered to the participants, they were presented with two training tasks, where everything was the same as in the experiment, but different stimuli were used.

In the aesthetic value treatment the participants were asked to rate how aesthetic they find the images using a single item:

1. How aesthetic do you find this image?

A deliberate decision was made to omit 'on a scale from 1-10' from the aesthetic value treatment, in order to adopt the exact phrasing from the Simulacra Aesthetic Captions dataset (JD-P, personal communication on Discord, 2023). For both treatments, the rating was conducted on a Likert scale from 1 to 10, with the values unlabeled except for the extreme values (1 and 10) with "strongly disagree" and "strongly agree." This experiment examined if a different formulation of the rating question significantly influenced the ratings participants prescribed for the stimuli.

Below, Figure 5.1 shows the data which is collected for each participant. The bold items were collected for this experiment.

Experiment subgroups	<i>How much do you like this image on a scale from 1-10?</i>	<i>How aesthetic do you find this image?</i>
Control treatment sub-stimulus set 1.1	YES	NO
Control treatment sub-stimulus set 1.2	YES	NO
Aesthetic value treatment sub-stimulus set 1.1	NO	YES
Aesthetic value treatment sub-stimulus set 1.2	NO	YES

Table 5.1: Table with the data which was collected per participant for Experiment 2.

5.3.3. Analysis

The data was analysed using JMP software.

H3: Aesthetic Value

The manner in which participants are asked a question significantly impacts their responses.

To analyse whether there was a significant difference between the aesthetic value treatment and the control treatment per image, T-tests were conducted. The complete results can be found in Appendix J. The T-tests indicated that the p-value was not significant for any stimulus (all p-values > 0,0705), for this data.

Internal consistency between subjects

To look into the inter-rater reliability the Cronbach's alpha was calculated. As can be seen in table 5.2, The recognised guideline where an alpha value > 0.7 is acceptable was adhered here. When the alpha values of the control treatment were compared to the aesthetic value treatment, the reliability was not systematically increased in the aesthetic value treatment.

Aesthetic value treatment	Sub-stimulus set 1.1	$\alpha = 0,8428$
Aesthetic value treatment	Sub-stimulus set 1.2	$\alpha = 0,9252$

Table 5.2: Experiment 2 Cronbach's alpha results.

Qualitative results

When examining images with substantial differences in mean ratings between the two treatments, it is evident that the left image shown in Figure 5.2 receives notably lower ratings for image liking compared to aesthetic value. Contrarily, the right image in Figure 5.2 receives considerably higher ratings for image liking than aesthetic value. When considering the content of these images, it suggests that the subject depicted in the image may be influencing the ratings. For instance, an image featuring a woman beside a building with bullet holes in what appears to be a war zone has an average liking score 1.2333333 lower than its aesthetic value rating. Conversely, an image depicting what resembles a Disney palace with a wedding setup in front of it receives higher average liking ratings than aesthetic value ratings, with a difference of 1.355556. Of course, these are just two stimuli where this is found, so further research needs to be performed to see if this might be happening systematically. These findings contribute to the rationale for conducting the experiment described in Chapter 8, where it is examined whether there is a significant difference between image liking ratings of participants who were instructed to evaluate the content of the image versus participants who were instructed to evaluate the overall image on liking. Next to this, a future research recommendation is proposed to investigate the impact of morally negative/positive topics in stimuli on participants' ability to assess the overall aesthetics of the images (Section 10.4.3).



Figure 5.2: Left: image assigned a considerably lower average liking than aesthetic value rating. Right: image assigned a considerably lower aesthetic value than average liking rating.

5.4. Conclusions

As for H3: Aesthetic Value, this data showed that there was no significant difference in participants' responses between the two treatments for any of the stimuli.

For this data, the control treatment and the aesthetic value treatment both show high internal consistencies.

The qualitative findings suggest that the content depicted in an image might possibly influence participants' ratings. Further investigation of this aspect is warranted, as discussed in Section 8 and 10.4.3.

Overall, there is no significant difference between asking participants to rate images on aesthetic value and image liking for this data. Rating images on aesthetic value is viewed as equivalently appropriate as rating images on liking.

6

Experiment 3: ranking images on aesthetic value

6.1. Introduction

Chapter 6 describes Experiment 3, conducted to compare ranking of images on aesthetics with ratings of image liking on a scale from 1-10.

The motivation for this study arises from existing literature on measuring aesthetic preference through ranking. Previous studies have highlighted possible advantages of ranking for providing quantitative outputs as well as its relevance in training generative models. Continuing on these insights, the H4: Ranking posits significant differences between participants' rankings of image aesthetics and their subjective ratings of image liking.

In the method section, it is outlined how the participants are recruited, what stimulus set is used, and it describes the survey deployed through Qualtrics on Prolific. It also describes the experimental procedure, which involves two treatments: the control treatment and the ranking treatment which uses ranking of the stimuli. In the analysis section, the results of the two treatments are compared to each other, and the internal consistency is examined both on group and participant level.

This experiment aims to explore if the data collected indicates that there might be a significant difference between the treatments. Therefore, the number of participants and stimuli is limited.

6.2. Experiment 3 - hypothesis and expected outcome

This hypothesis is based on literature described in Section 2.5.

H4: Ranking

Participants' rankings of image aesthetics will show significant differences when compared to their subjective ratings of image liking on a scale of 1-10.

The null hypothesis posits that there will be no significant difference in rankings between participants' assessments of image aesthetics and their subjective ratings of image liking on a scale of 1-10.

Expected Outcome

I expect that the participants' rankings of the stimuli will be significantly different to the relative ranking of the participants' ratings on image liking, because the ranking modality obliges participants to assign an image for every value between 1-10.

6.3. Method

6.3.1. Materials

Stimulus set

Here, stimulus set 2 is implemented, which consists of sub-stimulus sets 2.1 and 2.2. For further description, please refer to Section 3.2.6.

Survey

Experiment 3: ranking images on aesthetic value

Figure 6.1 shows the interface of the survey. The ranking interaction has been performed by drag and drop. Due to the modality of the question, the stimuli in the ranking treatment were smaller than in the control treatment (for this interface, please visit Figure 3.4).

Can you rank the following images from low aesthetic value (LEFT) to high aesthetic value (RIGHT)?

Very low aesthetic value: LEFT

Very high aesthetic value: RIGHT

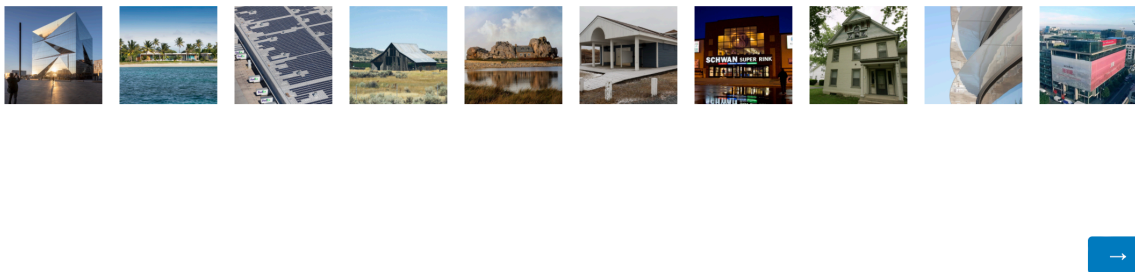


Figure 6.1: Ranking treatment question.

6.3.2. Procedure

Experiment 3 aimed to compare the control treatment used as described in Section 3.5 with a different modality: ranking, which will be referred to as ranking treatment. In the ranking treatment, the participants have ranked 10 stimuli based on their aesthetic value. Experiment 3 used the stimulus set as described in Section 3.2. As stated, the stimulus set consisted of two groups: 2.1 and 2.2. The participants have been randomly allocated to one of the two groups within the ranking treatment.

In the ranking treatment, participants were first presented with a training question. The setup in the training question was exactly the same as in the actual experiment question, but with different stimuli. In the actual experiment question, the participants have been tasked with ranking 10 stimuli based on their aesthetic value, from low to high. The stimuli have been presented to the participants in a randomised order. In the analysis, this question were compared to the control treatment.

After this, the participants were presented with a distraction task to divert their attention from the actual experiment question. This distraction task has involved ranking stimuli from the other stimulus set based on their aesthetic value, from low to high.

After the distraction task, the participants have been presented with the same experiment question again, with the same stimuli as before, to assess the internal consistency of the participants' ranking. The stimuli have been presented again in randomised order.

By comparing the control treatment and the ranking treatment, Experiment 3 aimed to investigate possible differences in participants' aesthetic evaluations and internal consistency across the two conditions.

6.3.3. Analysis

The analysis was conducted using JMP software.

H4: Ranking

Participants' rankings of image aesthetics will show significant differences when compared to their subjective ratings of image liking on a scale of 1-10.

Experiment 3 involved a comparison between two modalities, ranking and rating. To compare the two, one of the modalities needed to be converted to the other.

To compare ranking data with rating data, the mean score per stimulus per treatment was calculated. In the ranking treatment, the participant put each image in a position between 1-10. This was seen here as the rating per participant per stimulus. For the mean score of the stimuli for the ranking treatment, all these scores were averaged for all participants per stimulus. A linear regression analysis was performed with these scores. The results showed a weak positive correlation (slope = 0.13213270), but it was not statistically significant (standard error = 0,147997, p-value = 0.3837). The R-squared value is deemed unacceptable (R-squared = 0,042406), for Ozili (2023) states that in social sciences an R-squared value between 0 - 0.09 is deemed inadequate, between 0.10 - 0.50 is considered acceptable if some or most variables are statistically significant, and 0.51 - 0.99 is considered acceptable, particularly if most variables are statistically significant.

Next to this a comparison between ranking data and rating data was conducted using the relative rank per stimulus, calculated with the mean scores per treatment. The linear regression revealed a weak positive (slope = 0.2666667), which was not statistically significant (standard error = 0,227167, p-value = 0.2557). The R-squared was not acceptable (R-squared = 0,071111).

In Appendix K an overview of both treatments and how the corresponding stimulus scores are included.

Internal consistency between subjects

The Cronbach's alpha was employed to analyse the inter-rater reliability, of which the results were depicted in table 6.1. The commonly accepted rule of thumb which states that an alpha value > 0.7 is acceptable. The alpha values of the control treatment were high. In contrast, the poor inter-rater reliability of the ranking treatment was striking. The participants that were exposed to sub-stimulus set 2.2 even showed a negative Cronbach's alpha value, which is noteworthy. Having reviewed the data and consulted several sources (Hanafiah, 2015; Shehata, 2018), it seems that this may be due to the small sample size. The ranking treatment did not meet the threshold set beforehand, implicating that the alpha value was not acceptable.

Ranking treatment	Sub-stimulus set 2.1	$\alpha = 0,4124$
Ranking treatment	Sub-stimulus set 2.2	$\alpha = -0,8153$

Table 6.1: Experiment 3 Cronbach's alpha results.

Internal consistency within subjects

Correlations were examined between the regular ranking question and its corresponding internal consistency twin question for each participant. The complete individual internal consistency results can be found in Appendix K.

- In group 2.1, a significant positive correlation between both ranking twin questions was observed in only 1 participant.
- In group 2.2, a significant correlation between both ranking twin questions was observed in 3 participants. On top of this, 2 participants provided identical answers for both ranking questions.

The internal consistency among participants is very low in group 2.1 and also low in group 2.2. Overall, only 30 percent of the participants (6 out of 19) demonstrated a high level of internal consistency.

Qualitative interpretations

One possible explanation for the very low internal consistency within subject could be attributed to the stimulus presentation method. The stimuli were randomly presented to participants using a drag and drop interface, where they had to rearrange ten stimuli in a randomised row. It could be that participants clicked through the stimuli without bothering to reposition them. Another explanation could be that because participants were presented with 10 stimuli at the same time, this was too much for them to evaluate the stimuli on aesthetics. Palmer, Schoss and Sammartino (2013) also caution against a ranking modality for aesthetic preference studies. Possibly, this experiment would show improved internal consistencies within subject if participants had to evaluate fewer stimuli simultaneously. The stimuli of the drag-and-drop modality were displayed in a row. Participants were expected to drag and drop them into new places within this row, to give a preferred ranking. It may be that because the stimuli were already presented in a randomised order, it was more difficult for the participants to envision a new ranking. Unfortunately, I did not measure which randomised order each participant received during the experiment, so I cannot test this post hoc (Section 10.4.4). This is also noted as a limitation of this research (Section 10.3).

6.4. Conclusions

As for H4: Ranking, the analysis showed for this data no statistically significant correlation for the mean scores of the stimuli, as well as the relative ranking of the stimuli between the treatments. This means that the null hypothesis is rejected, and the two treatments are significantly different from each other.

The internal consistency on group level for this data showed an unacceptably low Cronbach's alpha for the ranking treatment. From this it can be concluded that for this data the responses within the group does not show a strong agreement on the ranking of the stimuli.

For the control treatment the internal consistencies on group level for this data are quite high, which shows that for this data the responses within the group show a strong agreement on the rating of the stimuli.

When looking at the internal consistency of the ranking treatment on the participant level for this data only a small number of participants demonstrated a significant correlation between their responses, indicating a low internal consistency per participant. For the control treatment, the internal consistency on participant level shows a statistically significant strong positive correlation between the ratings for the stimulus and its internal consistency twin stimulus.

Retrospectively, in consideration of the qualitative interpretations, the modality applied in this experiment may have been designed unfortunately. Future research where fewer stimuli are presented to the participants at a time, and where the stimuli are not already presented on a line which may make it more difficult for the participants to imagine the stimuli in a new order are therefore advised in Section 10.4.4.

Ranking images on aesthetic value does not seem to be a more appropriate alternative than the current situation due to the very low internal consistency of the participants both at group and individual level.

7

Experiment 4: image preference with two alternative forced choice

7.1. Introduction

Chapter 7 describes Experiment 4, conducted to compare two alternative forced choice (2AFC), a modality where participants indicate their preference between an image duo with ratings of image liking on a scale from 1-10.

The motivation for this study arises from existing literature on measuring aesthetic preference through 2AFC. Previous studies have highlighted possible advantages of 2AFC in the context of measuring aesthetics of visual stimuli as well as its relevance in training generative models. Continuing on these insights, H5: 2AFC posits significant differences between participants' preferences indicated with 2AFC and their subjective ratings of image liking.

In the method section, it is outlined how the participants are recruited, what stimulus set is used, and it describes the survey deployed through Qualtrics on Prolific. It also describes the experimental procedure, which involves two treatments: the control treatment and the 2AFC treatment. In the analysis section, the results of the two treatments are compared to each other, and the internal consistency is examined both on group and participant level.

The purpose of this study is for exploration rather than validation. Therefore, the amount of stimuli and participants is intentionally small.

7.2. Experiment 4 - hypothesis and expected outcome

This hypothesis is based on literature described in Section 2.6.

H5: 2AFC

The alternative forced choice annotations of image aesthetics will result in significantly different outcome scores compared to participants' subjective ratings of image liking on a scale of 1-10.

The null hypothesis posits that there will be no significant difference in outcome scores between the alternative forced choice annotations of image aesthetics and participants' subjective ratings of image liking on a scale of 1-10.

Expected outcome

I expect that the outcomes of the 2AFC experiment will significantly differ from the outcomes of ratings of image liking, because the modality of the 2AFC experiment obliges participants to decide between two images. This way, I expect the relative ranking to be different from ratings on image liking.

7.3. Method

7.3.1. Materials

Stimulus set

For Experiment 4, stimulus set 2 was used, which consists of sub-stimulus sets 2.1 and 2.2. For more information, see Section 3.2.6.

Survey

Experiment 4: image preference with two alternative forced choice

Figure 7.1 shows the interface of the survey. The participants have selected either the "LEFT" button or the "RIGHT" button. When the button was clicked, the survey automatically proceeded to the next duo. Participants have had the option to navigate back and forth between different duos using the arrows below. For the interface of the control treatment I would like to refer you to Figure 3.4.

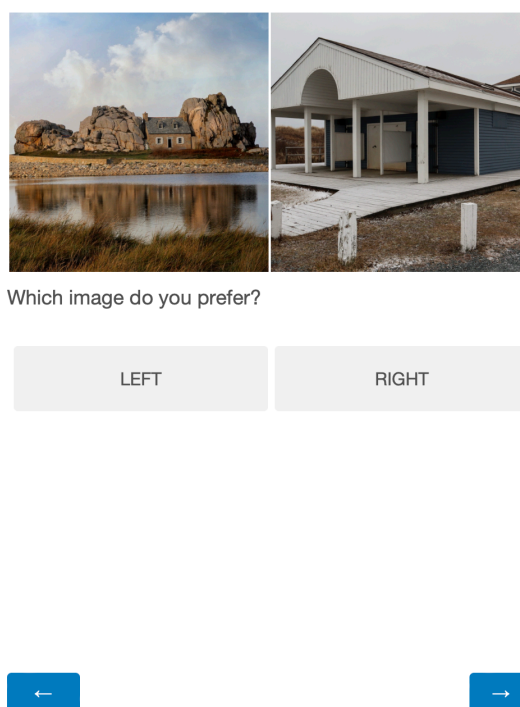


Figure 7.1: 2AFC treatment question.

7.3.2. Procedure

In the 2AFC treatment, participants have first been presented with two training questions. The setup in the training questions is exactly the same as in the actual experiment questions, but with different stimuli. In the actual experiment question, the 2AFC stimuli duos have been presented to the participants in a randomised order.

Experiment 4 aimed to compare the control treatment with a different modality: 2AFC, which is referred to as the 2AFC treatment. In the 2AFC treatment, the preference between two stimuli was indicated by the participants for all of the 10 stimuli per sub-stimulus set (2.1 and 2.2, respectively). The participants were randomly allocated to one of the two groups within the 2AFC treatment. This meant that each participant was presented with 45 unique combinations of the images from the stimulus set. Each image was shown a total of 9 times. As this experiment required participants to answer more questions than the previous experiments, it was decided to incorporate two attention checks. Participants were presented with simple maths, as also deployed by Lomas et al. (2023). Before the participants started the survey, it was communicated that the survey would test whether they were paying attention or not. An example of one of the two attention checks is depicted in Figure 7.2.

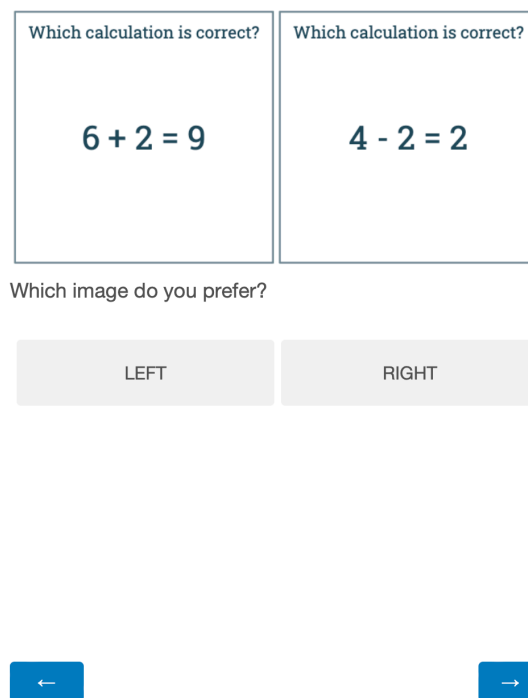


Figure 7.2: Experiment 4 example of an attention check.

To check the internal consistency within subjects, one stimulus set duo has been presented twice to each participant. This internal consistency question has been randomised in the loop order with the 45 treatment question duos.

By comparing the control treatment and the 2AFC treatment, Experiment 4 aimed to investigate possible differences in participants' aesthetic evaluations and internal consistency across the two conditions.

7.3.3. Analysis

JMP software was used to analyse the data of Experiment 4.

H5: 2AFC

The alternative forced choice annotations of image aesthetics will result in significantly different outcome scores compared to participants' subjective ratings of image liking on a scale of 1-10.

Experiment 4 involved a comparison between two modalities, 2AFC and rating. To compare 2AFC data with rating data, the total amount of times a stimulus was clicked as preferred per participant was re-scaled to 1-10, which was also the case in control treatment. To analyse if there is a statistically significant difference per image per treatment, t-tests were conducted. From this, it could be concluded that 30 percent of the images showed a statistically significant difference between the treatments. Of the six stimuli where a significant difference was found between the treatments, for four of the stimuli the 2AFC treatment had a higher mean. Please refer to Appendix M for the results in detail of this analysis.

Next to this, both treatments were also compared on mean score per image of all participants with a t-test. This showed no significant difference (p-value = 0,7457) between both treatments.

Internal consistency between subjects

To look at the inter-rater reliability, the Cronbach's alpha was examined, which is depicted in table 4.2. The widely accepted rule of thumb of interpreting alpha values was adhered to, which states an

acceptable reliability = >0.7 . When comparing the alpha values of the control treatment to the 2AFC treatment, the reliability was not systematically increased for the 2AFC treatment 7.1.

2AFC treatment	Sub-stimulus set 2.1	$\alpha = 0,8218$
2AFC treatment	Sub-stimulus set 2.2	$\alpha = 0,9184$

Table 7.1: Experiment 4 Cronbach's alpha results.

Internal consistency within subjects

To measure internal consistency, each participant was exposed to one image duo twice in the randomised loop order of the experiment. 95 percent of participants passed the internal consistency check. Please refer to Appendix M for the results in detail of this analysis.

Qualitative results

As discussed earlier in this section, there is a significant difference between the treatments for 30 percent of the images. These stimuli were examined, but no common denominator can be concluded from this. (for the full table, please refer to Appendix M.1).

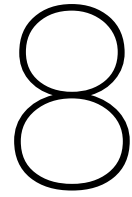
7.4. Conclusions

As for H5: 2AFC, the mean scores of all participants per image are compared between the treatments, and no significant difference is found. This leads to the conclusion that overall for this data the ratings of image liking do not significantly differ from indicating preference per image. When zooming in on single images, for a small number of images, a significant difference was found between the treatments. No common denominator was found between these images.

For the 2AFC treatment and the control treatment the internal consistencies on group level for this data are both quite high, which shows that for this data the responses within the group show a strong agreement on the rating of the stimuli.

Both treatments show high within subject internal consistency.

Overall, there is no significant difference between asking participants to indicate their preference between images and rating images on their liking for this data. 2AFC is viewed as equivalently appropriate as rating images on liking.



Experiment 5: content vs overall image liking

8.1. Introduction

This chapter describes an experiment run to examine H6: Content VS Overall Image. This is based on the rationale described in Section 2.7. As described there, high internal consistency between subjects is desirable for an annotation method in this context because it suggests that it can be used to annotate training datasets that are widely deployed. Although aesthetic experiences are highly subjective, previous experiments in this thesis and studies by Berghman and Hekkert (2017) and Blijlevens et al. (2017) have shown high internal consistency between subjects. Incidentally, these studies have product stimuli similar to the image class used here: buildings. These image classes all have a functional aspect. According to Kant's classification, free beauty is derived from the contemplation of the stimulus itself, while dependent beauty refers to how well the stimulus fulfils its intended purpose. It is possible that functional stimuli elicit primarily functional contemplation, valuing primarily the capabilities of the stimulus, rather than aesthetic features in aesthetic contemplation.

To investigate this, this chapter compares two groups of participants: one group is instructed before annotating to annotate the images on content liking. The second group is instructed to annotate on overall image liking. After this, they both go through the annotation process as was carried out in the control treatment.

8.2. Experiment 5 - hypothesis and expected outcome

To see whether the affordances of the buildings interfere with the liking judgements of the participants, the following hypothesis was formulated:

H6: Content VS Overall Image

When participants rate their liking of the image content, there will be significant difference with when they indicate the image liking of the overall image.

The null hypothesis posits that there will be no significant difference in participants' ratings of image liking between the evaluation of image content and the evaluation of the overall image.

Expected outcome

I anticipate a significant difference in liking ratings due to interference on the liking ratings of functional contemplation. This would cause a difference when participants indicate their overall image liking rating as when they give only what is depicted, the building, an image liking rating.

8.3. Method

8.3.1. Materials

Stimulus set

For this study, stimulus set 2 was used, which consists of sub-stimulus set 2.1 and sub-stimulus set 2.2, which is explained in more detail in Section 3.2.

Survey

The participants were initially presented with the instruction depicted in Figure 8.1 for 30 seconds. After this, they could proceed with the training and regular tasks, similar to the interface of the annotation method from the control treatment, as described in Section 3.5.

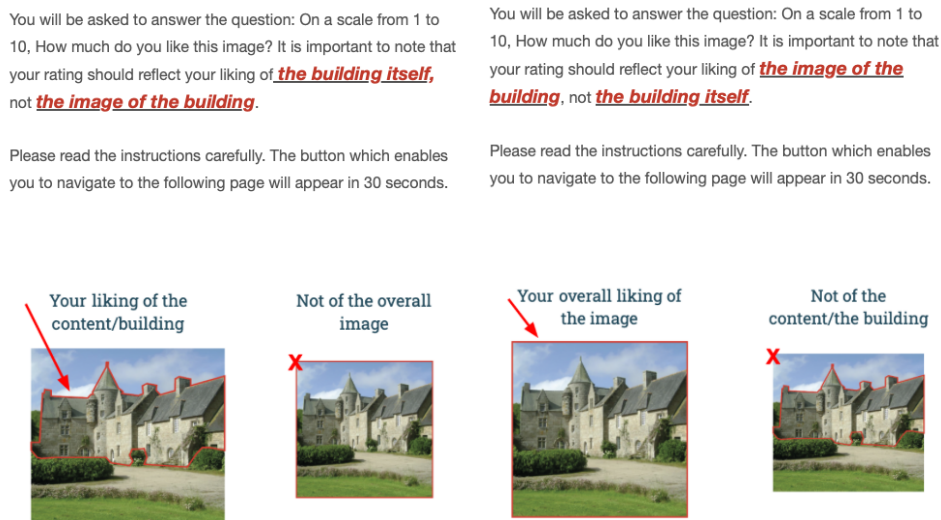


Figure 8.1: Left: instruction for the content treatment. Right: instruction for the overall image treatment.

8.3.2. Procedure

This experiment compared two treatments: the content treatment and the overall image treatment. In the content treatment, participants were instructed to rate specifically what was depicted on liking. In the overall image treatment, participants were instructed to rate the entire image on liking. Participants were exposed to this instruction for 30 seconds before they could move on to the training and experiment questions. This instruction was the only thing that was varied in this study. After the instruction, participants went through the same procedure as in the control treatment, also discussed in Section 3.5.

8.3.3. Analysis

The data was analysed using JMP software.

H6: Content VS Overall Image

When participants rate their liking of the image content, there will be significant difference with when they indicate the image liking of the overall image.

To analyse if there was a significant difference between the content treatment and the overall image treatment, T-tests were conducted for each stimulus. All results are enclosed in Appendix O. Based on the T-tests conducted, for this data a significant difference was found for only one stimulus (p -value = 0.0338).

Internal consistency between subjects

The Cronbach's alpha was examined to look into the inter-rater reliability, as depicted in table 8.1. The established guideline under which an alpha value > 0.7 is considered acceptable was adopted here.

When the alpha values of both treatments are compared, both the content treatment and the overall image treatment show satisfactory levels of inter-rater reliability.

Group 1	content treatment	sub-stimulus set 2.1	$\alpha = 0,8314$
Group 2	content treatment	sub-stimulus set 2.2	$\alpha = 0,8882$
Group 3	overall image treatment	sub-stimulus set 2.1	$\alpha = 0,8241$
Group 4	overall image treatment	sub-stimulus set 2.2	$\alpha = 0,8910$

Table 8.1: Content vs overall image Cronbach's alpha results.

Internal consistency within subjects

To analyse the internal consistency within subjects, a linear regression is used. Here, it is looked at if image 10 from sub-stimulus set 2.1 correlates significantly with its internal consistency twin image, and the same for image 20 from sub-stimulus set 2.2. The analysis showed that for this data, all groups show a significant correlation between the regular experimental image and its internal consistency twin image. The results can be found in table 8.2.

group	treatment	p-value	slope
group 1	content treatment	p-value = 0,0011	slope = 0,9288256
group 2	content treatment	p-value = 0,0005	slope = 0,7884615
group 3	overall image treatment	p-value = <0,0001	slope = 0,9430255
group 4	overall image treatment	p-value = 0,0129	slope = 0,8474576

Table 8.2: Linear regressions for the internal consistency within subjects

8.4. Conclusions

When considering the H6: Content VS Overall Image, it appears that for this data, there is no significant difference between participants instructed to rate the content of images on liking, and participants instructed to rate the liking of the overall image on liking.

For this data, both treatments show satisfactory levels of internal consistency both between subject and within subject level.

This suggests that the functional contemplation of stimuli with a practical aspect might not interfere with liking ratings on overall image level. More research is needed, using other stimulus sets, to give a conclusive answer to this.

9

Post hoc analysis: comparison between participant ratings and predictor scores

9.1. Introduction

This chapter focuses on evaluating the performance of the LAION Aesthetics Predictor in comparison to human ratings. Specifically, the goal is to measure the differences between the predicted aesthetic scores assigned by the LAION Aesthetics Predictor and the average image liking scores assigned by participants in the control treatment.

Examining the content of the LAION Aesthetics V2 subsets, an interesting observation arises. The images with the highest "aesthetic" scores consist mainly of landscapes and portraits of women, resembling generated images (Baio, 2022). This is noteworthy because well-established aesthetic principles highlight the significance of diversity in aesthetic experiences (Berghman and Hekkert, 2017). In contrast with this, the higher-scoring image segments of the LAION Aesthetics V2 datasets exhibit a noticeable decrease in diversity as the scores increase. This observation suggests that the current datasets may deviate from the literature on aesthetics.

To explore the alignment between the LAION Aesthetics Predictor scores and human ratings, a post hoc analysis is conducted by comparing participants' ratings of image liking to the scores assigned by the predictor for the stimulus set, which is introduced in Section 3.2.6. Further details and analysis of this hypothesis can be found in this chapter.

9.2. Post hoc analysis - hypothesis and expected outcome

This hypothesis is based on the observation that the highest-scoring stimuli of the LAION Aesthetics Predictor show a high degree of homogeneity, which raises questions in regard to aesthetic theory.

H7: Predictor Comparison

The predicted aesthetic scores assigned by the LAION Aesthetic Predictor are significantly different from the average image liking scores assigned by participants.

The null hypothesis posits that the predicted aesthetic scores assigned by the LAION Aesthetic Predictor are not significantly different from the average image liking scores assigned by participants.

Expected outcome

I expect that the predicted aesthetic scores will exhibit significant differences with human ratings because of the lack of diversity in the higher scoring segments of the current LAION Aesthetics V2 subsets, as described in Section 1.1.

9.3. Analysis

Here, a comparison is made between the average scores on image liking that participants assigned to the stimuli in the control treatment and the predicted aesthetic scores by the LAION Aesthetic Predictor. This is done in an OLS regression in which the average image score is the dependent variable and the LAION predictor score is the independent variable. If the LAION predictor does a poor job, the coefficient would be indistinguishable from 0. On the other hand, if the LAION predictor is good, then the coefficient would be indistinguishable from or close to 1 while at the same time the coefficient of the constant would be indistinguishable from 0. Here, the R-squared (R^2) is examined because it provides a measure of how well the regression model fits the observed data. In the social sciences, the interpretation of R-squared can be approached as follows: an R-squared value between 0 and 0.09 is considered inadequate, an R-squared value between 0.10 and 0.50 is considered acceptable if some or most variables are statistically significant, an R-squared value between 0.51 and 0.99 is considered acceptable, especially if most variables are statistically significant (Ozili, 2023).

There is a significant positive coefficient (1.7927925), (standard error = 0,750901, p-value = 0.0281) for the LAION predictor score. Here, the R-squared value is deemed acceptable (R-squared = 0,240515).

Observed:

*AVG img like score = -3,966483 + 1,7927925*LAION Aesthetic predictor score*

Ideal case (As stated by Piñeiro et al. (2008)):

*obs=0.00+1.00*Pred*

The coefficient of 1.7927925 is quite far from 1, but it is not significantly different from 1. The analysis shows an intercept of -3.966483 which is not significantly different from 0 (p-value = 0.3165). This means that the null hypothesis is not rejected; the LAION predictor aligns to some extent with the participants ratings, although a caveat is that the coefficients are not precisely estimated. See Appendix L for the full analysis and overview of all predicted scores per image.

9.3.1. Qualitative results

In Figure 9.1, the ratings of image liking assigned by humans are plotted with the aesthetic scores assigned by the LAION Aesthetics predictor, per image. Although the null hypothesis cannot be rebutted based on the analysis, it is interesting that there are some images where there seems to be a considerable difference between human ratings and predictor scores. When examining the specific contents of images where ratings and scores differ more (for an overview of this, see Figure L.3), no clear connection can be found between the images. Figure 9.1 shows that the predictor does not assign outlier scores, where humans do rate images with more extreme ratings. Therefore, the reason that for some images there is more agreement between humans and the predictor does not seem to be related to the images themselves, but only to the fact that the predictor for this stimulus set only assigned scores between 4,447845936 and 6,149407387.

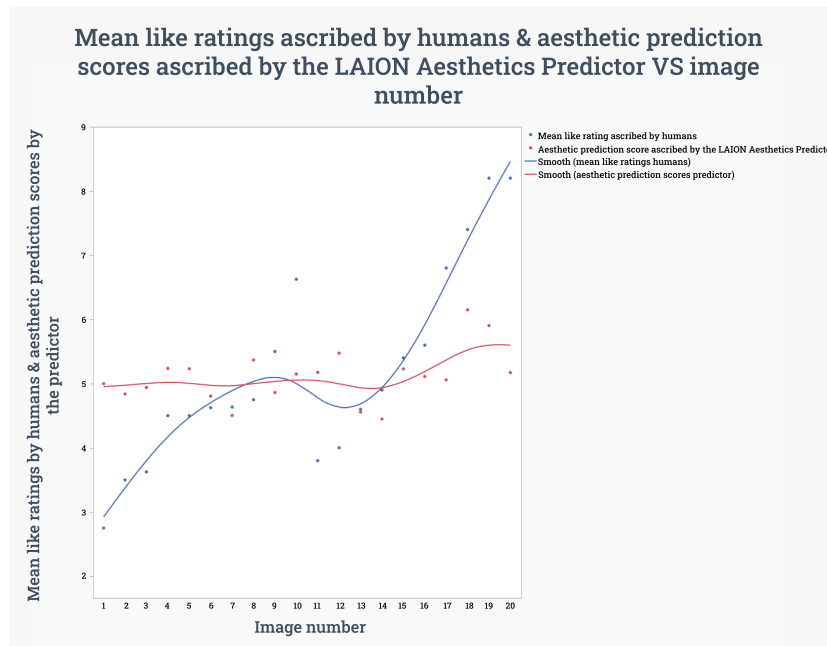


Figure 9.1: Human mean like ratings plotted against predictor scores. Each dot represents an image, the blue line shows a smooth curve through the human ratings, the red line through the predictor scores. Note that the rows were sorted based on relative rank for better graph clarity purposes, the numbers on the x-axis are unrelated to the numbers in the original stimulus set. X-axis: image number. Y-axis: average like ratings by humans and LAION Aesthetics Predictor score.

The LAION researchers have supplied a URL ([click here](#)) where you can take a look at samples from the LAION dataset, distributed into aesthetic buckets using the LAION-Aesthetics Predictor V2 (Schuhmann, n.d.). I have plotted a bell curve, as shown in Figure 9.2, where the distribution of the images of the buckets is depicted. From this it can be concluded that the images are not well distributed among the buckets. The predictor scores a disproportionate number of images between 4 and 5.5. This supports the finding from the analysis presented above that the LAION Aesthetics Predictor assigns images predominantly moderate scores.

Distribution of images of the LAION dataset over the aesthetic buckets by the LAION Aesthetics Predictor

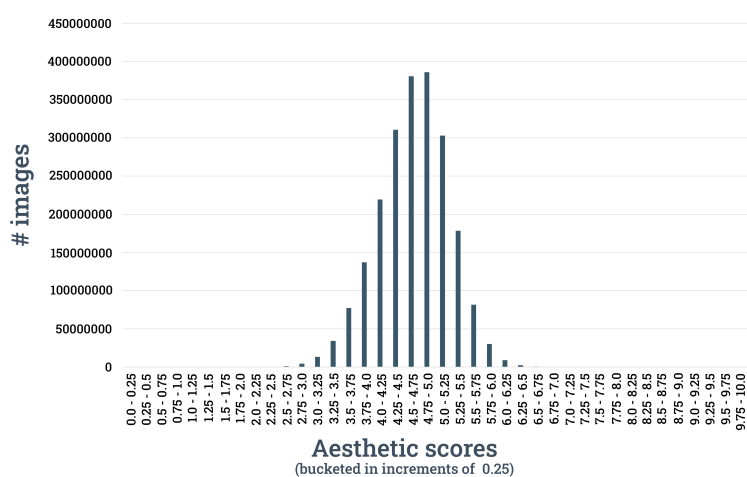


Figure 9.2: The distribution of images of the LAION dataset over the aesthetic buckets by the LAION-Aesthetics V2 predictor. X-axis: aesthetics score buckets. Y-axis: amount of images.

9.4. Conclusions

Taking the quantitative and qualitative results together, these findings suggest that while there is some agreement between the LAION Aesthetics Predictor and participants' ratings, there are also noticeable differences. The quantitative analysis indicates a positive relationship but is by no means a perfect match, and the qualitative analysis shows that there are moments of disagreement between the predictor and human ratings, when people assign more extreme ratings but the predictor still assigns moderate scores.

In conclusion, the findings suggest a partial alignment between the two, but also highlight the need for further research and room to improve the accuracy of the predictor for extreme ratings and take into account a wider range of aesthetic preferences.

10

Discussion

10.1. Summary of the findings

In this thesis, I examined whether there is a significant difference between participants' ratings on image liking and some promising alternatives which are backed by literature. I conducted four experiments where alternative treatments were compared to the control treatment. Additionally, one experiment aimed to gain more insight into annotating for aesthetics in the context of this thesis, and two post hoc analyses were deployed with the same purpose.

- Experiment 1 examined an alternative task whether participants exposed to semantic concept activation with the Unified Model of Aesthetics rated image liking differently from participants not exposed to it. Additionally, a post hoc analysis was conducted on the data of this experiment to examine if participants from different regions yielded different results (Chapter 4).
- Experiment 2 looked at whether an alternative metric, namely a different phrasing of the question *"how aesthetic do you find this image?"*, yielded different ratings compared to the question *"how much do you like this image on a scale from 1-10?"* (Chapter 5).
- Experiment 3 investigated whether using a different modality, specifically ranking, as an alternative metric, yields distinct results compared to ratings on image liking (Chapter 6).
- Experiment 4 delved into the question whether another modality as an alternative metric, 2AFC, provides different results compared to ratings on image liking (Chapter 7).
- Experiment 5 was conducted to gain further insights into annotating for aesthetics in this context. Specifically, it aimed to compare the results of participants who were instructed to indicate their image liking of the content with those who were asked to rate the stimulus based on their overall image liking (Chapter 8).
- A post hoc analysis was deployed to compare the mean image like ratings humans ascribed to the stimulus set to aesthetic predictions by the LAION Aesthetics Predictor (Chapter 9).

All of these research endeavours aimed to address the overarching research question: how does the annotation method used in the development of the LAION-Aesthetics V2 datasets compare to other annotation methods for measuring aesthetics, considering different approaches and annotating aesthetics in this context? On the basis of the literature, some promising alternatives were formulated. Surprisingly, none of the alternative methods performed significantly better than the LAION Aesthetics approach. In fact, one of the alternative approaches, the ranking method, even performed significantly worse. Region does not seem to influence image liking ratings for this data. Furthermore, there appears to be no difference between participants who indicate overall image liking compared to content liking. There is some level of agreement between human liking ratings and aesthetics prediction scores by the LAION Aesthetics predictor for this data, but it is by no means perfectly aligned. Where humans assign more extreme ratings, the predictor tends to keep assigning moderate scores.

10.1.1. Summary of the results

Control treatment

The internal consistency in the control treatment (*"how much do you like this image on a scale from 1-10?"*) shows a strong positive significant correlation for both groups of the stimulus set. The control treatment also shows high internal consistency at group level. Ex post, the stimulus set used may not have given the alternative approaches a fair chance to show their ability to improve matters. Moreover, considering the perspective that buildings are designed with the intention that everyone should find them somewhat pleasing to look at, this could have impacted the assessment of diverse aesthetic perspectives. A future recommendation is formulated to look into how the treatments behave with different image classes (Section 10.4.3).

H1: Semantic Concept Activation: exposing participants to semantic concept activation through questions about typicality and novelty, unity and variety, and relatedness, will influence their ratings of how much they like images.

Experiment 1 (Chapter 4) examined the impact of semantic concept activation using the Unified Model of Aesthetics (UMA) on participants' image liking ratings. The results indicate that there is no significant difference in image liking ratings between participants exposed to semantic concept activation with the UMA and those who were not. Both treatments showed satisfactory internal consistencies, suggesting that the hypothesis was not supported. Qualitative results suggest that aesthetic appreciation metrics may be more aligned with image liking than the different levels of the UMA. Further research is needed to explore the applicability of the UMA beyond product aesthetics. It should be noted that both the UMA and semantic concept activation have been validated primarily for product aesthetics and may not be generalisable to other image classes. Recommendations for future research on this topic can be found in Section 10.4.1 and 10.4.2.

H2: Region: participants from different regions experience different images as aesthetic.

The results of Experiment 1 do not show a significant influence of the region on participants' ratings, which aligns with the null hypothesis. However, caution should be taken when interpreting these results due to unequal participant distribution across regions and varying group sizes (division between sub-stimulus sets). These limitations are discussed further in Section 10.3, along with a recommendation for future research regarding this issue in Section 10.4.6.

H3: Aesthetic Value: the manner in which participants are asked a question on aesthetics significantly impacts their responses.

Experiment 2 (Chapter 5) compared participant ratings based on aesthetic value and image liking. The results indicate that there is no significant difference between the two treatments. Both treatments exhibited satisfactory internal consistencies. H3: Aesthetic Value was not supported. Qualitative results suggested that the depicted content potentially influences how images are rated in terms of liking. This observation motivated the formulation of H6: Content VS Overall Image. However, further research is needed to draw reliable conclusions, as discussed in the future recommendations in Section 10.4.3. One possible explanation for not being able to rebut the null-hypothesis is the limited variation between the question phrasings. It is possible that participants interpreted both questions similarly, resulting in a lack of differentiation between ratings based on aesthetic value and image liking.

H4: Ranking: participants' rankings of image aesthetics will show significant differences when compared to their subjective ratings of image liking on a scale of 1-10.

Experiment 3 (Chapter 6) compared image ranking on aesthetic value to image liking ratings, revealing significant differences between the two. The ranking treatment showed low inter-rater reliability at the group level, with only a small number of participants showing significant correlation in their responses. In contrast, the control treatment demonstrated strong within-participant agreement. Therefore, the hypothesis was not supported. The drag-and-drop ranking modality, where stimuli were presented in a randomised order, may have contributed to participants not thoroughly analysing each stimulus or considering alternative arrangements. This limitation should be addressed in future research (see Section 10.4.4). Additionally, ranking poses the challenge of participants being exposed to a large number of stimuli simultaneously, which is considered a drawback according to Palmer, Schoss and Sammartino (2013).

H5: 2AFC: the alternative forced choice annotations of image aesthetics will result in significantly different outcome scores compared to participants' subjective ratings of image liking on a scale of 1-10.

Experiment 4 (Chapter 7), compared preference using 2AFC with image liking ratings. The mean scores per image across all participants did not show a significant difference between the treatments. Both treatments demonstrated high internal consistencies at both the group and individual levels. The hypothesis was not confirmed. One potential explanation for this outcome is that 2AFC may limit participants' ability to express nuanced preferences. The binary choice format may overlook subtle distinctions in their preferences. A recommendation for future research (Section 10.4.5) is to explore n-AFC modalities (e.g., 4 or 6) where participants can make more deliberate choices based on their preferences. Another suggested direction for future research is Gibbs Sampling with People (Section 10.4.7), which offers a continuous sampling paradigm using a slider, enabling participants to manipulate a stimulus on a single dimension. This modality allows for greater flexibility in indicating preferences.

H6: Content VS Overall Image: when participants rate their liking of the image content, there will be significant difference with when they indicate the image liking of the overall image.

As described in Chapter 8, there is no significant difference between the liking ratings of participants who were instructed to rate the content of the images and those who were instructed to rate the overall image. Both treatments have satisfactory internal consistency within and between subjects. This indicates that the null hypothesis is not rebutted and that functional contemplation of stimuli with a practical aspect might not interfere with liking ratings on the overall image level. A possible explanation for this could be that functional artifacts are mostly produced to appeal to the general public, as also highlighted above in the summary of the control treatment results. It is possible that when using more controversial stimuli, the participants' liking ratings exhibit a greater divergence. A future recommendation regarding this matter has been formulated and is detailed in Section 10.4.3.

H7: Predictor Comparison: the predicted aesthetic scores assigned by the LAION Aesthetic Predictor are significantly different from the average image liking scores assigned by participants.

The combined quantitative and qualitative results indicate some agreement between the LAION Aesthetics Predictor and participants' ratings, but also reveal noticeable differences. The quantitative analysis shows a positive relationship, although not a perfect match. The qualitative analysis uncovers instances of disagreement, particularly when participants give more extreme ratings, and the predictor keeps assigning moderate scores.

The results indicate a partial alignment between the two, emphasising the need for further research. Replicating this analysis with a more controversial stimulus set, where participants may assign more extreme ratings, would be interesting to observe how the scores align with the ratings (as discussed in Section 10.4.3). If a lack of correlation between human ratings and predictor scores is found with more extreme ratings, it would suggest the necessity for additional research to enhance the predictor's accuracy for extreme ratings and consider a broader range of aesthetic preferences.

10.2. Interpretations of the findings and implications for the problem context

As mentioned earlier, there is a positive correlation between individuals' image ratings and the aesthetic scores provided by the LAION Aesthetic Predictor. However, it is important to note that this correlation is not flawless. Particularly, I observe that people sometimes give extreme ratings to stimuli, whereas the predictor tends to assign more moderate scores in those cases. When the question used by the LAION researchers in the development of the predictor's training dataset, "*how much do you like this image on a scale from 1-10?*", is compared with alternative questions and modalities that have support from the literature, no more appropriate alternative is found.

As discussed in Section 2.1.1, literature emphasises that aesthetic experiences consist of various facets. These include the embodied sensory experience (Merleau-Ponty, 1960: 1012), leading to a

disinterested encounter (Kant, 1790: 2000), which involves appraisal (Simpson, 1975), and induces a contemplative state in the beholder (Schopenhauer, 1818:2010; Chatterjee, 2002, 2003; Vartanian and Skov, 2014). From the perspective of philosophical aesthetics (and with substantiation from neuropsychological studies), it is remarkable that such a complex experience could be captured with the simple concept of '*liking*'. As described in the previous section, the experiments performed for this thesis are only a start of research in this direction. The findings of this thesis indicate that to make a conclusive statement about the appropriateness of the LAION researchers' approach, more in-depth research is needed.

10.3. Limitations of the experiments conducted in this thesis

In this thesis, it is important to acknowledge the limitations that exist within the scope of the conducted experiments. The objective of these experiments, as discussed in chapters 4, 5, 6, 7, and 8 is for exploration rather than validation. For this reason, small amounts of stimuli and participants were used, which makes it ill-considered to assign conclusive scientific value to the results.

Moreover, It should be noted that the high inter-rater reliability (Cronbach's alpha) for the stimulus sets already found in the control treatment, as reported in Section 10.1.1, may have been influenced by the selection of a specific image class, namely *buildings*. Buildings, which are often designed with the intention to please the masses, might not be the ideal representation for investigating diverse aesthetic perceptions. To encourage greater contrast between likings, task designs, and aesthetic metrics, future research is recommended to consider alternative image classes, such as the visual equivalent of Nickelback music.

It is also worth mentioning that for stimulus set 2, both sub-stimulus sets did not exhibit equal aesthetic distributions, which might have impacted the results, as mentioned in Section 3.5.2. In addition, the scope of the experiments was limited to stimuli from the image class *buildings*, making it unclear whether the results can be generalised to other image classes. Next to this, it is important to recognise that participants for the development of stimulus set 2 were recruited through personal connections, which could potentially introduce bias.

Next to this, the choice of a drag and drop ranking modality, described in Section 10.1.1, may have hindered participants from optimally ranking the stimuli. To explore this aspect more thoroughly, a different approach to investigating ranking modalities is suggested for future studies.

As no experiment yielded better results than the control treatment, no comparison was carried out with expert evaluations. Therefore, no conclusions can be drawn about the reliability of crowdworkers as a source of aesthetic judgements.

The analysis of whether participants' regions influenced their image liking ratings, as mentioned in Section 4.5, is considered unreliable and requires further research to draw any valuable conclusions on this front.

It is worth mentioning that experiments 1 and 2, as well as experiments 3 and 4, were administered with two different iterations of the stimulus set. Experiments 1 and 2 were compared to the control treatment conducted with stimulus set 1, and experiments 3 and 4 were compared to the control treatment conducted with stimulus set 2. However, this variation is not anticipated to significantly impact the research outcomes.

The internal consistency within subjects was not measured for the first two experiments. Nonetheless, since both experiments did not show significant differences from the control treatment, this is not deemed a major concern. Next to this, Experiment 3 failed to record the initial randomised ranking order of the stimuli which were presented to each participant, so the adjustments each participants made cannot be analysed post hoc.

For suggestions and recommendations for future research, I would like to refer you to Section 10.4.

10.4. Suggestions for future research

This section will go into more details on the suggestions for future research.

10.4.1. Semantic concept activation with the original semantic concepts

For the original experiment where the effect of semantic concept activation on the dynamics of aesthetic appreciation was investigated, Faerber et al. (2010) exposed participants to the following semantic concepts: *attractiveness*, *arousal*, *interestingness*, *valence*, *boredom* and *innovativeness*. It was decided to use the Unified Model of Aesthetics (Berghman and Hekkert, 2017) as semantic concept exposure in Experiment 1, as described in Section 2.3.4. Although Experiment 1 did not show that semantic concept activation with the UMA affects crowdworkers' likeness ratings, it can still be examined whether the original metrics of Faerber et al. (2010) affect likeness ratings.

10.4.2. Semantic concept activation with the Unified Model of Aesthetics with the original UMA stimulus set

Experiment 1 indicates that for this data, exposing participants to the UMA as semantic concept activation does not significantly affect participants' likeness ratings. What is noteworthy is that the qualitative results indicate that mainly aesthetic appreciation seems to correlate with liking ratings, and the different levels of the UMA do so only to a lesser extent. This observation raises the possibility that the UMA might be limited in its applicability to product aesthetics, the context in which it was initially verified by Berghman and Hekkert (2017). Experiment 1 was conducted with a stimulus set which contains images of buildings. To see if the UMA might be used for semantic concept activation in its original context, the study can be conducted again with the original stimulus set of Berghman and Hekkert (2017). This experiment was not conducted within this thesis because, after correspondence with the researchers, it appeared to not be possible to access the original stimulus set.

10.4.3. Different image classes

The findings of this thesis are limited to the specific stimulus set, *buildings*. To further investigate modalities under different conditions, different image classes, including non-functional depictions and potentially more controversial topics, should be considered. Exploring less curated stimulus sets, particularly when transitioning to annotating crawled datasets with high randomness, is crucial in the context of this thesis.

An experiment looking into the effects of a more controversial stimulus set

The influence of controversial image classes could be examined using a controlled stimulus set with categories: low aesthetic value of morally good scenes, high aesthetic value of morally good scenes, low aesthetic value of morally bad scenes, and high aesthetic value of morally bad scenes. This stimulus set will explore whether participants annotate topics or overall aesthetics of images. While writing this thesis, I came across the posters depicted in Figure 10.1 in a museum in Amsterdam. These propaganda posters are of high aesthetic quality (in my opinion). They were distributed by the German occupiers during the Second World War. The posters called on the Dutch population to work in Germany. This could possibly be an interesting entry point in developing such a type of stimulus set.



Figure 10.1: Propaganda posters created and distributed by the German occupiers during the Second World War. Artist unknown. Photograph taken by me.

10.4.4. More iterations on ranking modalities

As discussed in Section 10.1.1, it may be interesting to rerun Experiment 3 with different types of the ranking modality. In the current experiment, a drag and drop interaction was used, but possibly other interactions such as radio buttons, text boxes or select boxes are more suitable in this context. Additionally, it is interesting to compare the effects of presenting different amounts of stimuli simultaneously on participants and determine if it leads to improved internal consistency both within and between subjects.

10.4.5. More iterations on alternative-forced choice

Besides 2AFC, there are also studies that invoke 4AFC and 6AFC. For example, Yannakakis (2009) describes a protocol for preference learning, where participants are asked in a survey to evaluate pairs of alternative choices, in a 4AFC modality. Nakauchi et al. (2018) employed 4AFC to examine which features of colour compositions in judgement for art paintings are influential. Yu et al. (2020) employed a 6AFC modality, among others, to investigate the colour preference of participants. This is just a glimpse of n-alternative forced choice experiments deployed for preference learning in the realm of aesthetics. Hence, it may be interesting to see how different amounts of stimuli in this modality affect participant preferences.

10.4.6. An additional study examining the impact of regions on aesthetics.

In Experiment 1, the region was examined. No influence of region on image liking was found for this data. However, this analysis is considered unreliable due to several limitations. There is an unequal amount of participants across the regions. There are differing numbers of participants per sub-stimulus set group. This implicates that the sample is very small to investigate differences between regions. The results should therefore be interpreted with caution. It is important to repeat this study, with the region as the experiment variable. The reason for this is that it may have major implications for text-to-image models such as Stable Diffusion, should the experiment indicate that region affects aesthetic experiences. If this is the case it would be worthwhile to consider whether it would be useful to deploy different models for different regions.

10.4.7. Gibbs sampling with people for image aesthetics

Harrison et al. (2020) describe a generalised version of Markov Chain Monte Carlo with People (MCMCP) that they call Gibbs Sampling with People (GSP). This method can be used to investigate how humans derive semantic representation from stimuli. Examples include the colour of a crema layer of a cup of coffee but also the aesthetic value of an art work. GSP is a continuous sampling paradigm, in which participants manipulate one dimension of a stimulus at each turn by means of a slider, to optimise the stimulus with a given criterium (such as aesthetic value). This paradigm appears promising in the context of this thesis.

10.4.8. Different contexts and demographical groups

This thesis has solely focused on crowdworkers within a crowdsourcing context. It is worth considering that crowdsourcing might not be the most appropriate method for measuring aesthetic experiences. Several studies suggest that participants tend to have a heightened aesthetic experience within a museum environment compared to a laboratory context where stimuli are presented on a computer (Brieber, Nadal and Leder, 2015; Locher, Smith and Smith, 1999; Locher, Smith and Smith, 2001). In addition to exploring different contexts, it's important to acknowledge that even though crowdworkers may come from all over the world, it still remains a segment of our society. For future research, I recommend involving participants from diverse demographic groups to annotate images based on aesthetic metrics and compare their responses with those of crowdworkers. Furthermore, investigating the impact of different annotation contexts, such as online versus offline settings, on the results could yield valuable insights. Personally, I am particularly curious about how the control treatment compares to the same Likert scale question in a museum context. Exploring these aspects could enhance our understanding of aesthetic experiences in a broader sense.

10.5. Personal reflection

Regrettably, all good things must come to an end, and so must this thesis. Throughout my journey as a student (including high school), I have noticed that when I fail to see the purpose of a project or find the subject uninteresting, I can sometimes struggle to put in the effort required. On the flip side, when I am genuinely interested in the work I am doing, I become completely absorbed in it. That is precisely what happened during my thesis. I soon realised that I was profoundly enthusiastic about the subject, and that I genuinely enjoy conducting research.

I learnt a lot during my thesis, thanks in large to my supervisory team. Their expertise has steered my enthusiasm in the right direction and because of the pleasant collaboration, I consider this project of sufficient quality and am happy with the final result.

Because I enjoyed working on my thesis so much, there were times when it proved challenging to close my project in my mind at 18:00 too. For instance, during family gathering my mother implemented the following rule: no *'nerding'* allowed (meaning no research talk). While some people questioned my hyperfocus, I personally did not perceive it as a problem, as this work did not really feel like work to me. However, I did have to learn not to take it personally when things went less well with the project, and to separate criticism of the project from criticism of me personally. Next to this, I was able to further sharpen my research skills, which was my main learning goal for taking on this thesis.

In high school, I earned high grades in the subjects I enjoyed (maths, physics, chemistry, Dutch language) but almost did not pass because I found the German language so uninteresting that I could hardly motivate myself to put in the effort required to pass. My mentor, Mr Teuben, then spoke the grand words, "*Céline, there will always be a German language in life.*" What he meant by this is that there is always something you do not like, but you still have to work for it to achieve your goals. This has always held true until now, but with my thesis I found out that there are exceptions to this rule after all.

11

Conclusion

This thesis aimed to explore alternative annotation methods for measuring aesthetics in images and their applicability in developing visual training datasets for AI-driven models, via the research question:

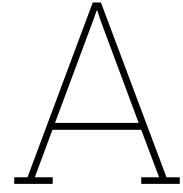
How does the annotation method used in the development of the LAION-Aesthetics V2 datasets compare to other annotation methods for measuring aesthetics?

The thesis delved into different levels of inquiry: alternative task design and alternative metrics, aiming to determine the most suitable approach for annotating aesthetics in the LAION-Aesthetics V2 dataset. To explore this, a series of experiments compared the Likert scale question *"how much do you like this image on a scale from 1-10?"* with semantic concept activation, alternative question phrasing, ranking, and alternative-forced choice (2AFC). Additionally, three hypotheses investigated aesthetics annotation for visual training datasets, comparing participants' ratings based on content liking to ratings on overall image liking. Research conducted also compared the LAION Aesthetics Predictor scores with participant liking ratings and examined the influence of region on image liking ratings.

The comparison between the LAION aesthetic approach and scientifically based alternatives showed no significant difference in performance. The ranking treatment even performed significantly worse. Interestingly, participants' geographical region had no observable influence on image ratings. Moreover, there was no significant difference between overall image preference and content preference. Partial alignment was found between human liking ratings and the LAION Aesthetics predictor scores. There is an indication that misalignment occurs when humans assign more extreme ratings but the predictor keeps aligning moderate scores. These findings suggest the need for further research to determine whether the concept of simple "liking" is a relevant and appropriate method to capture aesthetics.

This thesis acknowledges several limitations in the conducted experiments, such as small sample sizes, limited image classes, a possibly unfortunately chosen image class, and unequal distribution of participants across regions. These factors may have influenced the results and call for further research. The exploration of different image classes and contexts is recommended to validate the findings across various demographic groups and environments. Additionally, experimenting with different modalities for ranking and alternative-forced choice could offer valuable insights into eliciting aesthetic experiences. Although the research has shed light on the alignment between the LAION Aesthetic Predictor and human ratings, the influence of regions on aesthetic experiences requires more extensive investigation.

In conclusion, this thesis provides valuable insights into the annotation methods for measuring aesthetics in visual datasets and emphasises the importance of continued research to better understand and cater to diverse aesthetic preferences and experiences in the realm of text-to-image generators.



Bibliography

Althuizen, Niek. "Revisiting Berlyne's Inverted U-shape Relationship between Complexity and Liking: The Role of Effort, Arousal, and Status in the Appreciation of Product Design Aesthetics." *Psychology and Marketing* 38, no. 3 (2021): 481–503. <https://doi.org/10.1002/mar.21449>.

Alwin, Duane F., and Jon A. Krosnick. "The Measurement of Values in Surveys: A Comparison of Ratings and Rankings." *The Public Opinion Quarterly* 49, no. 4 (1985): 535–52. <https://www.jstor.org/stable/2748921>.

Augustin, M.D., Claus-Christian Carbon, and Johan Wagemans. "Measuring Aesthetic Impressions of Visual Art." *Perception*, 2011, 40–219.

Baio, Andy. "Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator." *Waxy.Org* (blog), August 30, 2022. <https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/>.

Bandura, Albert. "Self-Efficacy: Toward a Unifying Theory of Behavioral Change." *Psychological Review* 84, no. 2 (1977): 191–215. <https://doi.org/10.1037/0033-295X.84.2.191>.

Bara, Ionela, Richard J. Binney, and Richard Ramsey. "Investigating the Role of Executive Resources across Aesthetic and Non-Aesthetic Judgments." Preprint. *PsyArXiv*, November 2, 2021. <https://doi.org/10.31234/osf.io/ydmbr>.

Barkow, Jerome H., Leda Cosmides, and John Tooby. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. [2nd. ed.], 1st issued as an Oxford university press paperback, cop. 1992. New York (N.Y.): Oxford university press, 1995.

Baumgarten, Alexander Gottlieb. "Esthetica/Ästhetik." Edited by Dagmar Mirbach 2 (2007). <https://doi.org/DOI:10.33134/eeja.54>.

Berghman, Michaël, and Paul Hekkert. "Towards a Unified Model of Aesthetic Pleasure in Design." *New Ideas in Psychology* 47 (2017): 136–44. <https://doi.org/10.1016/j.newideapsych.2017.03.004>.

Berleant, Arnold. "Aesthetic Sensibility." *Ambiances*, March 30, 2015. <https://doi.org/10.4000/ambiances.526>.

Berlyne, D. E. "Novelty, Complexity, and Hedonic Value." *Perception and Psychophysics* 8, no. 5 (September 1, 1970): 279–86. <https://doi.org/10.3758/BF03212593>.

Berlyne, D.E. "Aesthetics and Psychobiology." *Appleton-Century-Crofts*, 1971. <https://psycnet.apa.org/record/1973-00821-000>.

Bhattacharya, Subhabrata, Rahul Sukthankar, and Mubarak Shah. "A Framework for Photo-Quality Assessment and Enhancement Based on Visual Aesthetics." In Proceedings of the 18th ACM International Conference on Multimedia, 271–80. Firenze Italy: ACM, 2010. <https://doi.org/10.1145/1873951.1873990>.

Bryk, Erdem, Nicolas Huynh, Mykel J. Kochenderfer, and Dorsa Sadigh. "Active Preference-Based Gaussian Process Regression for Reward Learning," 2020. <https://doi.org/10.48550/ARXIV.2005.02575>.

Blijlevens, Janneke, and Paul Hekkert. "'Autonomous, yet Connected': An Esthetic Principle Explaining Our Appreciation of Product Designs." *Psychology and Marketing* 36, no. 5 (2019): 530–46. <https://doi.org/10.1002/mar.21195>.

Blijlevens, J., P. Hekkert, and C. Thurgood. "The Joint Effect of Typicality and Novelty on Aesthetic Pleasure for Product Designs: Influences of Safety and Risk," 2014. <https://www.semanticscholar.org/paper/The-joint-effect-of-typicality-and-novelty-on-for-Blijlevens-Hekkert/a1249f9df3b267ed7e0246b4d66e6381c622bdf1>.

Blijlevens, Janneke, Clementine Thurgood, Paul Hekkert, Lin-Lin Chen, Helmut Leder, and T. W. Allan Whitfield. "The Aesthetic Pleasure in Design Scale: The Development of a Scale to Measure Aesthetic Pleasure for Designed Artifacts." *Psychology of Aesthetics, Creativity, and the Arts* 11, no. 1 (2017): 86–98. <https://doi.org/10.1037/aca0000098>.

Blizek, William, and D. E. Berlyne. "Aesthetics and Psychobiology." *The Journal of Aesthetics and Art Criticism* 31, no. 4 (1973): 553. <https://doi.org/10.2307/429334>.

Boeijen, Annemiek van, Jaap Daalhuizen, Jelle Zijlstra, Roo van der Schoor, and Technische Universiteit Delft, eds. *Delft Design Guide: Design Methods*. Revised 2nd edition. Amsterdam: BIS Publishers, 2014.

Bornstein, Robert F. "Exposure and Affect: Overview and Meta-Analysis of Research, 1968–1987." *Psychological Bulletin* 106, no. 2 (1989): 265–89. <https://doi.org/10.1037/0033-2909.106.2.265>.

Brieber, David, Marcos Nadal, and Helmut Leder. "In the White Cube: Museum Context Enhances the Valuation and Memory of Art." *Acta Psychologica* 154 (2015): 36–42. <https://doi.org/10.1016/j.actpsy.2014.11.004>.

Brisco, Ross, Laura Hay, and Sam Dhami. "EXPLORING THE ROLE OF TEXT-TO-IMAGE AI IN CONCEPT GENERATION." *Proceedings of the Design Society* 3 (July 2023): 1835–44. <https://doi.org/10.1017/pds.2023.184>.

Brown, James Dean. "Questions and Answers about Language Testing Statistics: The Cronbach Alpha Reliability Estimate," 2002. <https://hosted.jalt.org/test/bro13.htm>.

Bruens, G., 2007. *Form/Color anatomy*. Utrecht: Lemma.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6, no. 1 (2011): 3–5. <https://doi.org/10.1177/1745691610393980>.

Cai, Kenrick, Iain Martin. "The AI Founder Taking Credit For Stable Diffusion's Success Has A History Of Exaggeration." *Forbes*, 2023. <https://www.forbes.com/sites/kenrickcai/2023/06/04/stable-diffusion-emad-mostaque-stability-ai-exaggeration/>.

Carbon, Claus-Christian, Thomas Grüter, Martina Grüter, Joachim E. Weber, and Andreas Lueschow.

“Dissociation of Facial Attractiveness and Distinctiveness Processing in Congenital Prosopagnosia.” *Visual Cognition* 18, no. 5 (2010): 641–54. <https://doi.org/10.1080/13506280903462471>.

Cela-Conde, Camilo J., Luigi Agnati, Joseph P. Huston, Francisco Mora, and Marcos Nadal. “The Neural Foundations of Aesthetic Appreciation.” *Progress in Neurobiology* 94, no. 1 (2011): 39–48. <https://doi.org/10.1016/j.pneurobio.2011.03.003>.

Chamorro-Premuzic, T., Furnham, A., and Reimers, S. (2007). *The artistic personality*. *The Psychologist*, 20(2), 84–87.

Chatterjee, Anjan. “Prospects for a Cognitive Neuroscience of Visual Aesthetics: (514602010-003).” American Psychological Association, 2003. <https://doi.org/10.1037/e514602010-003>.

Chatterjee, Md, Anjan. *The Aesthetic Brain: How We Evolved to Desire Beauty and Enjoy Art*. Oxford University Press, 2013. <https://doi.org/10.1093/acprof:oso/9780199811809.001.0001>.

Chittilappilly, Anand Inasu, Lei Chen, and Sihem Amer-Yahia. “A Survey of General-Purpose Crowdsourcing Techniques.” *IEEE Transactions on Knowledge and Data Engineering* 28, no. 9 (September 1, 2016): 2246–66. <https://doi.org/10.1109/TKDE.2016.2555805>.

Christoph, Schumann. “LAION Aesthetics.” Github, 2022. <https://github.com/LAION-AI/laion-datasets/blob/main/laion-aesthetic.md>.

Chwilla, Dorothee J., Peter Hagoort, and C.M. Brown. “The Mechanism Underlying Backward Priming in a Lexical Decision Task: Spreading Activation versus Semantic Matching.” *The Quarterly Journal of Experimental Psychology Section A* 51, no. 3 (1998): 531–60. <https://doi.org/10.1080/713755773>.

Collins, Allan M., and Elizabeth F. Loftus. “A Spreading-Activation Theory of Semantic Processing.” *Psychological Review* 82, no. 6 (1975): 407–28. <https://doi.org/10.1037/0033-295X.82.6.407>.

Cotter, Katherine N., Rebekah M. Rodriguez-Boerwinkle, Alexander P. Christensen, Anna Fekete, Jeffrey K. Smith, Lisa F. Smith, Pablo P. L. Tinio, and Paul J. Silvia. “Updating the Aesthetic Fluency Scale: Revised Long and Short Forms for Research in the Psychology of the Arts.” Edited by Frantisek Sudzina. *PLOS ONE* 18, no. 2 (February 8, 2023): e0281547. <https://doi.org/10.1371/journal.pone.0281547>.

Crilly, Nathan, James Moultrie, and P. John Clarkson. “Seeing Things: Consumer Response to the Visual Domain in Product Design.” *Design Studies* 25, no. 6 (2004): 547–77. <https://doi.org/10.1016/j.destud.2004.03.001>.

Daniel, Florian, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Al-lahbakhsh. “Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions.” *ACM Computing Surveys* 51, no. 1 (January 31, 2019): 1–40. <https://doi.org/10.1145/3148148>.

Datta, Ritendra, Dhiraj Joshi, Jia Li, and James Z. Wang. “Studying Aesthetics in Photographic Images Using a Computational Approach.” In *Computer Vision – ECCV 2006*, edited by Aleš Leonardis, Horst Bischof, and Axel Pinz, 288–301. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, 2006. <https://doi.org/10.1007/11744078>.

Datta, Ritendra, jia Li, and James Z. Wang. “Algorithmic Inferencing of Aesthetics and Emotion in Natural Images: An Exposition.” In *15th IEEE, 2008*. <https://doi.org/10.1109/ICIP.2008.4711702>.

Deci, Edward L., and Richard M. Ryan. “The ‘What’ and ‘Why’ of Goal Pursuits: Human Needs and the Self-Determination of Behavior.” *Psychological Inquiry* 11, no. 4 (2000): 227–68. https://doi.org/10.1207/S15327965PLI1104_01.

- Faerber, Stella J., Helmut Leder, Gernot Gerger, and Claus-Christian Carbon. "Priming Semantic Concepts Affects the Dynamics of Aesthetic Appreciation." *Acta Psychologica* 135, no. 2 (2010): 191–200. <https://doi.org/10.1016/j.actpsy.2010.06.006>.
- Fang, Yuming, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. "Perceptual Quality Assessment of Smartphone Photography," 3677–86, 2020.
- Feitosa, Jennifer, Dana L. Joseph, and Daniel A. Newman. "Crowdsourcing and Personality Measurement Equivalence: A Warning about Countries Whose Primary Language Is Not English." *Personality and Individual Differences* 75 (2015): 47–52. <https://doi.org/10.1016/j.paid.2014.11.017>.
- Gero, John S., and Udo Kannengiesser. "The Function-Behaviour-Structure Ontology of Design." In *An Anthology of Theories and Models of Design*, edited by Amaresh Chakrabarti and Lucienne T. M. Blessing, 263–83. London: Springer London, 2014. https://doi.org/10.1007/978-1-4471-6338-1_13.
- Giacalone, Davide, Mette Duerlund, Jannie Bøegh-Petersen, Wender L.P. Bredie, and Michael Bom Frøst. "Stimulus Collative Properties and Consumers' Flavor Preferences." *Appetite* 77 (2014): 20–30. <https://doi.org/10.1016/j.appet.2014.02.007>.
- Goldstein, E. Bruce. *Sensation and Perception*. 6th ed. Australia; [Pacific Grove, CA]: Wadsworth-Thomson Learning, 2002. Government of Canada, Statistics Canada. "Classification of Gender - 1 - Man," October 18, 2021. <https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1326727&CVD=1326727&CLV=0&MLV=1&D=1>.
- Gracyk, Theodore. "Hume's Aesthetics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University, 2021. <https://plato.stanford.edu/archives/win2021/entries/hume-aesthetics/>.
- Guyer, Paul. "18th Century German Aesthetics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2020. Metaphysics Research Lab, Stanford University, 2020. <https://plato.stanford.edu/archives/fall2020/entries/aesthetics-18th-german/>.
- Hanafiah, Mohd. (2015). Re: What can I do if I got negative cronbach alpha value?. Retrieved from: <https://www.researchgate.net/post/What-can-I-do-if-I-got-negative-cronbach-alpha-value/564621e86225ff45ff8b4588/citation/download>.
- Harrison, Peter M. C., Raja Marjeh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. "Gibbs Sampling with People," 2020. <https://doi.org/10.48550/ARXIV.2008.02595>.
- Hassenzahl, Marc, and Andrew Monk. "The Inference of Perceived Usability From Beauty." *Human-Computer Interaction* 25, no. 3 (August 31, 2010): 235–60. <https://doi.org/10.1080/07370024.2010.500139>.
- Hayes, Ben, and Alben Kuyumdzhieva. "Ethics and Data Protection." European Commission, 2021.
- He, Shuai, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. "Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks." (IJCAI-22). Austria, 2022. <https://www.ijcai.org/proceedings/2022/0132.pdf>.
- Hekkert, Paul. "Design Aesthetics: Principles of Pleasure in Design." *Psychology Science* 6, no. 2 (2006): 157–72. <https://research.tudelft.nl/en/publications/design-aesthetics-principles-of-pleasure-in-design>.

Hekkert, Paul, and Helmut Leder. "PRODUCT AESTHETICS." In *Product Experience*, 259–85. Elsevier, 2008. <https://doi.org/10.1016/B978-008045089-6.50013-7>.

Hekkert, Paul, Dirk Snelders, and Piet C. W. Wieringen. "Most Advanced, yet Acceptable: Typicality and Novelty as Joint Predictors of Aesthetic Preference in Industrial Design." *British Journal of Psychology* 94, no. 1 (2003): 111–24. <https://doi.org/10.1348/000712603762842147>.

Hesslinger, Vera M., Claus-Christian Carbon, and Heiko Hecht. "Social Factors in Aesthetics: Social Conformity Pressure and a Sense of Being Watched Affect Aesthetic Judgments." *I-Perception* 8, no. 6 (2017): 204166951773632. <https://doi.org/10.1177/2041669517736322>.

Hettiachchi, Danula, Senuri Wijenayake, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. "How Context Influences Cross-Device Task Acceptance in Crowd Work." *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8* (October 1, 2020): 53–62. <https://doi.org/10.1609/hcomp.v8i1.7463>.

Higgins, J.M., 1994. *101 Problem Solving Techniques*. New York: New Management Publishing Company.

Hosu, Vlad, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. "KoniQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment," 2019. <https://doi.org/10.48550/ARXIV.1910.06180>.

Hung, Wei-Ken, and Lin-Lin Chen. "Effects of Novelty and Its Dimensions on Aesthetic Preference in Product Design." *International Journal of Design* 6, no. 2 (2012): 81–90. <http://www.ijdesign.org/index.php/IJDesign/article/viewFile/1146/474>.

Johnston, Victor. "The Origin and Function of Pleasure." *Cognition and Emotion* 17, no. 2 (2003): 167–79. <https://doi.org/10.1080/02699930302290>.

Joshi, Dhiraj, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Wang, Jia Li, and Jiebo Luo. "Aesthetics and Emotions in Images." *IEEE Signal Processing Magazine* 28, no. 5 (2011): 94–115. <https://doi.org/10.1109/MSP.2011.941851>.

Kang, C, G Valenzise, and F Dufaux. "EVA: An Explainable Visual Aesthetics Dataset," 5–13. Seattle, United States, 2020. <https://doi.org/ff10.1145/3423268.3423590ff>.

Kant, Immanuel. *Critique of the Power of Judgment*. Edited by Paul Guyer. Translated by Eric Matthews. 1st ed. Cambridge University Press, 2000. <https://doi.org/10.1017/CB09780511804656>.

King, Alexandra. "Aesthetic Attitude | Internet Encyclopedia of Philosophy." Peer reviewed blog. *The Aesthetic Attitude*(blog). Accessed March 13, 2023. <https://iep.utm.edu/aesthetic-attitude/>.

King, Alexandra. "The Aesthetic Attitude." In *Internet Encyclopedia of Philosophy*, edited by J. Feiser and B. Dowden, 2012.

Kleine, Robert E., Susan Schultz Kleine, and Jerome B. Kernan. "Mundane Consumption and the Self: A Social Identity Perspective." *Journal of Consumer Psychology* 2, no. 3 (1993): 209–35. [https://doi.org/10.1016/S1057-7408\(08\)80015-0](https://doi.org/10.1016/S1057-7408(08)80015-0).

Knox, Israel. *The Aesthetic Theories of Kant, Hegel, and Schopenhauer*. Thames and Hudson, 1958.

Kong, Shu, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. "Photo Aesthetics

Ranking Network with Attributes and Content Adaptation,” 2016. <https://doi.org/10.48550/ARXIV.1606.01621>.

Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen. “Aesthetic Critiques Generation for Photos.” In 2017 IEEE International Conference on Computer Vision (ICCV), 3534–43. Venice: IEEE, 2017. <https://doi.org/10.1109/ICCV.2017.380>.

Langlois, Judith H., Lisa Kalakanis, Adam J. Rubenstein, Andrea Larson, Monica Hallam, and Monica Smoot. “Maxims or Myths of Beauty? A Meta-Analytic and Theoretical Review.” *Psychological Bulletin* 126, no. 3 (2000): 390–423. <https://doi.org/10.1037/0033-2909.126.3.390>.

Locher, Paul, Lisa Smith, and Jeffrey Smith. “Original Paintings versus Slide and Computer Reproductions: A Comparison of Viewer Responses.” *Empirical Studies of the Arts* 17, no. 2 (1999): 121–29. <https://doi.org/10.2190/R1WN-TAF2-376D-EFUH>.

Locher, Paul J, Jeffrey K Smith, and Lisa F Smith. “The Influence of Presentation Format and Viewer Training in the Visual Arts on the Perception of Pictorial and Aesthetic Qualities of Paintings.” *Perception* 30, no. 4 (2001): 449–65. <https://doi.org/10.1068/p3008>.

Li, Taylor. “Aesthetic Disinterestedness,” 2018. <https://era.ed.ac.uk/handle/1842/35655>.

Loewy, Raymond. *Never Leave Well Enough Alone*. Simon and Schuster, 1951.

Lomas, Derek, Willem van der Maden, Giovanni Lion, Sohom Bandyopadhyay, Yanna Litowsky, Haiyan Xue, Pieter Desmet. “The Alignment of AI Emotions: human ratings of the emotions expressed by GPT-3, DALL-E and Stable Diffusion” (2023) [manuscript in preparation].

Luo, Wei, Xiaou Tang, and Xiaogang Wang. “Content-Based Photo Quality Assessment.” *IEEE Transactions on Multimedia* 15, no. 8 (2011): 1930–43. <https://doi.org/10.1109/TMM.2013.2269899>.

Mandoki, Katya. *Everyday Aesthetics: Prosaics, the Play of Culture and Social Identities* (2007, 2016 Routledge), 2007. https://www.academia.edu/1304855/Everyday_Aesthetics_Prosaics_the_Play_of_Culture_and_Social_Identities_2007_2016_Routledge_.

Marković, Slobodan. “Components of Aesthetic Experience: Aesthetic Fascination, Aesthetic Appraisal, and Aesthetic Emotion.” *I-Perception* 3, no. 1 (2012): 1–17. <https://doi.org/10.1068/i0450aap>.

Matthews, Gerald, Ian J. Deary, and Martha C. Whiteman. *Personality Traits*. Cambridge University Press, 2003.

McManus, I. C., and A. Furnham. “Aesthetic Activities and Aesthetic Attitudes: Influences of Education, Background and Personality on Interest and Involvement in the Arts.” *British Journal of Psychology* 97, no. 4 (2006): 555–87. <https://doi.org/10.1348/000712606X101088>.

Merleau-Ponty, Maurice, Claude Lefort, Jenny Slatman, Rens Vlasblom, and Ineke van der Burg. *Oog en geest*. Heruitg. Amsterdam: Parrèsia, 2011. Meyers-Levy, Joan, and Alice M. Tybout. “Schema Congruity as a Basis for Product Evaluation.” *Journal of Consumer Research* 16, no. 1 (1989): 39. <https://doi.org/10.1086/209192>.

Muller, W., 2001. *Order and Meaning in Design*. Utrecht: Lemma.

Murray, Naila, Luca Marchesotti, and Florent Perronnin. “AVA: A Large-Scale Database for Aesthetic Visual Analysis.” Providence, RI, USA, 2012. <https://doi.org/10.1109/CVPR.2012.6247954>.

Nakauchi, Shigeki, Taisei Kondo, Hiroshi Higashi, João Linhares, and Sérgio Nascimento. “Color

Statistics Underlying Preference Judgement for Art Paintings.” *Journal of Vision* 18, no. 10 (September 1, 2018): 867. <https://doi.org/10.1167/18.10.867>.

Nanda, Upali, Debajyoti Pati, Hessam Ghamari, and Robyn Bajema. “Lessons from Neuroscience: Form Follows Function, Emotions Follow Form.” *Intelligent Buildings International* 5, no. sup1 (2013): 61–78. <https://doi.org/10.1080/17508975.2013.807767>.

Nemecek, Jan, and Etienne Grandjean. “Noise in Landscaped Offices.” *Applied Ergonomics* 4, no. 1 (1973): 19–22. [https://doi.org/10.1016/0003-6870\(73\)90006-9](https://doi.org/10.1016/0003-6870(73)90006-9).

Nguyen, Tam V., Si Liu, Bingbing Ni, Jun Tan, Yong Rui, and Shuicheng Yan. “Sense Beauty via Face, Dressing, and/or Voice.” In *Proceedings of the 20th ACM International Conference on Multimedia*, 239–48. Nara Japan: ACM, 2012. <https://doi.org/10.1145/2393347.2393385>.

Ozili, Peterson K. “The Acceptable R-Square in Empirical Modelling for Social Science Research.” MPRA Paper, 2023. <https://mpra.ub.uni-muenchen.de/115769/>.

P, J-D. “Simulacra-Aesthetic-Captions.” Github, 2022. <https://github.com/JD-P/simulacra-aesthetic-captions>.

Page, Christine, and Paul M. Herr. “An Investigation of the Processes by Which Product Design and Brand Strength Interact to Determine Initial Affect and Quality Judgments.” *Journal of Consumer Psychology* 12, no. 2 (2002): 133–47. https://doi.org/10.1207/S15327663JCP1202_06.

Palmer, Stephen E., Karen B. Schloss, and Jonathan Sammartino. “Visual Aesthetics and Human Preference.” *Annual Review of Psychology* 64, no. 1 (January 3, 2013): 77–107. <https://doi.org/10.1146/annurev-psych-120710-100504>.

Piñeiro, Gervasio, Susana Perelman, Juan P. Guerschman, and José M. Paruelo. “How to Evaluate Models: Observed vs. Predicted or Predicted vs. Observed?” *Ecological Modelling* 216, no. 3–4 (2008): 316–22. <https://doi.org/10.1016/j.ecolmodel.2008.05.006>.

Post, R.A.G., J. Blijlevens, and P. Hekkert. “‘To Preserve Unity While Almost Allowing for Chaos’: Testing the Aesthetic Principle of Unity-in-Variety in Product Design.” *Acta Psychologica* 163 (2016): 142–52. <https://doi.org/10.1016/j.actpsy.2015.11.013>.

Postrel, Virginia I. *The Substance of Style: How the Rise of Aesthetic Value Is Remaking Commerce, Culture, and Consciousness*. 1. Perennial ed., [Nachdr.]. New York: Perennial, 2005.

Reber, Rolf, Norbert Schwarz, and Piotr Winkielman. “Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver’s Processing Experience?” *Personality and Social Psychology Review* 8, no. 4 (2004): 364–82. https://doi.org/10.1207/s15327957pspr0804_3.

Redi, Judith Alice, Tobias Hoßfeld, Pavel Korshunov, Filippo Mazza, Isabel Pova, and Christian Keimel. “Crowdsourcing-Based Multimedia Subjective Evaluations: A Case Study on Image Recognizability and Aesthetic Appeal.” In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, 29–34. Barcelona Spain: ACM, 2013. <https://doi.org/10.1145/2506364.2506368>.

Ren, Jian, Xiaohui Shen, Zhe Lin, Radomír Měch, and David J. Foran. “Personalized Image Aesthetics.” In *IEEE*, 638–47, 2017. https://openaccess.thecvf.com/content_ICCV_2017/papers/Ren_Personalized_Image_Aesthetics_ICCV_2017_paper.pdf.

Roozenburg, N.F.M. and Eekels, J., 1995. *Product Design: Fundamentals and Methods*. Utrecht: Lemma.

Sadigh, Dorsa, Anca Dragan, Shankar Sastry, and Sanjit Seshia. "Active Preference-Based Learning of Reward Functions." In *Robotics: Science and Systems XIII*. Robotics: Science and Systems Foundation, 2017. <https://doi.org/10.15607/RSS.2017.XIII.053>.

Saito, Yuriko. "Aesthetics of the Everyday." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2021. Metaphysics Research Lab, Stanford University, 2021. <https://plato.stanford.edu/archives/spr2021/entries/aesthetics-of-everyday/>.

Saito, Yuriko. *Aesthetics of the Familiar: Everyday Life and World-Making*. First edition. Oxford, United Kingdom: Oxford University Press, 2017.

Saito, Yuriko. *Everyday Aesthetics*. Oxford; New York: Oxford University Press, 2010.

Schneider, Silke L. "The International Standard Classification of Education 2011." In *Comparative Social Research*, edited by Gunn Elisabeth Birkelund, 30:365–79. Emerald Group Publishing Limited, 2013. [https://doi.org/10.1108/S0195-6310\(2013\)0000030017](https://doi.org/10.1108/S0195-6310(2013)0000030017).

Schopenhauer, Arthur. *The World as Will and Representation*. Edited by Judith Norman, Alistair Welchman, and Christopher Janaway. The Cambridge Edition of the Works of Schopenhauer. Cambridge; New York: Cambridge University Press, 2010.

Schuhmann, Christoph. "Aesthetic Subsets in LAION 2170337258 Samples." captions.christoph-schuhmann. Accessed June 5, 2023. http://captions.christoph-schuhmann.de/aesthetic_viz_laion_sac+logos+ava1-114-linearMSE-en-2.37B.html.

Schuhmann, Christoph. "LAION-Aesthetics | LAION," 2022. <https://laion.ai/blog/laion-aesthetics>.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, et al. "LAION-5B: An Open Large-Scale Dataset for Training next Generation Image-Text Models." arXiv, October 15, 2022. <https://doi.org/10.48550/arXiv.2210.08402>.

Semin, Gün R., and Christianne J. De Poot. "Bringing Partiality to Light: Question Wording and Choice as Indicators of Bias." *Social Cognition* 15, no. 2 (1997): 91–106. <https://doi.org/10.1521/soco.1997.15.2.91>.

Semin, Gün R., and Christianne J. De Poot. "The Question–Answer Paradigm: You Might Regret Not Noticing How a Question Is Worded." *Journal of Personality and Social Psychology* 73, no. 3 (1997): 472–80. <https://doi.org/10.1037/0022-3514.73.3.472>.

Shapshay, Sandra. "Schopenhauer's Aesthetics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University, 2021. <https://plato.stanford.edu/archives/win2021/entries/schopenhauer-aesthetics/>.

Shehata, Mohamed. (2018). Re: What are the reasons behind negative Cronbach Alpha for a variable, while doing the pilot study analysis in SPSS?. Retrieved from: https://www.researchgate.net/post/What_are_the_reasons_behind_negative_Cronbach_Alpha_for_a_variable_while_doing_the_pilot_study_analysis_in_SPSS/5b7b3f38a5a2e2a74a1dd9fc/citation/download.

Siahaan, Ernestasia, Alan Hanjalic, and Judith Redi. "A Reliable Methodology to Collect Ground Truth Data of Image Aesthetic Appeal." In *IEEE Transactions on Multimedia*, 1338–50, 2016.

Silberman, M. Six, Lilly Irani, and Joel Ross. "Ethics and Tactics of Professional Crowdwork." *XRDS: Crossroads, The ACM Magazine for Students* 17, no. 2 (2010): 39–43. <https://doi.org/10.1145/1869086.1869100>.

- Simpson, Evan. "Aesthetic Appraisal." *Philosophy* 50, no. 192 (1975): 189–204. <https://www.jstor.org/stable/3749507>.
- Skov, Martin, and Marcos Nadal. "There Are No Aesthetic Emotions: Comment on Menninghaus et al. (2019)." *Psychological Review* 127, no. 4 (2020): 640–49. <https://doi.org/10.1037/rev0000187>.
- Smith, L.F., and J.K. Smith. "The Nature and Growth of Aesthetic Fluency." *New Directions in Aesthetics, Creativity and the Arts* In P. Locher, C. Martindale, L. Dorfman (Eds.), no. 47–58 (2006). <https://psycnet.apa.org/record/2006-03935-004>.
- Sonderegger, Andreas, and Juergen Sauer. "The Influence of Design Aesthetics in Usability Testing: Effects on User Performance and Perceived Usability." *Applied Ergonomics* 41, no. 3 (2010): 403–10. <https://doi.org/10.1016/j.apergo.2009.09.002>.
- Sun, Peng, and Kathryn T. Stolee. "Exploring Crowd Consistency in a Mechanical Turk Survey." In *2016 IEEE/ACM 3rd International Workshop on CrowdSourcing in Software Engineering (CSI-SE)*, 8–14, 2016. <https://doi.org/10.1145/2897659.2897662>.
- Surowiecki, James. *The Wisdom of Crowds*. Nachdr. New York, NY: Anchor Books, 2005.
- Swanson R, D. Escoffery and A. Jhala, "Learning visual composition preferences from an annotated corpus generated through gameplay," *2012 IEEE Conference on Computational Intelligence and Games (CIG)*, Granada, Spain, 2012, pp. 363-370, doi:10.1109/CIG.2012.6374178.
- Szubielska, Magdalena, Kamil Imbir, and Anna Szymańska. "The Influence of the Physical Context and Knowledge of Artworks on the Aesthetic Experience of Interactive Installations." *Current Psychology* 40, no. 8 (August 1, 2021): 3702–15. <https://doi.org/10.1007/s12144-019-00322-w>.
- Thurgood, Clementine, Paul Hekkert, and Janneke Blijlevens. "The Joint Effect of Typicality and Novelty on Aesthetic Pleasure for Product Designs: Influences of Safety and Risk.," 2014.
- Toadvine, Ted. "Maurice Merleau-Ponty." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019. Metaphysics Research Lab, Stanford University, 2019. <https://plato.stanford.edu/archives/spr2019/entries/merleau-ponty/>.
- Umeda, Y., and T. Tomiyama. "Functional Reasoning in Design." *IEEE Expert* 12, no. 2 (1997): 42–48. <https://doi.org/10.1109/64.585103>.
- Urdan, Tim, and Frank Pajares. *Self-Efficacy Beliefs of Adolescents*. IAP, 2006.
- Van Der Heijden, Hans. "Factors Influencing the Usage of Websites: The Case of a Generic Portal in The Netherlands." *Information and Management* 40, no. 6 (2003): 541–49. [https://doi.org/10.1016/S0378-7206\(02\)00079-4](https://doi.org/10.1016/S0378-7206(02)00079-4).
- Vartanian, Oshin, and Martin Skov. "Neural Correlates of Viewing Paintings: Evidence from a Quantitative Meta-Analysis of Functional Magnetic Resonance Imaging Data." *Brain and Cognition* 87 (2014): 52–56. <https://doi.org/10.1016/j.bandc.2014.03.004>.
- Vincent, James. "The Scary Truth about AI Copyright Is Nobody Knows What Will Happen Next." *The Verge*, November 15, 2022. <https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data>.
- Wagemans, Johan, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R. Pomerantz, Peter A. Van Der Helm, and Cees Van Leeuwen. "A Century of Gestalt Psychology in Visual Perception: II. Conceptual and Theoretical Foundations." *Psychological Bulletin* 138, no. 6 (2012): 1218–52. <https://doi.org/10.1037/a0029334>.

Wang, Yihong, Konstantinos Papangelis, Michael Saker, Ioanna Lykourantzou, Alan Chamberlain, and Vassilis-Javed Khan. "Crowdsourcing in China: Exploring the Work Experiences of Solo Crowdworkers and Crowdfarm Workers." In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–13. Honolulu HI USA: ACM, 2020. <https://doi.org/10.1145/3313831.3376473>.

TU Delft. "Research Design 1: Minimising risk." Accessed June 12, 2023. <https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics/research-design-1-minimising-risk>.

U.S. Copyright Office. "U.S. Copyright Office Fair Use Index," 2023. <https://www.copyright.gov/fair-use/>.

Wu, Xiaoshi, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. "Better Aligning Text-to-Image Models with Human Preference," 2023. <https://doi.org/10.48550/ARXIV.2303.14420>.

Yannakakis, G. N. "Preference learning for affective modeling," 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, Netherlands, 2009, pp. 1-6, doi:10.1109/ACII.2009.5349491.

Yannakakis, Georgios N., and John Hallam. "Ranking vs. Preference: A Comparative Study of Self-Reporting." In Affective Computing and Intelligent Interaction, edited by Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, 437–46. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011. https://doi.org/10.1007/978-3-642-24600-5_47.

Youmans, Robert J., and Thomaz Arciszewski. "Design Fixation: Classifications and Modern Methods of Prevention." AI EDAM 28, no. 2 (May 2014): 129–37. <https://doi.org/10.1017/S0890060414000043>.

Yu, Luwen, Stephen Westland, and Zhenhong Li. "Analysis of Experiments to Determine Individual Colour Preference." Color Research and Application 46, no. 1 (2021): 155–67. <https://doi.org/10.1002/col.22589>.

Yuan, Zheng, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. "RRHF: Rank Responses to Align Language Models with Human Feedback without Tears," 2023. <https://doi.org/10.48550/ARXIV.2304.05302>.

Zhang, Tingting, Harold T. Nefs, Judith Redi, and Ingrid Heynderickx. "The Aesthetic Appeal of Depth of Field in Photographs." In 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), 81–86. Singapore, Singapore: IEEE, 2014. <https://doi.org/10.1109/QoMEX.2014.6982300>.

B

Exploring aesthetic experiences from a philosophical perspective

[Hyperlink to Section 2.1.](#)

B.0.1. A brief overview of philosophical viewpoints

Aesthetics is a subject written about for centuries. The term aesthetics, derived from the Greek 'aesthesis', implies sensuous knowledge, or sensory perception and understanding (Hekkert, 2006). Baumgarten, a German philosopher of the 18th century contributed much to how we look at aesthetics today. In his book *Aesthetica*, published in 1750, he is the first to describe this new discipline. According to Baumgarten, art and beauty are not merely rational, and he describes this as sensuous delight, or aesthetic pleasure (Goldman, 2001) (Guyer, 2020).

Like Baumgarten, Schopenhauer is a German philosopher. He thinks the aesthetic experience is a key aspect of our human existence. His take on aesthetics is still relevant, and arguably the most lasting part of his philosophy (Knox, 1958). Schopenhauer speaks of aesthetic contemplation, in which he differentiates three different variations.

- '*The beautiful*' is embodied in an aesthetic experience triggered by objects that are sensory pleasing, and satisfy our desire for harmony and order. An example he gives of such a trigger is a flower (Shapshay, 2021), which you can appreciate for its beauty *an sich*. You do not need anything from the flower, and can appreciate it for the beauty it possesses without ulterior motives. Schopenhauer here describes, is more commonly known as *aesthetic disinterestedness*, a brainchild of Kant (Li, 2018).
- Next to the beautiful, he outlines '*the stimulating*'. If an aesthetic trigger is stimulating it can put us into a state of aesthetic contemplation by attracting our attention and engaging our senses in a pleasurable way (Shapshay, 2021). Stimulating triggers can include food or sexual arousal.
- The third form of aesthetic contemplation that the philosopher recognizes is '*the sublime*'. A trigger of the sublime can be described as overwhelming, and can even induce fear, or a feeling of transcendence. A sublime experience can be awe-inspiring, and so powerful that they evoke a sense of reverence in the beholder. Being in a sublime aesthetic contemplating state can result in a sense of elevation (Shapshay, 2021). Triggers of a sublime experience can be volatile landscapes or powerful oceans.

Where Schopenhauer's theory of contemplation takes a more rational perspective on aesthetic experience, Hume centralizes emotion. He argues that our feelings, rather than our thoughts, determine whether we find something beautiful or not. He further reasons that when we make an aesthetic appraisal we are saying something about how the trigger affects our emotions. He argues that our emotions and aesthetic appraisals influence how we make decisions and take action (Gracyk, 2021).

In his book, *The Structure of Behavior* describes Merleau-Ponty how the world around us cannot be regarded as merely a collection of objective facts which can be described through scientific research. According to him, the world is a complex and dynamic system in which objects relate to each other at different levels of organization (Toadvine, 2019). In the book *Phenomenology of Perception*, he elaborates on how people relate to this complex world described above. He emphasizes the important role of the body in perceiving the world around us. Our aesthetic perception also falls under this. To elaborate on this: Merleau-Ponty reasons that perception is a reciprocal interaction between our body and an object. According to Merleau-Ponty, our senses are not passive receivers of information. Rather, he considers them active participants in the process of perception. They participate by adapting and adjusting themselves to best perceive the relevant triggers. This implies that we humans are not passive observers in relation to the world around us, but active participants who are intimately connected to it. Our bodies play a central role in this, they shape our perception of the world around us. And our aesthetic experiences originate from this active engagement with the world around us.

Simpson (1975) describes that aesthetic appraisal is subjective and context-dependent. He further reasons that aesthetic judgments are not universal, but depend on personal preferences and cultural norms. Your aesthetic preferences are influenced by both your cultural background and personal experiences. This can cause you to make different judgments than others. An example of this is that I can really enjoy a Rothko painting because my parents used to bring me to a lot of modern museums, whereas a friend of mine who grew up in China might make very different judgments about his paintings.

B.0.2. Everyday aesthetics

Chatterjee (2003) and Vartanian and Skov (2014) are both good examples of the contemporary focus on fine arts as an aesthetic trigger. However, any (in)tangible trigger can evoke an aesthetic experience. The following statement encapsulates this well:

“There is virtually no limit to what can become a source of aesthetic experience.” Saito (2021)

Mandoki (2007) and Saito (2010) expound on this rationale in everyday aesthetics, attempting to broaden the scope of the Western focus of aesthetics on fine arts. They do so by incorporating everyday objects and activities. Mandoki and Saito are returning aesthetics to their original meaning: sensory experience. An aesthetic experience can have different intensities: breathtaking scenery, but also disturbing news or a sensual image can make a very strong aesthetic impact. But even boring, unremarkable or plain triggers can still make a weak aesthetic impact and thus fall under everyday aesthetics (Saito, 2021).

Negative aesthetics should not be disregarded. Saito (2021) describes how many unappealing aspects affect your quality of life. The theory of negative aesthetics emphasizes the importance of considering these as well, in contrast to the current focus on positive aesthetics. Saito argues that the negative qualities of these experiences are essential tools to ultimately establish a positive aesthetic experience. Berleant (2012) and Saito (2021) describe the activist dimension of negative aesthetics. They expand on how negative aesthetic experiences can help us identify elements that are detrimental to our quality of life, and act as catalysts for change.

Everyday aesthetics requires a richer variety of aesthetic qualities. Besides beauty and sublimity, for example (as proposed by Schopenhauer), qualities such as *“messiness, shabbiness and cuteness”* (Saito, 2021) also deserve consideration. To appreciate the aesthetics of everyday triggers one does not need the sensibility that does play a role in appreciating art.

B.0.3. Exploring aesthetic attitude, aesthetic sensibility, and aesthetic fluency

King (2012) describes aesthetic attitude as a way of approaching experiences. It is a state of mind, a way of relating and orienting oneself in the world around us. When you adopt an aesthetic attitude you focus on the features of a trigger that are relevant to aesthetic appraisal. For some people, it is easier to adopt this attitude than for others (similar to adopting an optimistic attitude).

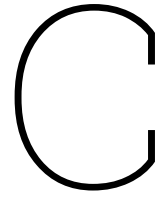
Aesthetic sensibility describes the ability to perceive the aesthetic qualities of a trigger. You can think of it as perceptual awareness, which is based on heightened aesthetic awareness. This ability is often shaped by past experiences (Berleant, 2015).

Related to this is the notion of aesthetic fluency. This is one's ability to understand and appreciate different forms of aesthetic expression. It relies on the understanding of composition, materials, and colours, among other aspects. So part of aesthetic fluency is that someone has knowledge of different movements in art, literature, and music and that they can put into words why they find some of these genres appealing. It is developed through exposure to cultural and artistic experiences, and education and practice (Reber et al., 2004).

Mcmanus and Furnham (2006) propose a way to measure aesthetic attitude, and Reber et al. (2004) found a way to assess aesthetic fluency. Since the two notions affect the aesthetic experience, it is important to see if and how they influence how participants annotate images. To my knowledge, aesthetic sensibility has so far been a purely philosophical concept, which has the consequence that there are no metrics to determine it. Since it concerns heightened aesthetic awareness, this notion is partially described by aesthetic attitude and aesthetic fluency. I expect that sensibility will be taken into account as a result, even if we do not explicitly ask for it.

B.0.4. Controversy around aesthetic emotions

Above, aesthetic emotions has been mentioned by various thinkers. However, the topic of aesthetic emotions is a controversial one in the world of aesthetics. Several papers argue that aesthetic pleasure is not an emotion, since an aesthetic experience is derived only from the sensory perception of the stimulus, and does not involve our personal concerns or emotions. An aesthetic response is disinterested, since it only focuses on perceiving the object itself without any ulterior motives. However, this does not mean that an aesthetic experience cannot evoke emotions. These only arise after the aesthetic experience (Hekkert and Leder, 2008; Blijlevens, Thurgood and Hekkert, 2017;). Next to this, Skov and Nadal (2020) comment on work on aesthetic emotions by Menninghaus and colleagues (2019). Menninghaus et al. (2019) claim that aesthetic emotions differ from regular emotions because they include aesthetic evaluation, they rely on specific aesthetic values, and they involve pleasure and displeasure, among other things. The paper by Skov and Nadal refutes the rationale of Menninghaus et al. with empirical evidence. They demonstrate that affective states observed during aesthetic appreciation are not distinctly different from affective states that can be observed during other sensory evaluations.



Exploring aesthetic experiences from a neuroscientific perspective

[Hyperlink to Section 2.1.](#)

C.0.1. The neuroscientific basis of the aforementioned philosophical theories

Chatterjee (2002), highlights the subject from a neuroscientific angle. He concludes that an aesthetic experience is derived from reactions to various components of a visual stimulus. This aligns with Merleau-Ponty's perspective, which emphasizes the importance of the reciprocal interaction between our body and an object. Chatterjee continues that the process by which people respond to stimuli and activate neural circuits that respond to rewarding stimuli (e.g. pleasure) may be a probe into the neural foundation for "liking without wanting", which infers "the beautiful" aesthetic contemplation described by Schopenhauer (or: aesthetic disinterestedness).

Cela-Conde et al. (2011) found that aesthetic experiences are strongly influenced by our affective states, our emotions. This is in line with Hume's philosophy, where aesthetic experiences are shaped by our emotional response to stimuli.

An example of this is provided by Nanda et al. (2013), where the researchers conclude that certain shapes can evoke certain emotions.

Chatterjee (2003) differentiates five stages of information processing involved in visual aesthetic preference. First, as with other objects, the basic visual properties of a work of art are processed in the primary and secondary visual brain regions. After this, attention processes in the frontal-parietal networks are employed to recognize prominent visual information such as shape, colour, and composition. Hereafter, properties of the visual stimulus, the artwork, are grouped and recognized (e.g., a woman, a dress) in the temporal lobe.

Once the visual stimulus is identified, the trigger will evoke an experience in you, through feedback and feedforward processes that connect attribution and attention circuits. Finally, Chatterjee (2003) describes that the parts of the brain related to emotions play a role in the aesthetic experience.

Vartanian and Skov (2014) find similar results when they expose participants to paintings while in an fMRI machine. These researchers find that while viewing the paintings not only the brain regions responsible for visual perception and object identification are activated in their participants, but also areas that we relate to emotions and personal thoughts. These findings combine Hume's view: that an aesthetic experience is an emotional one, and Schopenhauer's view, that an aesthetic experience puts you in a contemplative state.

Although critique can be made on the approach to link philosophical theories that are inherently

not empirical to neurobiological studies, it provides a deeper understanding of the workings of aesthetic experiences.

The literature described above bridges philosophical theories and the underlying neurological processes involved in having an aesthetic experience. These empirical outcomes can serve as evidence for the connection between philosophical theories and neural mechanisms.

D

An overview of empirical aesthetics literature

[Hyperlink to Section 2.2.](#)

An overview of existing literature on empirical aesthetics is presented on page number 80 and 81.

Authors + date	Title	Field	Determinants or 'as such'??*	Scales	Method	What was measured?	Objective or subjective aesthetics?
Hassenzahl and Monk, 2010	The Inference of Perceived Usability From Beauty	HCI	Measures determinants of aesthetics	Pragmatic value, beauty, goodness, hedonic value -> captivating, stylish, premium, and creative. **	Scales	Websites	Subjective
Moshagen and Thielsch, 2010	Facets of visual aesthetics	HCI	Determinants (Simplicity and Diversity, colour, craftsmanship)	The layout is too dense, the colors are attractive, the layout is pleasantly varied	Scales	Websites	Objective
Lavie and Tractinsky, 2004	Assessing dimensions of perceived visual aesthetics of web sites	HCI	Determinants: classical aesthetics and expressive aesthetics	Pleasing, sophisticated, symmetrical, and modern	Scales	Websites	Subjective and objective
Augustin et al., 2011	Measuring aesthetic impressions of visual art	Art	Measures determinants of aesthetics	Attractive, beautiful, fascinating, happy, incomprehensible, innovative, ordinary, original, overwhelming and warm	Scales	Visual art	Subjective
Blijlevens et al., 2017	The Aesthetic Pleasure in Design Scale: The development of a scale to measure aesthetic pleasure for designed artifacts.	Design	Measures aesthetics	Beautiful, attractive, pleasing to see, nice to see, and like to look at.			Subjective

Hung and Chen, 2012	Effects of Novelty and Its Dimensions on Aesthetic Preference in Product Design	Design	Influence of dimensions of product semantics on novelty and aesthetic pref.	Novelty: typical-unique aesthetic preference: beautiful-ugly	Scales	Drawings/images of 88 chairs	Subjective
Sonderegger and Sauer, 2010	The influence of design aesthetics in usability testing: Effects on user performance and perceived usability	Design	Measures aesthetics	“The design of the mobile phone is very appealing”	Likert scale	Computer simulation of a mobile phone	Subjective
Page and Herr, 2002	An Investigation of the Processes by Which Product Design and Brand Strength Interact to Determine Initial Affect and Quality Judgments	Design	Measures aesthetics	Highly unattractive - highly attractive	11 point scale	Pictures of laptops	Subjective

Berghman and Hekkert, 2017	Towards a unified model of aesthetic pleasure in design	Design	Measures aesthetics	<p>Aesthetic appreciation:</p> <ul style="list-style-type: none"> - This product is pleasing to see. - The design of this product is beautiful. - This product has an attractive design. <p>Unity:</p> <ul style="list-style-type: none"> - The product is unified. - The product is coherent. <p>Variety:</p> <ul style="list-style-type: none"> - The product conveys variety. - The product is rich in elements. <p>Typicality:</p> <ul style="list-style-type: none"> - The design is typical for this kind of product. - This is a standard design for this type of product. <p>Novelty:</p> <ul style="list-style-type: none"> - This product is original. - The design of this product is novel. <p>Connectedness:</p> <ul style="list-style-type: none"> - The design of this product makes me feel connected to people like me. - The design of this product shows that I am similar to people like me. <p>Autonomy:</p> <ul style="list-style-type: none"> - This product design helps me to be unique in reference to people like me. - The design of this product helps me to distinguish myself from others. 	7 point scale	Images of: bicycles, sunglasses, dining tables, espresso makers, table lamps	Subjective
----------------------------	---	--------	---------------------	---	---------------	--	------------

E

Existing visual datasets claiming to reflect aesthetics

[Hyperlink to Section 2.2.](#)

An overview of existing literature on empirical aesthetics is presented on page number 82, 83 and 84.

Name of dataset	Researchers+date	Paper name	#images	How filtered on aesthetics?	'Why' did they do what they did?	How was the crowdsourced output computed?	control variables	Obj./subj.?
EVA: Explainable Visual Aesthetics	Kang et al. 2020	EVA: An Explainable Visual Aesthetics Dataset	4070	<p>Scales:</p> <p>"What is the overall aesthetic quality of this picture?"</p> <p>least beautiful - most beautiful</p> <p>"How difficult is it for you to judge this image's aesthetic quality?"</p> <p>Likert: very difficult - difficult - easy - very easy</p> <p>Getting crowdsourcers to measure objective aesthetic qualities</p> <p>"How do you like this attribute?"</p> <p>where we consider four attributes: light and colour, composition and depth</p>	<p>"Considering the attributes in previous work [1, 8, 10, 11, 13] and inspired by methods for subjective quality assessment experiments in laboratory conditions [11], we design the survey considering four main attributes and one measure of the difficulty to judge image aesthetic quality."</p>	<p>"Finally, by averaging and normalizing the binary votes over attributes, we can get a continuous, per-image probability distribution of importance weights."</p> <p>-> all images in the dataset are annotated (no computation)</p>	<p>year of birth, region, gender, and whether they are colour blind or wearing glasses.</p> <p>Self-assessment: their experience in photography, as either beginner (without any specific knowledge about photography); intermediate (a casual photographer without specific training); or advanced (having followed some specific training in photography).</p>	Subj. and obj.
AVA: aesthetic visual analysis	Murray et al. 2012	AVA: A large-scale database for aesthetic visual analysis	250,000	<p>Upvotes</p> <p>This dataset is derived from an online community where (amateur) photographers upload their images, and score them (the aesthetic rating that is used to develop AVA). The photos are uploaded as 'challenges' (e.g. make the sky the subject of your photo this week).</p>	N/A.	<p>Analysis of score distributions: Gaussian functions are adequate for modelling score distributions of images with mean scores between 2-8, which make up 99.77% of the dataset.</p>	N/A.	Subj. <input type="checkbox"/>

TAD66K: Theme and Aesthetics Dataset with 66K images	He et al. 2022	Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks	66,327	range 1-10: lowest aesthetics - highest aesthetics, Problem with data annotation: participants have no baseline, so they do not know how to add value to something. This research provided anchor images for different rating scales	N/A.	The third component performs two functions. First, APNet directly extracts aesthetic features x_{aes} from the input x ; we use MobileNetV2 as the backbone, and the output is processed by L5. Second, three features are fused to predict an aesthetic score, and the output is processed by L6. We describe the whole process as: $p = \text{Faes}(x_{theme} \oplus x_{rgb} \oplus x_{aes}, \theta_{aes})$, where θ_{aes} represents all the parameters of Faes.	N/A.	Subj.
FLICKR-AES	Ren et al. 2017	Personalized Image Aesthetics	40,000	Aesthetics score determined by asking MTurkers to rate images 1-5 from lowest to highest aesthetics level.	N/A.	"we train a deep neural network to predict genetic aesthetic scores. With the generic aesthetic scores, we can compute residual scores (offsets) for the example images by subtracting them from ratings by each user. Our goal is then reduced to learn a regressor to predict the residual score given any new image. Due to the lack of annotated examples from each user, training such regressor directly from an image is not practical. Therefore, we propose to use high-level image attributes related to image aesthetics to form a compact feature representation for residual learning."	Context factors	Subj.

DP Challenge	Datta et al. 2008	Algorithmic inferring of aesthetics and emotion in natural images: An exposition	16,509	The aesthetic evaluation is based on peer-rating of overall quality of the images on a scale of 1-10.	N/A.	"When a photograph is rated by a set of n people on a 1 to D scale on the basis of its aesthetics, the average score can be thought of as an estimator for its intrinsic aesthetic quality. More specifically, we assume that an image I has associated with it a true aesthetics measure $q(I)$, which is the asymptotic average if the entire population rated it. The average over the size n sample of ratings, given by $\hat{q}(I) = \frac{1}{n} \sum_{i=1}^n r_i(I)$ is an estimator for the population parameter $q(I)$, where $r_i(I)$ is the i th rating given to image I . Intuitively, a larger n gives a better estimate. A formulation for aesthetics score prediction is therefore to infer the value of $q(I)$ by analyzing the content of image I , which is a direct emulation of humans in the photo rating process."	N/A.	Subj.
PCCD: Photo Critique Captioning Dataset	Chang et al. 2017	Aesthetic Critiques Generation for Photos	4,235	The dataset contains pairwise data of images and sentences, where an image could have multiple sentences related to different aspects of aesthetics. Each image is attached with comments and scores of 7 aesthetic attributes: colour lighting, composition, depth of field, focus, general impression, subject of	This dataset focusses on aesthetic images and corresponding text	The approach combines individual aspect-oriented captioning systems and aspects predictor to create captions that target different aspects of the aesthetics of an image.	N/A.	Obj.

DPC-Captions	Jin et al. 2019	Aesthetic Attributes Assessment of Images	154,384	5 aesthetic attributes tested on a scale: color and lighting, composition, depth and focus, impression and subject, use of camera.	unclear	unclear	N/A.	Obj.
AADB: Aesthetics with Attributes Database	Kong et al. 2016	Photo Aesthetics Ranking Network with Attributes and Content Adaptation	10,000	11 aesthetic attributes, rated on a scale of 1-5: interesting content, object emphasis, good lighting, color harmony, vivid color, shallow depth of field, motion blur, rule of thirds, balancing element, repetition, and symmetry.	N/A.	"For each image, we average the ratings of five raters as the ground-truth aesthetic score."	N/A.	Obj.
CUHK-PQ	Luo et al. 2011	Content-Based Photo Quality Assessment	17,673	Binary labels: high quality/low quality photos	N/A.	A photo is classified as high or low quality only if eight out of the ten viewers agree on its assessment. The procedure is content-based quality assessment of images, so the computation focus is on content.	N/A.	Subj.
Photo.net	Joshi et al., 2011	Aesthetics and emotions in images	20,278	The images are peer rated on a one to seven scale of aesthetics. Images and annotations are derived from the photography website photo.net	N/A.	Images are directly derived from a website where the images were already annotated, not much is computed with this data.	N/A.	Subj.

KonIQ-10k	Hosu et al. 2019	KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment	10,073	"The subjects were then instructed to consider the following types of noise, JPEG artifacts, aliasing, lens and motion blur, over-sharpening, wrong exposure, color fringing, and over-saturation. We used the standard 5-point Absolute Category Rating (ACR) scale, i.e. bad (1), poor (2), fair (3), good (4), and excellent (5)."	N/A.	Here, the MOS is also implemented. In this experiment, 11 expert MOS values are compared to the MOS value of the crowd. "We compensated for the difference in the range of the MOS between the two data sources by fitting a linear model: $MOS_{experts} = 1.12 \cdot MOS_{crowd} - 10.43$. Note that the alignment of the crowd to the expert MOS scale is not applied to transform the final scores that are part of the database. After the re-alignment, we compared the two data sources: experts and crowd. For the crowd data, we bootstrapped MOS computations for subsamples of size of up to 120 – the number of ratings in KonIQ-10k. Each set of MOS values was compared to the ground truth MOS, giving rise to a root-mean-square error (RMSE), see Fig. 5(a). We found that this RMSE quickly converges to a lower bound of 11.35 on the 100 point scale. We also bootstrapped expert groups of size 11 and found a standard deviation of bootstrapped MOS values of 6.63."	N/A.	Obj.
-----------	------------------	---	--------	---	------	---	------	------

SPAQ: Smartphone Photography Attribute and Quality	Fang et al., 2020	Perceptual Quality Assessment of Smartphone Photography	11,125	EXIF tags: 1) focal length, 2) f-number (inversely proportional to aperture size), 3) exposure time, 4) ISO (light sensitivity of sensor), 5) brightness value (brightness of focus point in the scene), 6) flash (flash fired or not), 7) time (when image was recorded). MOS (Mean Opinion Score): a continuous score in [0, 100] to represent the overall quality of the image. Image attribute scores, including 1) brightness, 2) colorfulness, 3) contrast, 4) noisiness, and 5) sharpness.	N/A.	the 'aesthetic' measurement here is MOS: Mean Opinion Score. Meaning that they to the mean of all scores of an image as aesthetic value.	N/A.	Obj. and subj.
--	-------------------	---	--------	---	------	--	------	----------------

F

HREC approved consent

[Hyperlink to Section 2.8.](#)

F.1. HREC submission

F.2. HREC Revisions

Delft University of Technology
HUMAN RESEARCH ETHICS
CHECKLIST FOR HUMAN RESEARCH
(Version January 2022)

IMPORTANT NOTES ON PREPARING THIS CHECKLIST

1. An HREC application should be submitted for every research study that involves human participants (as Research Subjects) carried out by TU Delft researchers
2. Your HREC application should be submitted and approved **before** potential participants are approached to take part in your study
3. All submissions from Master's Students for their research thesis need approval from the relevant Responsible Researcher
4. The Responsible Researcher must indicate their approval of the completeness and quality of the submission by signing and dating this form OR by providing approval to the corresponding researcher via email (included as a PDF with the full HREC submission)
5. There are various aspects of human research compliance which fall outside of the remit of the HREC, but which must be in place to obtain HREC approval. These often require input from internal or external experts such as [Faculty Data Stewards](#), [Faculty HSE advisors](#), the [TU Delft Privacy Team](#) or external [Medical research partners](#).
6. You can find detailed guidance on completing your HREC application [here](#)
7. Please note that incomplete submissions (whether in terms of documentation or the information provided therein) will be returned for completion **prior to any assessment**
8. If you have any feedback on any aspect of the HREC approval tools and/or process you can leave your comments [here](#)

I.

I. Applicant Information

PROJECT TITLE:	Developing a process for annotation of images on the basis of aesthetics with the purpose of training generative models
Research period: <i>Over what period of time will this specific part of the research take place</i>	13/02/'23 - 07/07/'23
Faculty:	IDE
Department:	HCD
Type of the research project: <i>(Bachelor's, Master's, Dream Team, PhD, PostDoc, Senior Researcher, Organisational etc.)</i>	Master's thesis
Funder of research: <i>(EU, NWO, TUD, other – in which case please elaborate)</i>	-
Name of Corresponding Researcher: <i>(If different from the Responsible Researcher)</i>	Céline Offerman
E-mail Corresponding Researcher: <i>(If different from the Responsible Researcher)</i>	celineofferman@live.nl
Position of Corresponding Researcher: <i>(Masters, Dream Team, PhD, PostDoc, Assistant/ Associate/ Full Professor)</i>	Masters
Name of Responsible Researcher: <i>Note: all student work must have a named Responsible Researcher to approve, sign and submit this application</i>	Willem van der Maden
E-mail of Responsible Researcher: <i>Please ensure that an institutional email address (no Gmail, Yahoo, etc.) is used for all project documentation/ communications including Informed Consent materials</i>	W.L.A.vanderMaden@tudelft.nl
Position of Responsible Researcher : <i>(PhD, PostDoc, Associate/ Assistant/ Full Professor)</i>	PhD

II. Research Overview

NOTE: You can find more guidance on completing this checklist [here](#)

a) Please summarise your research very briefly (100-200 words)

What are you looking into, who is involved, how many participants there will be, how they will be recruited and what are they expected to do?

Add your text here – (please avoid jargon and abbreviations)

We will look into how crowdsourced participants can be used in the evaluation of aesthetics of datasets. The amount of participants is to be determined, they will be recruited via mechanical turk, they will be expected to evaluate images on the basis of their aesthetics.

b) If your application is an additional project related to an existing approved HREC submission, please provide a brief explanation including the existing relevant HREC submission number/s.

Add your text here – (please avoid jargon and abbreviations)

-

- c) If your application is a simple extension of, or amendment to, an existing approved HREC submission, you can simply submit an [HREC Amendment Form](#) as a submission through LabServant.

III. Risk Assessment and Mitigation Plan

NOTE: You can find more guidance on completing this checklist [here](#).

Please complete the following table in full for all points to which your answer is "yes". Bear in mind that the vast majority of projects involving human participants as Research Subjects also involve the collection of **Personally Identifiable Information (PII)** and/or **Personally Identifiable Research Data (PIRD)** which may pose potential risks to participants as detailed in Section G: Data Processing and Privacy below.

To ensure alignment between your risk assessment, data management and what you agree with your Research Subjects you can use the last two columns in the table below to refer to specific points in your Data Management Plan (DMP) and Informed Consent Form (ICF) – **but this is not compulsory**.

It's worth noting that **you're much more likely to need to resubmit your application if you neglect to identify potential risks**, than if you identify a potential risk and demonstrate how you will mitigate it. If necessary, the HREC will always work with you and colleagues in the Privacy Team and Data Management Services to see how, if at all possible, your research can be conducted.

ISSUE	If YES please complete the Risk Assessment and Mitigation Plan columns below.		Please provide the relevant reference #			
	Yes	No	RISK ASSESSMENT – what risks could arise? <i>Please ensure that you list ALL of the actual risks that could potentially arise – do not simply state whether you consider any such risks are important!</i>	MITIGATION PLAN – what mitigating steps will you take? <i>Please ensure that you summarise what actual mitigation measures you will take for each potential risk identified – do not simply state that you will comply with regulations.</i>	DMP	ICF
A: Partners and collaboration						
1. Will the research be carried out in collaboration with additional organisational partners such as: <ul style="list-style-type: none"> One or more collaborating research and/or commercial organisations Either a research, or a work experience internship provider? <i>If yes, please include the contractual agreement in this application.</i>	✓					
2. Is this research dependent on a Data Transfer or Processing Agreement with a collaborating partner or third party supplier? <i>If yes, please provide a copy of the signed DPA form.</i>	✓					
3. Has this research been approved by another (external) research ethics committee (e.g.: HREC and/or MREC/METIC)? <i>If yes, please provide a copy of the approval (if possible) and summarise any key points in your Risk Management section below.</i>	✓					
B: Location						

		If YES please complete the Risk Assessment and Mitigation Plan columns below.			Please provide the relevant reference #	
ISSUE	Yes	No	RISK ASSESSMENT – what risks could arise? Please ensure that you list ALL of the actual risks that could potentially arise – do not simply state whether you consider any such risks are important!	MITIGATION PLAN – what mitigating steps will you take? Please ensure that you summarise what actual mitigation measures you will take for each potential risk identified – do not simply state that you will e.g. comply with regulations.	DMP	ICF
4. Will the research take place in a country or countries, other than the Netherlands, within the EU?	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
5. Will the research take place in a country or countries outside the EU?	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
6. Will the research take place in a place/region or of higher risk – including known dangerous locations (in any country) or locations with non-democratic regimes?	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
C: Participants						
7. Will the study involve participants who may be vulnerable and possibly (legally) unable to give informed consent? (e.g., children below the legal age for giving consent, people with learning difficulties, people living in care or nursing homes).	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
8. Will the study involve participants who may be vulnerable under specific circumstances and in specific contexts, such as victims and witnesses of violence, including domestic violence; sex workers; members of minority groups, refugees, irregular migrants or dissidents?	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
9. Are the participants, outside the context of the research, in a dependent or subordinate position to the investigator (such as own children, own students or employees of either TU Delft and/or a collaborating partner organisation)? It is essential that you safeguard against possible adverse consequences of this situation (such as allowing a student's failure to participate to your satisfaction to affect your evaluation of their coursework).	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
10. Is there a high possibility of re-identification for your participants? (e.g., do they have a very specific job which there are only a small number in a given country, are they members of a small community, or employees from a partner company collaborating in the research? Or are they one of only a handful of (expert) participants in the study?)	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
D: Recruiting Participants						
11. Will your participants be recruited through your own, professional, channels such as conference attendance lists, or through specific network/s such as self-help groups.	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
12. Will the participants be recruited or accessed in the longer term by a (legal or customer) (pink)eeper? (e.g., an adult professional working with children, a community leader or family member who has this customary role – within or outside the EU, the data producer of a long-term cohort study).	<input checked="" type="checkbox"/>	<input type="checkbox"/>				

		If YES please complete the Risk Assessment and Mitigation Plan columns below.			Please provide the relevant reference #	
ISSUE	Yes	No	RISK ASSESSMENT – what risks could arise? Please ensure that you list ALL of the actual risks that could potentially arise – do not simply state whether you consider any such risks are important!	MITIGATION PLAN – what mitigating steps will you take? Please ensure that you summarise what actual mitigation measures you will take for each potential risk identified – do not simply state that you will e.g. comply with regulations.	DMP	ICF
13. Will you be recruiting your participants through a crowd-sourcing service and/or involve a third party data-gathering service, such as a survey platform?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Risks could be that the participants will not answer the questions truthfully to save their own time. Another risk could be that a participant quits halfway while filling in the survey.	Exclude all participants who only ticked one box throughout the whole survey, exclude participants who did not fill in all the answers.		
14. Will you be offering any financial, or other, remuneration to participants, and might this induce or bias participation?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	We will offer a financial reward, but it will not be dependent on their answers. This will not influence their ability to provide answers.			
E: Subject Matter Research related to medical questions/health may require special attention. See also the website of the CCMO before contacting the WVTC .						
15. Will your research involve any of the following: • Medical research and/or clinical trials • Invasive sampling and/or medical imaging • Medical and in Vitro Diagnostic Medical Devices Research	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
16. Will drugs, placebos, or other substances (e.g., drinks, foods, food or drink constituents, dietary supplements) be administered to the study participants? If yes see here to determine whether medical ethical approval is required	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
17. Will blood or tissue samples be obtained from participants? If yes see here to determine whether medical ethical approval is required	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
18. Does the study risk causing psychological stress or anxiety beyond that normally encountered by the participants in their life outside research?	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
19. Will the study involve discussion of personal sensitive data which could put participants at increased legal, financial, reputational, security or other risk? (e.g. financial data, location data, data relating to children or other vulnerable groups) Definitions of sensitive personal data, and special cases are provided on the TU Delft Privacy team website.	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
20. Will the study involve disclosing commercially or professionally sensitive, or confidential information? (e.g., relating to decision-making processes or business strategies which might, for example, be of interest to competitors)	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
21. Has your study been identified by the TU Delft Privacy Team as requiring a Data Processing Impact Assessment (DPIA)? If yes please attach the advice/approval from the Privacy Team to the application.	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
22. Does your research investigate causes or areas of conflict?	<input checked="" type="checkbox"/>	<input type="checkbox"/>				

		If YES please complete the Risk Assessment and Mitigation Plan columns below.			Please provide the relevant reference #	
ISSUE	Yes	No	RISK ASSESSMENT – what risks could arise? Please ensure that you list ALL of the actual risks that could potentially arise – do not simply state whether you consider any such risks are important!	MITIGATION PLAN – what mitigating steps will you take? Please ensure that you summarise what actual mitigation measures you will take for each potential risk identified – do not simply state that you will e.g. comply with regulations.	DMP	ICF
<i>If yes please confirm that your fieldwork has been discussed with the appropriate safety/security advisors and approved by your Department/Faculty.</i>						
23. Does your research involve observing illegal activities or data processed or provided by authorities responsible for preventing, investigating, detecting or prosecuting criminal offences? If so please confirm that your work has been discussed with the appropriate legal advisors and approved by your Department/Faculty.	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
F: Research Methods						
24. Will it be necessary for participants to take part in the study without their knowledge and consent at the time? (e.g., covert observation of people in non-public places).	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
25. Will the study involve actively deceiving the participants? (For example, will participants be deliberately falsely informed, will information be withheld from them or will they be misled in such a way that they are likely to object or show unease when debriefed about the study).	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
26. Is pain or more than mild discomfort likely to result from the study? And/or could your research activity cause an accident involving (non-) participants?	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
27. Will the experiment involve the use of devices that are not 'CE' certified? Only if 'yes' continue with the following questions: • Was the device built in-house? • Was it inspected by a safety expert at TU Delft? If yes, please provide a signed device report • If it was not built in-house and not CE-certified, was it inspected by some other, qualified authority in safety and approved? If yes, please provide records of the inspection	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
28. Will your research involve face-to-face encounters with your participants and if so how will you assess and address Covid considerations?	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
29. Will your research involve either: a) 'big data', combined datasets, new data-gathering or new data-merging techniques which might lead to re-identification of your participants and/or b) artificial intelligence or algorithm training where, for example biased datasets could lead to biased outcomes?	<input checked="" type="checkbox"/>	<input type="checkbox"/>				
G: Data Processing and Privacy						

ISSUE			If YES please complete the Risk Assessment and Mitigation Plan columns below.		Please provide the relevant reference #	
	Y	N	RISK ASSESSMENT – what risks could arise? <i>Please ensure that you list ALL of the actual risks that could potentially arise – do not simply state whether you consider any such risks are important!</i>	MITIGATION PLAN – what mitigating steps will you take? <i>Please ensure that you summarise what actual mitigation measures you will take for each potential risk identified – do not simply state that you will e.g. comply with regulations.</i>	DMP	IC
30. Will the research involve collecting, processing and/or storing any directly identifiable PI (Personally Identifiable Information) including name or email address that will be used for administrative purposes only? (eg: obtaining Informed Consent or disbursing remuneration)		✓				
31. Will the research involve collecting, processing and/or storing any directly or indirectly identifiable PIRD (Personally Identifiable Research Data) including videos, pictures, IP address, gender, age etc and what other Personal Research Data (including personal or professional views) will you be collecting?		✓				
32. Will this research involve collecting data from the internet, social media and/or publicly available datasets which have been originally contributed by human participants		✓				
33. Will your research findings be published in one or more forms in the public domain, as e.g., Master's thesis, journal publication, conference presentation or wider public dissemination?	✓		Since no personal data of the participants will be noted, the online publication of the thesis will not be a risk.			
34. Will your research data be archived for re-use and/or teaching in an open, private or semi-open archive?	✓		Since no personal data of the participants will be noted, the online publication of the thesis will not be a risk.			

H: More on Informed Consent and Data Management

NOTE: You can find guidance and templates for preparing your Informed Consent materials) [here](#)

Your research involves human participants as Research Subjects if you are recruiting them or actively involving or influencing, manipulating or directing them in any way in your research activities. This means you must seek informed consent and agree/ implement appropriate safeguards regardless of whether you are collecting any PIRD.

Where you are also collecting PIRD, and using Informed Consent as the legal basis for your research, you need to also make sure that your IC materials are clear on any related risks and the mitigating measures you will take – including through responsible data management.

Got a comment on this checklist or the HREC process? You can leave your comments [here](#)


IV. Signature/s

Please note that by signing this checklist list as the sole, or Responsible, researcher you are providing approval of the completeness and quality of the submission, as well as confirming alignment between GDPR, Data Management and Informed Consent requirements.

Name of Corresponding Researcher (if different from the Responsible Researcher) (print)

Céline Offerman

Signature of Corresponding Researcher:




Date: 13/02/'23

Name of Responsible Researcher (print)

Willem van der Maden

Signature (or upload consent by mail) Responsible Researcher:



Date: 21/02/'23

--

V. Completing your HREC application

Please use the following list to check that you have provided all relevant documentation

Required:

- o **Always:** This completed HREC checklist
- o **Always:** A data management plan (reviewed, where necessary, by a data-steward)
- o **Usually:** A complete Informed Consent form (including Participant Information) and/or Opening Statement (for online consent)

Please also attach any of the following, if relevant to your research:

Document or approval	Contact/s
Full Research Ethics Application	After the assessment of your initial application HREC will let you know if and when you need to submit additional information
Signed, valid Device Report	Your Faculty HSE advisor
Ethics approval from an external Medical Committee	TU Delft Policy Advisor, Medical (Devices) Research
Ethics approval from an external Research Ethics Committee	Please append, if possible, with your submission
Approved Data Transfer or Data Processing Agreement	Your Faculty Data Steward and/or TU Delft Privacy Team
Approved Graduation Agreement	Your Master's thesis supervisor
Data Processing Impact Assessment (DPIA)	TU Delft Privacy Team
Other specific requirement	Please reference/explain in your checklist and append with your submission

Delft University of Technology
HUMAN RESEARCH ETHICS
REVISIONS TEMPLATE
(Version: January 2022)

This revisions template should be used to address queries raised by the Human Research Ethics Committee (HREC) in an ongoing ethics approval and uploaded into your live submission.

If you have any questions about your applying for HREC approval which are not dealt with on the [Research Ethics webpages](#), please contact HREC@tudelft.nl

I. **Response to HREC queries:**

Query 1:

HREC Query	Provide an opening statement and/or Informed Consent form.
Response	<p>Thank you for pointing out that I was still missing a way to check the consent of the participants. Since the study will be an anonymized online survey I will opt for the following opening statement:</p> <p>With this survey I will contribute with the collection of my evaluation of images on the basis of their aesthetics. When clicking through the survey I will answer the various questions concerning this topic truthfully. I can leave this survey anytime I want, with the only implication that I will not receive compensation for the questions I did answer. By clicking through this survey I give permission for my anonymous data to be collected and used for research purposes. I also give permission that this data may be used for scientific publication.</p> <p>O Yes, I agree with this statement, and will continue to the survey. O No, I do not agree with this statement, and will not be filling in the survey.</p>

Query 2:

HREC Query	Provide a data management plan.
Response	<ol style="list-style-type: none"> 1. Data type: The data that will be produced are annotations on existing and publicly accessible photographs. All NSFW content will be filtered out of the dataset before it is presented to participants. The annotations in question will be labels regarding the aesthetics of the photo (e.g. answers to questions such as "Is this image coherent?"). 2. Anonymization: All responses will be anonymized for the rest of the study, ensuring that specific responses cannot be traced back to the participants. 3. Data collection: The data that will be collected is raw data derived from experiments. The data will be obtained from Qualtrics, an online crowd-sourcing tool. 4. Data export and processing: The data will be exported from Qualtrics in the form of an Excel file and later processed with SPSS or a program with the same functionalities and approved by TU Delft. 5. External parties: There are no external parties who have requirements regarding the data.

	<p>6. Privacy and security: The data will be stored securely on TU Delft's servers, with access controls in place to ensure that only authorized personnel can access the data. The servers will be regularly backed up to prevent data loss in the event of a system failure. Any security incidents will be reported to the relevant authorities and to the affected participants.</p> <p>7. Data retention: The generated data will be kept on TU Delft's servers for ten years after the project, in accordance with TU Delft's data retention policies.</p> <p>8. Responsibility: The responsibility for the management of the project will be transferred from the researchers to the person in charge of the TU Delft servers.</p> <p>9. Data sharing: The data may be made available to other researchers upon request, subject to appropriate data use agreements and ethical clearance.</p>
--	---

Query 3:

HREC Query	Please reflect on the expected amount of participants that you will involve, this should be defined before starting your research.
Response	50.

Query 4:

HREC Query	Give more details on what exactly the research project is about.
Response	<p>Developing a process for annotation of images on the basis of aesthetics with the purpose of training generative models:</p> <p>The age of AI is here. Biases are a big problem for this technology (Google AI, n.d.). The mislabeling of data for machine learning models is one of the causes of this problem, which will be the focus of this project. The project will specify on the LAION-5B dataset, and develop a more valuable process to evaluate its aesthetics. Aesthetics is a rich and complex concept, but it is important to keep the labeling simple. To achieve this goal, methods of crowd-sourced labeling will be tested to take into account the unified model of aesthetics developed by Berghman & Hekkert (2017).</p> <p>With 10 million daily users, Stable Diffusion, the image generating-system by Stability AI has a considerable reach (Mehta, 2022). Stable Diffusion is trained on three sub-sets of the LAION-5B data set. These sub-sets are pairs of images and their alt-text, derived from the internet. The three LAION sub-sets are laion2B-en, laion-high-resolution, and laion-aesthetics v2 5+. The evaluation of the 'aesthetics' of the pictures was done by asking participants the following question: "How much do you like this image on a scale from 1 to 10?" (Schuhmann, 2022). The reason the current evaluation was conducted this way is most likely due to its simplicity. This oversimplified question lacks the richness to adequately describe aesthetics. For a data set that is deployed on such a large scale and has a touch point with millions of people a day, it is bad that aesthetics are so superficially reflected. This is where a significant improvement can be made, and that will be the goal of this project.</p> <p>There are several models regarding aesthetics that are universal but relate differently in context. One promising model to apply here is Berghman & Hekkert's (2017) unified model of aesthetics, because it is a profound theory based around three dichotomous facets. Berghman & Hekkert's model relies not on a superficial notion of beauty (or "likeability," as the current dataset reflects),</p>

	<p>but on deeper aesthetic value. As the dataset becomes richer based on aesthetics, it may be a good step in the right direction to reduce biases. These researchers proved the following universal impacts on aesthetics: connectedness - autonomy, typicality - novelty, unity - variety (see figure 2 in this project brief).</p> <p>Connectedness - autonomy: describes how humans feel the need to be part of a group, whilst preserving a sense of individuality. Typicality - novelty: describes our need for the familiar in contrast with our evolutionary want for new. Unity - variety: describes the craving to feel a sense of one whole, where things fit together, and our interest in contrasting elements within this whole</p> <p>Part of the design challenge will be to communicate these abstract and complex principles to the participants.</p> <p>The goal of this research project is to develop a more nuanced procedure on how to annotate images on the basis of aesthetics. A challenge here will be to maintain simplicity, the strength of the current analysis method. Although the existing scales, developed by Berghman and Hekkert (2017) are proven to evaluate aesthetics, the items are complex. I foresee that it will be difficult for an average person to rate a photo on a 'unity' scale. It is therefore important that the new process is still based on simplicity, while accounting for the proven impacts. To accomplish this goal, I will test multiple methods inspired by Berghman & Hekkert's framework to determine how best to analyze aesthetics of a dataset. I will also design an interface to support the participants in understanding the universal aesthetics impacts.</p>
--	--

Query 5:

HREC Query	Clarify the situation concerning the financial compensation of participants (how much? for what? when?).
Response	The appropriate amount of money, from the hourly wage as discussed by Gadiraju et al. (2017) apply that workers should be compensated at USD 7.5 per hour. With the final prototype, we will see exactly how much time the participant needs to perform the tasks and calculate the final compensation.

Query 6:

HREC Query	Note that Mechanical Turk is not on the approved tooling list. Please change the tool to one on the approved list: https://brightspace-support.tudelft.nl/educational-tooling/
Response	Thank you for the feedback. We will opt for using Qualtrics instead.

II. **Signature/s**

Please note that by signing this checklist list as the sole, or Responsible, researcher you are providing approval of the completeness and quality of the submission, as well as confirming alignment between GDPR, Data Management and Informed Consent requirements.


Name of Corresponding Researcher: Céline Offerman

Signature of Corresponding Researcher:


Date:
15/03/2023

Name of Responsible Researcher: Willem van der Maden

Signature (or upload consent by mail) Responsible Researcher:



Date:
15/03/2023



Stimulus set

[Hyperlink to Section 3.2.](#)

G.1. Rationale for the approach

The first experiment was conducted with a different stimulus set. In retrospect, the stimulus set from Experiment 1 can be considered questionable. Therefore, a new approach was taken. **It should be noted that this approach is not strictly scientific but rather opportunistic in nature.**

G.1.1. Goal

I aim to create a stimulus set with a wide range of expected aesthetic scores. To achieve this, I conducted two experiments to identify extreme photos from the dataset. I sought photos that received very low, average, and very high aesthetic scores.

G.1.2. Participants

To accomplish this objective, I opportunistically recruited participants from my personal network, specifically second-year master's design students from TU Delft, both Dutch and international. I chose this particular group based on the assumption that design students, being at an advanced stage in their education, possess a sufficient understanding of aesthetic theory to provide relatively objective evaluations compared to the average individual. Additionally, this group was easily accessible for sampling, utilizing an opportunistic approach. The strength of this analysis lies in the substantial number of design student respondents, allowing us to confidently assert that the photos ultimately included in the experiment represent diverse aesthetic qualities.

G.1.3. Summary of the two experiments

In the first experiment (n=60), I asked participants to categorize the photos into low, medium, and high aesthetic categories. They were also given the option to indicate if a photo should not be included in the stimuli set.

The second experiment (n=12) was conducted to identify photos that received the most extreme ratings. Participants were asked to rate the images on "*How much do you like this image on a scale from 1-10?*", the same question as in Experiment 1.

G.1.4. Analysis

I have reviewed the feedback provided by the participants on the images. Irrelevant criticism refers to comments made by participants indicating that they interpreted the evaluation to be focused on the aesthetics of the building itself, rather than the aesthetics of the image *an sich*. Such criticism has been disregarded. Feedback specifically related to the aesthetics of the images has been taken into account, and these images have been excluded from the analysis.

A comparison of the means for the images from both experiments is presented in the graph. As can be observed, the means per photo show a degree of agreement across the various treatments. The photos selected for the stimuli set are highlighted with orange circles. I emphasized the low and high aesthetics image groups. This choice was made because it is preferable to have a controlled dataset with extremes, as it facilitates our exploratory experiments into potential effects that we are seeking to uncover.

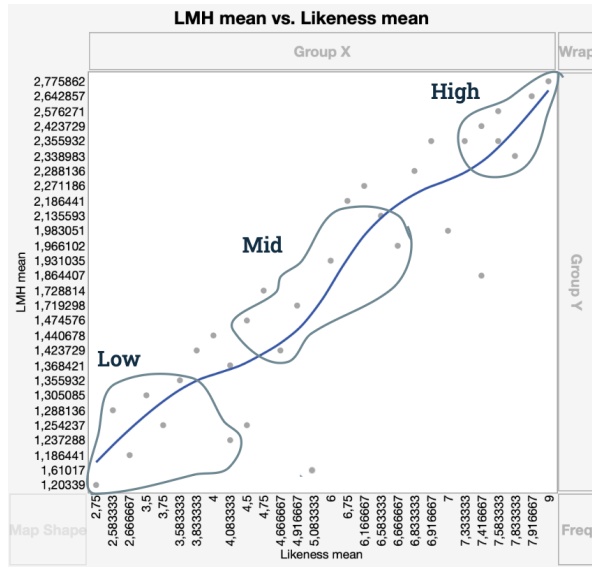


Figure G.1: means of the surveys plotted

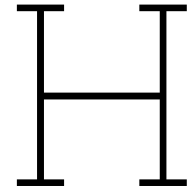
I included an overview of the three brackets in the Excel below, in Figure G.2. The rows are filtered in this overview on the specific brackets. I also included an overview of the same Excel with the rows sorted on likeness means low to high, in Figure G.3. The images that belong in the two groups can be found in Figure 3.3 .

n=12				n=60			
Image nr	Likeness			Image nr	LMH		
	Mean	St. Dev.	Variance		Mean	St.Dev.	Variance
30	2,58333	1,37895	1,90152	30	1,28814	0,49309	0,24313
33	2,66667	1,43548	2,06061	33	1,18644	0,50769	0,25774
3	2,75	1,21543	1,47727	3	1,20339	0,44643	0,1993
27	3,5	2,0226	4,09091	27	1,30509	0,5649	0,31911
26	3,58333	1,67649	2,81061	26	1,35593	0,54969	0,30216
31	3,75	1,60256	2,56818	31	1,25424	0,43917	0,19287
7	4,08333	1,50504	2,26515	7	1,23729	0,50306	0,25307
32	3,83333	2,16725	4,69697	32	1,42373	0,67475	0,45529
37	4	2,2563	5,09091	37	1,44068	0,59513	0,35418
25	4,08333	1,72986	2,99242	25	1,36842	0,6162	0,3797
29	4,5	1,93062	3,72727	29	1,25424	0,54435	0,29632
28	4,75	2,13733	4,56818	28	1,72881	0,69059	0,47691
2	4,5	1,5667	2,45455	2	1,47458	0,568	0,32262
35	4,66667	1,37069	1,87879	35	1,42373	0,67475	0,45529
24	4,91667	2,4293	5,90152	24	1,7193	0,81841	0,6698
11	6	1,53741	2,36364	11	1,93104	0,81353	0,66183
8	6,58333	2,02073	4,08333	8	2,13559	0,65542	0,42957
4	6,66667	1,77525	3,15152	4	1,9661	0,69397	0,48159
38	6,16667	1,52753	2,33333	38	2,27119	0,63871	0,40795
36	5,08333	1,83196	3,35606	36	1,61017	0,69523	0,48334
5	6,75	1,91288	3,65909	5	2,18644	0,65586	0,43016
17	6,83333	1,6967	2,87879	17	2,28814	0,55866	0,3121
18	6,91667	1,72986	2,99242	18	2,35593	0,6634	0,44009
16	7	2,17423	4,72727	16	1,98305	0,7069	0,49971
20	7,41667	1,88093	3,53788	20	1,86441	0,77588	0,60199
9	7,33333	1,557	2,42424	9	2,35593	0,71348	0,50906
6	7,41667	1,37895	1,90152	6	2,42373	0,62155	0,38632
13	7,58333	1,92865	3,7197	13	2,57627	0,56335	0,31736
15	7,58333	1,62135	2,62879	15	2,35593	0,6092	0,37113
12	7,83333	1,40346	1,9697	12	2,33898	0,70979	0,5038
14	7,91667	1,97523	3,90152	14	2,64286	0,48349	0,23377
22	9	1,04447	1,09091	22	2,77586	0,49712	0,24713

Figure G.2: table with rows sorted on the specific aesthetic brackets

n=12				n=60			
Image nr	Likeness			Image nr	LMH		
	Mean	St. Dev.	Variance		Mean	St.Dev.	Variance
30	2,58333	1,37895	1,90151	30	1,288136	0,493085	0,243133
33	2,66667	1,43548	2,06060	33	1,186441	0,507685	0,257744
3	2,75	1,21543	1,47727	3	1,20339	0,446429	0,199299
27	3,5	2,0226	4,09090	27	1,305085	0,5649	0,319112
26	3,58333	1,67648	2,81060	26	1,355932	0,549693	0,302163
31	3,75	1,60255	2,56818	31	1,254237	0,439169	0,19287
32	3,83333	2,16724	4,69697	32	1,423729	0,674751	0,455289
37	4	2,25630	5,09090	37	1,440678	0,595129	0,354179
7	4,08333	1,50504	2,26515	7	1,237288	0,503059	0,253068
25	4,08333	1,72986	2,99242	25	1,368421	0,616197	0,379699
2	4,5	1,56669	2,45456	2	1,474576	0,567995	0,322618
29	4,5	1,93061	3,72727	29	1,254237	0,544351	0,296318
35	4,66667	1,37068	1,87878	35	1,423729	0,674751	0,455289
28	4,75	2,13733	4,56818	28	1,728814	0,69059	0,476914
24	4,91667	2,42930	5,90151	24	1,719298	0,818413	0,6698
36	5,08333	1,83195	3,35606	36	1,61017	0,695229	0,483343
11	6	1,53741	2,36363	11	1,931035	0,813528	0,661827
38	6,16667	1,52752	2,33333	38	2,271186	0,638709	0,407949
8	6,58333	2,02072	4,08333	8	2,135593	0,655419	0,429573
4	6,66667	1,77525	3,15151	4	1,966102	0,693967	0,48159
5	6,75	1,91287	3,65909	5	2,186441	0,655864	0,430158
17	6,83333	1,69669	2,87878	17	2,288136	0,558658	0,312098
18	6,91667	1,72986	2,99242	18	2,355932	0,663395	0,440094
16	7	2,17422	4,72727	16	1,983051	0,7069	0,499708
9	7,33333	1,55698	2,42424	9	2,355932	0,713484	0,509059
6	7,41667	1,37895	1,90151	6	2,423729	0,62155	0,386324
20	7,41667	1,88092	3,53787	20	1,864407	0,775878	0,601987
13	7,58333	1,92865	3,71969	13	2,576271	0,563346	0,317358
15	7,58333	1,62135	2,62878	15	2,355932	0,609203	0,371128
12	7,83333	1,40345	1,96969	12	2,338983	0,709788	0,503799
14	7,91667	1,97522	3,90151	14	2,642857	0,483494	0,233766
22	9	1,04446	1,09090	22	2,775862	0,497118	0,247126

Figure G.3: table with rows sorted on likeness means low to high



Confounding variables

[Hyperlink to Section 4.3.](#)

H.1. Confounding variables included in Experiment 1

Following is an overview of all confounding variables, and the corresponding indication from the literature that they might be of influence in this context.

H.1.1. Demographical information provided by Prolific (Chamorro-Premuzic, Furnham and Reimers, 2007):

Literature:

Chamorro-Premuzic, Furnham and Reimers (2007) state that certain demographics of participants influences their aesthetic experience.

Hypothesis - Age:

There will be a correlation between age and ratings of how much they like images, when controlling for other variables, to examine if age is a potential confounding variable when we ask participants to rate images on how much they like them.

Hypothesis - Sex:

There will be a correlation between sex and ratings of how much they like images, when controlling for other variables, to examine if sex is a potential confounding variable when we ask participants to rate images on how much they like them.

Hypothesis - First language:

There will be a correlation between first language and ratings of how much they like images, when controlling for other variables, to examine if the first language is a potential confounding variable when we ask participants to rate images on how much they like them.

Hypothesis - Current country of residence:

There will be a correlation between the current country of residence and ratings of how much they like images, when controlling for other variables, to examine if the current country of residence is a potential confounding variable when we ask participants to rate images on how much they like them.

Hypothesis - Nationality:

There will be a correlation between nationality and ratings of how much they like images, when controlling for other variables, to examine if nationality is a potential confounding variable when we ask participants to rate images on how much they like them.

Hypothesis - Country of birth:

There will be a correlation between the country of birth and ratings of how much they like images, when controlling for other variables, to examine if the country of birth is a potential confounding variable when we ask participants to rate images on how much they like them.

Hypothesis - Student status:

There will be a correlation between student status and ratings of how much they like images, when controlling for other variables, to examine if student status is a potential confounding variable when we ask participants to rate images on how much they like them.

Hypothesis - Employment status:

There will be a correlation between employment status and ratings of how much they like images, when controlling for other variables, to examine if employment status is a potential confounding variable when we ask participants to rate images on how much they like them.

H.1.2. What is measured by Prolific:

- Age
- Sex
- First language
- Current country of residence
- Nationality
- Country of birth
- Student status
- Employment status

H.1.3. Self-efficacy**Literature:**

As described by Bandura (1977) and Urdan and Pajares (2006), self-efficacy, the confidence someone has that they can successfully complete a task, can ultimately influence performance.

Hypothesis-self-efficacy:

There will be a correlation between declared self-efficacy and ratings of how much they like images, when controlling for other variables, to examine if declared self-efficacy is a potential confounding variable when we ask participants to rate images on how much they like them.

Question:

Developed with the chapter provided by Urdan and Pajares (2006): How confident are you that you can accurately annotate images on the items asked?

Type of question:

Multiple choice

Answers:

Developed with the chapter provided by Urdan and Pajares (2006):

- Not at all confident
- Slightly confident
- Quite confident
- Extremely confident

H.1.4. Variable: Socioeconomic status and education level (Mcmanus and Furnham, 2006)**Literature:**

Mcmanus and Furnham (2006) found that education levels correlate to some extent with some aesthetic attitudes (e.g. they found a positive link between aesthetic attitude 4 (aesthetic relativism) and education level).

Hypothesis-education level:

There will be a correlation between education levels and ratings of how much they like images, when controlling for other variables, to examine if the education level is a potential confounding variable when we ask participants to rate images on how much they like them.

Question:

What level of education do you have?

Type of question:

Dropdown

Answers:

Based on International Standard Classification of Education (ISCED) 2011 (Schneider, 2013):

- Less than primary education
- Primary education
- Lower secondary education
- Upper secondary education
- Post-secondary non-tertiary education
- Short-cycle tertiary education
- Bachelor's or equivalent level
- Master's or equivalent level
- Doctoral or equivalent level
- Not elsewhere classified

H.1.5. Variable: noise levels**Literature:**

Wang et al. (2020) conducted a study on the work experiences of crowd workers in China. They found that some workers noted that they were easily distracted by noises emanating from their direct environment.

Hypothesis-noise levels:

There will be a correlation between noise disturbance while crowd working and ratings of how much they like images, when controlling for other variables, to investigate if noise disturbance acts as a potential confounding variable in participants' ratings of image liking.

Question:

Are you disturbed by noise?

Type of question:

Multiple Choice

Answers:

Based on a question by Nemecek and Grandjean (1973)

- Very much disturbed by noise
- Slightly disturbed by noise
- Not disturbed at all

H.1.6. Variable: working environment**Literature:**

Szubielska et al., 2021 found that the physical environment influences an aesthetic experience.

Wang et al. (2020) found that the work experiences and context of crowdwork and solo workers are substantially different.

Hypothesis-working environment:

There will be a correlation between working environment and ratings of how much they like images, when controlling for other variables, to investigate if working environment acts as a potential confounding variable in participants' ratings of image liking.

Question:

What type of working environment are you in?

Type of question:

Multiple Choice

Answers:

- At home
- In a crowd farm
- In a transportation vehicle (e.g. train)
- In a café or a similar social environment
- In a flex working area
- Other, namely...

H.1.7. Variable: conformity pressure and sense of being watched (Hesslinger et al., 2017)**Literature:**

Hesslinger et al. (2017) concluded from their experiments that social conformity pressure and a sense of being watched both significantly affected the aesthetic judgements of their participants.

Hypothesis-social context:

There will be a correlation between the social context in which the participant crowd works and ratings of how much they like images, when controlling for other variables, to investigate if the social context in which the participant crowd works acts as a potential confounding variable in participants' ratings of image liking.

Question:

In what context are you doing your crowdsourcing work?

Type of question:

multiple choice

Answers:

- Alone
- In a social environment, but working independently
- In a social environment, having social connections while working

H.1.8. Variable: colourblindness (Kang et al., 2020)**Literature:**

Kang et al. (2020) asked participants if they were colour blind or wearing glasses for the development of the Explainable Visual Aesthetics Dataset (EVA).

Hypothesis-colourblindness:

There will be a correlation between colourblindness and ratings of how much they like images, when controlling for other variables, to investigate if colourblindness acts as a potential confounding variable in participants' ratings of image liking.

Question:

Do you experience colourblindness?

Type of question:

Multiple Choice

Answers:

- Yes
- No
- I do not know

H.1.9. Variable: aesthetic fluency (Cotter et al., 2023)

Literature:

Smith and Smith (2006) describe aesthetic fluency as the knowledge foundation regarding art that facilitates aesthetic experiences in individuals.

Cotter et al. (2023) state that art knowledge is one of the most fundamental variables in the psychology of aesthetics and the arts.

Hypothesis-aesthetic fluency:

There will be a correlation between aesthetic fluency and ratings of how much they like images, when controlling for other variables, to investigate if aesthetic fluency acts as a potential confounding variable in participants' ratings of image liking.

Question:

Can you rate the following 10 items?

Type of question:

Based on Cotter et al. (2023): Rate items in a matrix on the following scale: On the following scale:

- 0 = I don't really know anything about this artist or term
- 1 = I'm familiar with this artist or term
- 2 = I know a lot about this artist or term.

Answers:

Based on Cotter et al. (2023): Pop art, Claude Monet, Cubism, Alessandro Botticelli, Gustav Klimt, Lithography, Gouache, Georgia O'Keeffe, Jean-Michel Basquiat, and Amedeo Modigliani.

H.1.10. Variable: aesthetic attitude Mcmanus and Furnham (2006)

Literature:

King (2012) describes aesthetic attitude as a way of approaching experiences. It is a state of mind, a way of relating and orienting oneself in the world around us. When you adopt an aesthetic attitude you focus on the features of a trigger that are relevant to aesthetic appraisal.

Mcmanus and Furnham (2006) outlined 5 aesthetic attitudes:

Anti-art: "High scores agreed that government funding for art should be redistributed to other services, that all kinds of art should be censored (as films are), that science is more important than art for our society, and that today's artists owe their success more to good marketing and publicity rather than talent, and they disagreed with statements that modern art is authentic, and that art can be created from any medium as long as the intention is to make art."

Aesthetic inclusivity: "high scorers saw art as being broadly defined, agreeing that science can be art, that sport is an art, that a child's drawing is art, that cordon bleu chefs are artists, and they agreed with the questions, 'Do you think the talents of Picasso can be compared on equal standing to those of the Beatles?'"

Emotion and understanding: "high scores agreeing that one has to understand the emotions of

the artist in order to appreciate the work, that the meaning behind art has to be obvious for it to have value, and that one needs to understand the background information of a piece of art to appreciate it properly. In addition, high scorers on this factor also agreed that art had to be controversial to make an impact, that in schools the arts are more important than the sciences, and that one has to like something to consider it as art."

Aesthetic relativism: *"high scorers agreeing that their appreciation of art has been influenced by academic tuition, by their education (school, college and university) and by their upbringing, that the media have a powerful influence over what is considered as good art, that the meaning of a piece of art changes with time, that art reflects the attitudes of society and that art is class restrictive."*

Aesthetic quality: *"high scores thought that art required skill, that art loses its value if mass produced, that it must provoke an emotional response, that artistic talent is innate, that children should be exposed to art through compulsory school trips to galleries, museums, theatres, etc., and that the ability of a piece of art to withstand the test of time is a better indicator of quality than its monetary value."*

Hypothesis-aesthetic attitude:

There will be a correlation between differing aesthetic attitudes and ratings of how much they like images, when controlling for other variables, to investigate if differing aesthetic attitudes act as a potential confounding variable in participants' ratings of image liking.

Question:

Which items apply to you?

Type of question:

Matrix for every attitude, where every item will be answered with yes/no

Items:

Based on the items of Mcmanus and Furnham (2006):

Attitude 1:

- Science can be art
- Sport is an art
- Children's drawings can be art
- Cordon blue chefs are artists
- The talents of Picasso can be viewed as on equal standing as the talents of the Beatles

Attitude 2:

- Your appreciation of art has been influenced by academic tuition, by your education (school, college and university) and by your upbringing
- The media have a powerful influence over what is considered as good art
- The meaning of a piece of art changes with time
- Art reflects the attitudes of society
- Art is class restrictive

Attitude 3:

- Art requires skill
- Art loses its value if it is mass produced
- Art must provoke an emotional response
- Artistic talent is innate
- children should be exposed to art through compulsory school trips to galleries, museums, theatres, etc.

- the ability of a piece of art to withstand the test of time is a better indicator of quality than its monetary value

Attitude 4:

- Government funding for art should be redistributed to other services
- All kinds of art should be censored (as films are)
- Science is more important than art for our society
- Today's artists owe their success more to good marketing and publicity rather than talent
- Modern art is not authentic
- Art cannot be created from any medium as long as the intention is to make art

Attitude 5:

- You have to understand the emotions of the artist in order to appreciate the work
- The meaning behind the art has to be obvious in order for the art to have value
- You need to understand the background information of a piece of art to appreciate it properly
- Art has to be controversial to make an impact
- In schools, the arts are more important than the sciences
- You have to like something to consider it as art

H.2. Confounding Variables Excluded in Experiment 1

H.2.1. Personal history

Why it could be relevant

As also described in the social aesthetic level of the Unified Model of Aesthetics, personal history is relevant to one's aesthetic experiences.

Why it is excluded from Experiment 1

Personal history is too broad and extensive to measure in a crowdsourcing experiment.

H.2.2. Personality traits

Why it could be relevant

Personality traits e.g.: "extraversion regarded extraverts as 'stimulus hungry', searching for new forms of sensory stimulation (Matthews, Deary, and Whiteman, 2003)."

Why it is excluded from Experiment 1

Too broad and extensive to measure in a crowdsourcing experiment

H.2.3. Aesthetic sensibility (Berleant, 2015)

Why it could be relevant

As described in Appendix B, is aesthetic sensibility described by the capacity to perceive, appreciate and evaluate sensory stimuli. Someone with high aesthetic sensibility has a heightened sensitivity towards aesthetic characteristics (e.g. Gestalt).

Why it is excluded from Experiment 1

Is a philosophical concept, and to the best of my knowledge there are no metrics of this notion. I assume aesthetic sensibility is measured (in part) by aesthetic preference and aesthetic fluency.

H.2.4. The device used for crowd working (Hettiachchi et al., 2020)

Why it could be relevant

Research indicates that the type of device used for crowdworking might influence the results.

Why it is excluded from Experiment 1

This confounding variable is mitigated by only letting participants work from a desktop.



Experiment 1 analysis

[Hyperlink to Section 4.4.3.](#)

I.1. Results difference per stimulus

ImageID	Mean control treatment	Mean SCA treatment	St. dev. control treatment	St. dev. SCA treatment	P-value
31	7.8	6.8	1.5491933	1.8737959	0.2098
32	5.9	5.7	2.3309512	2.3593784	0.8509
33	2.1	3.1	1.2866839	1.9119507	0.1869
34	7.4	7.9	2.3190036	1.66333	0.5864
35	7.3	6.1	1.8885621	1.3703203	0.1213
36	4.9	4.6	2.1317703	1.1737878	0.7012
37	4.1	5.0	1.7288403	2.2110832	0.324
38	3.1	3.3	1.2866839	2.3118055	0.8138
39	7.4	7.6	1.4298407	1.5776213	0.7698
40	6.7	7.2	2.3593784	2.0976177	0.6226
64	5.4	4.9	2.6331224	2.4698178	0.6666
65	4.7	4.7	2.5407785	1.7669811	1
66	7.4	7.4	1.0749677	2.1705094	1
67	4.3	4.4	1.8287822	2.0110804	0.9087
68	3.5	3.0	2.3687784	2.3094011	0.6384
69	5.5	5.6	2.4152295	2.5033311	0.9286
70	7.3	7.1	2.0027759	2.2335821	0.8354
71	6.1	6.9	1.197219	1.7919573	1.7919573
72	6.2	6.2	1.9888579	1.5491933	1
73	6.5	6.4	2.013841	2.1186998	0.915

Table I.1: Results mean comparison analysis experiment 1.

I.2. Results Variance

Foto id	St. dev. SCA treatment	St. dev. control treatment	F statistic	2-sided p-value
64	2.469818	2.633122	1.1366	0.8518
65	1.766981	2.540779	2.0676	0.2943
66	2.170509	1.074968	4.0769	0.0481
67	2.01108	1.828782	1.2093	0.7817
68	2.309401	2.368778	1.0521	0.941
69	2.366432	2.415229	1.0417	0.9525
70	2.233582	2.002776	1.2438	0.7505
71	1.791957	1.197219	2.2403	0.2453
72	1.549193	1.72884	1.2454	0.7491
73	2.1187	2.013841	1.1068	0.8823
31	1.873796	1.549193	1.463	0.5799
32	2.359378	2.330951	1.0245	0.9718
33	1.911951	1.286684	2.2081	0.2537
34	1.66333	2.319004	1.9438	0.3364
35	1.37032	1.888562	1.8994	0.3532
36	1.173788	2.13177	3.2984	0.0901
37	2.211083	1.72884	1.6357	0.475
38	2.311805	1.286684	3.2282	0.0958
39	1.577621	1.429841	1.2174	0.7743
40	2.097618	2.359378	1.2652	0.7318

Table I.2: Experiment 1 variance.

I.3. Results Region

ImageID	Portugal/UK p-value	Portugal/Poland p-value	Poland/UK p-value
64	0.0377	0.0467	0.3259
65	0.0165	0.0598	0.2415
66	0.4292	0.279	0.8967
67	0.798	0.9292	0.9511
68	0.4878	0.5214	0.9572
69	0.2602	0.4808	0.4764
70	0.3581	0.4761	0.6731
71	0.1778	0.2176	0.881
72	0.3304	0.7326	0.8041
73	0.5157	0.6951	1
31	0.3921	0.425	0.8614
32	0.7495	0.8578	0.8911
33	0.1533	0.2942	0.5801
34	0.5555	0.7463	0.7396
35	0.6215	0.6892	0.9728
36	0.4729	0.7127	0.7781
37	0.2073	0.5127	0.5274
38	0.0435	0.2486	0.3195
39	0.1573	0.3693	0.5975
40	0.3379	0.5265	0.7544

Table I.3: Results region analysis.

J

Experiment 2 analysis

[Hyperlink to Section 5.3.3.](#)

J.1. Results difference per stimulus

ImageID	Mean control treatment	Mean aesthetic value treatment	St. dev. control treatment	St. dev. SCA treatment	P-value
61	5.4	4.0	2.6331224	1.6035675	0.2065
62	4.7	5.25	2.5407785	2.1213203	0.6308
63	7.4	8.25	1.0749677	1.0350983	0.1096
64	4.3	3.75	1.8287822	2.1876275	0.569
65	3.5	3.125	2.3687784	1.7268882	0.7131
66	5.5	5.875	2.4152295	1.7268882	0.7168
67	7.3	8.25	2.0027759	1.5811388	0.2901
68	6.1	7.375	1.197219	1.5979898	0.0705
69	5.9	6.0	1.7288403	2.4494897	0.9203
70	6.5	7.635	2.013841	2.0658793	0.2613
61	7.8	6.4444444	1.5491933	2.0682789	0.1219
62	5.9	5.0	2.3309512	2.9580399	0.469
63	2.1	3.3333333	1.2866839	2.8722813	0.2353
64	7.4	8.1111111	2.3190036	1.9002924	0.4778
65	7.3	7.2222222	1.8885621	1.6414763	0.9252
66	4.9	5.3333333	2.1317703	1.8708287	0.6454
67	4.1	4.2222222	1.7288403	2.4381231	0.9003
68	3.1	3.6666667	1.2866839	1.5	0.3877
69	7.4	7.0	1.4298407	2.1794495	0.6388
70	6.7	5.8888889	2.3593784	1.6914819	0.4061




















Table J.1: Results analysis experiment 2, where the first block are the results of sub-stimulus set 1.1 and the second block are the results of sub-stimulus set 1.2.





















K

Experiment 3 analysis

[Hyperlink to Section 6.3.3.](#)

K.1. Overview of the two treatments

Relative ranking based on mean score	Control T Mean	Control T Images	Rank T mean score	Rank T images
1	2,75		3,8	
2	3,5		3,8	
3	3,625		4,7	
4	4,5		5	
5	4,5		5,1	
6	4,625		5,8	
7	4,635		6,2	
8	4,75		6,6	
9	5,5		7	
10	6,625		7	

Relative ranking based on mean score	Control T Mean	Control T Images	Rank T mean score	Rank T images
1	3.8		4,6666667	
2	4		4,7777778	
3	4,6		4,8888889	
4	4,9		5	
5	5,4		5,1111111	
6	5,6		5,3333333	
7	6,8		5,8888889	
8	7,4		5,8888889	
9	8,2		6,6666667	
10	8,2		6,7777778	

K.2. Internal consistency per participant for the Experiment 3 treatment

Participant	Slope/coefficient	p-value t test
p1	0.3818182	0.2763
p2	0.8909091	0.0005
p3	-0.1272727	0.7261
p4	-0.0424242	0.9074
p5	-0.1151515	0.7514
p6	0.1151515	0.7514
p7	0.0545455	0.881
p8	0.4060606	0.2443
p9	-0.0424242	0.9074
p10	0.369697	0.2931
p11	-0.0424242	0.9074
p12	0.6969697	0.0251
p13	0.2727273	0.4458
p14	0.9151515	0.0002
p15	0.5636364	0.0897
p16	-0.0909091	0.8028
p17	1	.
p18	1	.
p19	0.9393939	<0.0001

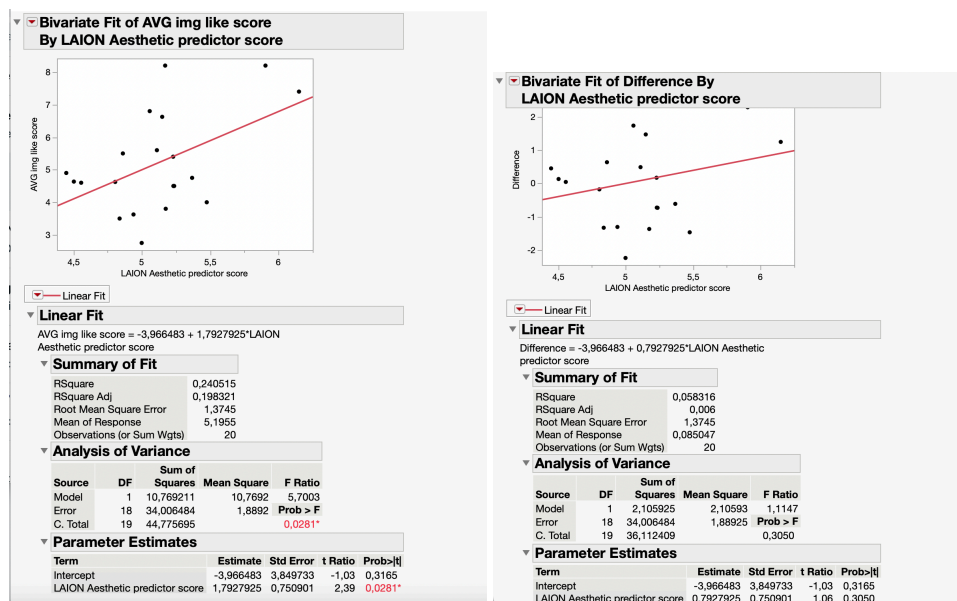
Table K.1: Results internal consistency within subjects



Comparison between aesthetic scores assigned by participants and the LAION aesthetic predictor analysis and overview








[Hyperlink to Section 9.3.](#)








L.1. The two linear regressions



L.2. Overview of the images, their scores assigned by humans and their scores assigned by the predictor

Images	Corresponding average image liking score	Corresponding LAION Aesthetic predictor score
 1	2,75	4.999312400817871
 2	3,5	4.837807655334473
 3	3,625	4.9393415451049805
 4	4,5	5.236815452575684
 5	4,5	5.231170177459717
 6	4,625	4.805502414703369

 7	4,635	4.503327369689941
 8	4,75	5.367993354797363
 9	5,5	4.861712455749512
 10	6,625	5.149194717407227
 11	3,8	5.175015926361084
 12	4	5.474945068359375
 13	4,6	4.556765079498291


 14	4,9	4.447845935821533
 15	5,4	5.229292392730713
 16	5,6	5.110494613647461
 17	6,8	5.057919979095459
 18	7,4	6.149407386779785
 19	8,2	5.904305934906006
 20	8,2	5.170899868011475

L.3. Qualitative results - stimuli where humans values stimuli noticeably lower/higher than the predictor

Stimuli where humans valued stimuli noticeably lower than the predictor

1. Rating: 2,750 Score: 4,999		2. Rating: 3,500 Score: 4,838		3. Rating: 3,625 Score: 4,939	
11. Rating: 3,800 Score: 5,175		12. Rating: 4,000 Score: 5,475			

Stimuli where humans valued stimuli noticeably higher than the predictor

10. Rating: 6,625 Score: 5,149		18. Rating: 7,400 Score: 6,149	
19. Rating: 8,200 Score: 5,904		20. Rating: 8,200 Score: 5,171	

M

Experiment 4 analysis

[Hyperlink to Section 7.3.3.](#)

M.1. Results difference per stimulus

To analyse if there is a statistically significant difference per image per treatment, t-tests are conducted (as is depicted in the table below). The bold p-value's are statistically significant.

ImageID	Mean control treatment	Mean 2AFC treatment	St. dev. control treatment	St. dev. 2AFC treatment	P-value
1	4.75	5.3	2.1876275	1.8885621	0.5748
2	4.5	4.5	1.9272482	2.2236107	1
3	6.625	8.4	1.9955307	2.4585452	0.1184
4	4.625	7.5	2.5599944	2.0682789	0.0179
5	2.75	2.5	1.2817399	1.7159384	0.7368
6	5.5	7	0.7559289	1.2472191	0.0088
7	3.5	4.1	1.7728105	2.7668675	0.6029
8	4.625	6.1	2.3867192	2.6012817	0.2332
9	3.625	3.9	1.9955307	2.5582112	0.8066
10	4.5	5.6	0.9258201	2.6331224	0.2786
11	7.4	9.3	2.0655911	0.9486833	0.0165
12	6.8	4.3	2.394438	1.8885621	0.0184
13	8.2	6.9	1.6865481	2.1317703	0.1478
14	8.2	8.1	1.8737959	1.7288403	0.9027
15	4	7	2.4944383	2.7487371	0.0199
16	3.8	2.3	1.3165612	2.1108187	0.0726
17	4.6	4.2	1.8973666	1.6193277	0.6182
18	5.4	3.7	1.5776213	1.1595018	0.0133
19	5.6	4.2	2.1186998	2.2010099	0.1645
20	4.9	4.7	2.6853512	2.3118055	0.8603

Table M.1: Results analysis experiment 2

Conclusion: 30 percent of the images show a statistically significant difference between the x

M.2. Internal consistency 2AFC treatment

To measure internal consistency, each participant is exposed to one image duo twice in the randomised loop order.

Participant	Experiment image	internal consistency twin image	Check?
p1	RIGHT	LEFT	NO
p2	RIGHT	RIGHT	YES
p3	RIGHT	RIGHT	YES
p4	RIGHT	RIGHT	YES
p5	RIGHT	RIGHT	YES
p6	RIGHT	RIGHT	YES
p7	RIGHT	RIGHT	YES
p8	LEFT	LEFT	YES
p9	RIGHT	RIGHT	YES
p10	RIGHT	RIGHT	YES
p11	LEFT	LEFT	YES
p12	RIGHT	RIGHT	YES
p13	RIGHT	RIGHT	YES
p14	RIGHT	RIGHT	YES
p15	LEFT	LEFT	YES
p16	RIGHT	RIGHT	YES
p17	RIGHT	RIGHT	YES
p18	LEFT	LEFT	YES
p19	LEFT	LEFT	YES
p20	LEFT	LEFT	YES

Table M.2: Results internal consistency within subject

Conclusion: 95 percent of participants passed the internal consistency check

M.3. Qualitative results


img nr	mean t1	mean t2	st dev t1	st dev t2	p-value	images
6	5,5	7	0,755929	1,247219	0,0088	
18	5,4	3,7	1,577621	1,159502	0,0133	
11	7,4	9,3	2,065591	0,948683	0,0165	
4	4,625	7,5	2,559994	2,068279	0,0179	
12	6,8	4,3	2,394438	1,888562	0,0184	
15	4	7	2,494438	2,748737	0,0199	
16	3,8	2,3	1,316561	2,110819	0,0726	
3	6,625	8,4	1,995531	2,458545	0,1184	
13	8,2	6,9	1,686548	2,13177	0,1478	
19	5,6	4,2	2,1187	2,20101	0,1645	
8	4,625	6,1	2,386719	2,601282	0,2332	
10	4,5	5,6	0,92582	2,633122	0,2786	
1	4,75	5,3	2,187628	1,888562	0,5748	
7	3,5	4,1	1,772811	2,766868	0,6029	
17	4,6	4,2	1,897367	1,619328	0,6182	
5	2,75	2,5	1,28174	1,715938	0,7368	
9	3,625	3,9	1,995531	2,558211	0,8066	
20	4,9	4,7	2,685351	2,311806	0,8603	
14	8,2	8,1	1,873796	1,72884	0,9027	
2	4,5	4,5	1,927248	2,223611	1	

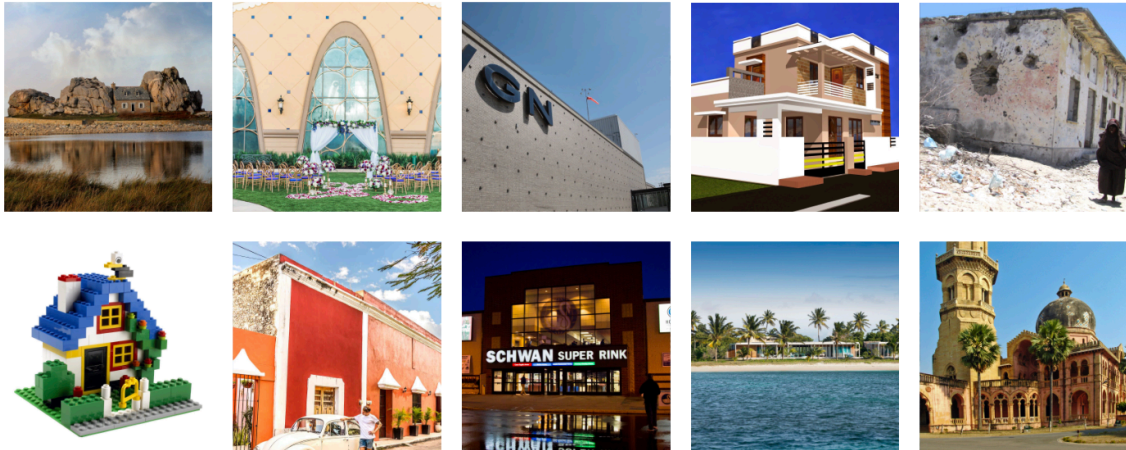
Figure M.1: Overview of the results of the 2AFC experiment, sorted on p-value of the t-test, with corresponding stimuli.

N

Earlier version of the stimulus set

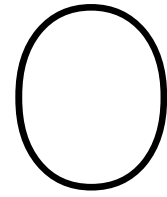
[Hyperlink to Section 3.2.5.](#)

Earlier version of the stimulus set



Earlier version of the stimulus set





Analysis content vs overall image liking

[Hyperlink to Section 8.3.3.](#)

Image nr	Mean content	Mean overall	St. dev. content	St. dev. overall	P-value t-test
1	6	7.5	3.0912062	2.6352314	0.2581
2	5.1	6.1	2.6012817	3.247221	0.4571
3	6.6	8	2.5905812	2.0548047	0.1973
4	7	7.2	1.7638342	1.4757296	0.7864
5	3	5.5	2.2607767	2.5927249	0.0338
6	6.1	5.9	1.7919573	1.9119507	0.812
7	4.5	3.6	2.013841	1.9550504	0.324
8	5	5.5	1.8257419	2.321398	0.5989
9	2.8	3.4	1.3165612	1.6465452	0.38
10	5.7	6	1.8885621	2.3570226	0.7571
11	8.3	8.3636364	1.6363917	1.5015144	0.9269
12	5.7	5.8181818	2.2135944	1.8877596	0.8963
13	7.1	8	2.0248457	1.2649111	0.2321
14	8	7.7272727	1.4142136	1.6787441	0.6934
15	5	5.6363636	2.7487371	2.5009089	0.5849
16	3.2	4.4545455	2.6997942	2.1148823	0.2481
17	3.9	4.2727273	2.4698178	1.6787441	0.6879
18	3.9	4.9090909	1.8529256	1.5135749	0.1859
19	4.7	5.1818182	1.7669811	1.6624188	0.5274
20	5	4.7272727	2.4037009	1.7939292	0.77

Table O.1: Results analysis experiment 1

P

Approved design brief

DESIGN
FOR OUR
future

TU Delft

IDE Master Graduation

Project team, Procedural checks and personal Project brief

This document contains the agreements made between student and supervisory team about the student's IDE Master Graduation Project. This document can also include the involvement of an external organisation, however, it does not cover any legal employment relationship that the student and the client (might) agree upon. Next to that, this document facilitates the required procedural checks. In this document:

- The student defines the team, what he/she is going to do/deliver and how that will come about.
- SSC E&SA (Shared Service Center, Education & Student Affairs) reports on the student's registration and study progress.
- IDE's Board of Examiners confirms if the student is allowed to start the Graduation Project.

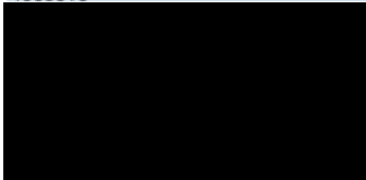
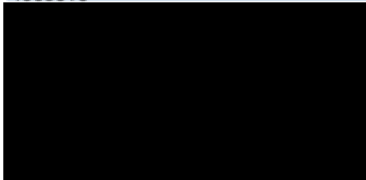
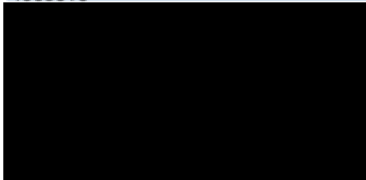
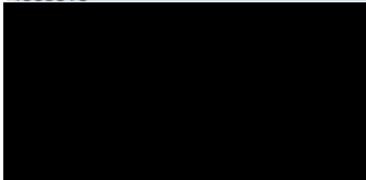
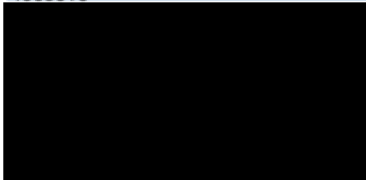
USE ADOBE ACROBAT READER TO OPEN, EDIT AND SAVE THIS DOCUMENT

Download again and reopen in case you tried other software, such as Preview (Mac) or a webbrowser.

STUDENT DATA & MASTER PROGRAMME

Save this form according the format "IDE Master Graduation Project Brief_familyname_firstname_studentnumber_dd-mm-yyyy".

Complete all blue parts of the form and include the approved Project Brief in your Graduation Report as Appendix 1 !

family name **Offerman** 6308
 initials **C.E.** given name **Céline**
 student number **4663616**
 street & no. 
 zipcode & city 
 country 
 phone 
 email 

Your master programme (only select the options that apply to you):

IDE master(s): IPD Dfi SPD

2nd non-IDE master: _____

individual programme: - - (give date of approval)

honours programme: Honours Programme Master

specialisation / annotation: Medisign

Tech. in Sustainable Design

Entrepreneurship

SUPERVISORY TEAM **

Fill in the required data for the supervisory team members. Please check the instructions on the right !

** chair **Alessandro Bozzon** dept. / section: **SDE**
 ** mentor **Willem van der Maden** dept. / section: **HCD**
 2nd mentor _____
 organisation: **TU Delft**
 city: **Delft** country: **The Netherlands**

comments
(optional)

⋮

Chair should request the IDE Board of Examiners for approval of a non-IDE mentor, including a motivation letter and c.v.



Second mentor only applies in case the assignment is hosted by an external organisation.



Ensure a heterogeneous team. In case you wish to include two team members from the same section, please explain why.

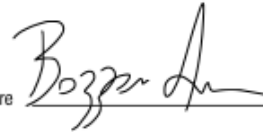
Procedural Checks - IDE Master Graduation

APPROVAL PROJECT BRIEF

To be filled in by the chair of the supervisory team.

 chair Alessandro Bozzon date 02 - 03 - 2023

signature


CHECK STUDY PROGRESS

To be filled in by the SSC E&SA (Shared Service Center, Education & Student Affairs), after approval of the project brief by the Chair. The study progress will be checked for a 2nd time just before the green light meeting.

 Master electives no. of EC accumulated in total: 27 EC

 Of which, taking the conditional requirements into account, can be part of the exam programme 27 EC

List of electives obtained before the third semester without approval of the BoE

 YES all 1st year master courses passed

 NO missing 1st year master courses are:

 name Robin den Braber date 07 - 03 - 2023

signature

 Robin
den
Braber
Digitaal ondertekend door Robin den Braber
Datum: 2023.03.07 08:29:13 +01:00
FORMAL APPROVAL GRADUATION PROJECT

To be filled in by the Board of Examiners of IDE TU Delft. Please check the supervisory team and study the parts of the brief marked **. Next, please assess, (dis)approve and sign this Project Brief, by using the criteria below.

- Does the project fit within the (MSc)-programme of the student (taking into account, if described, the activities done next to the obligatory MSc specific courses)?
- Is the level of the project challenging enough for a MSc IDE graduating student?
- Is the project expected to be doable within 100 working days/20 weeks ?
- Does the composition of the supervisory team comply with the regulations and fit the assignment ?

 Content: APPROVED **NOT APPROVED**

 Procedure: **APPROVED** NOT APPROVED

- the adapted version has been approved (separately)

comments

 name Monique von Morgen date 21 - 03 - 2023

signature

Personal Project Brief - IDE Master Graduation

Data annotation for aesthetics project title

Please state the title of your graduation project (above) and the start date and end date (below). Keep the title compact and simple. Do not use abbreviations. The remainder of this document allows you to define and clarify your graduation project.

start date 13 - 02 - 2023 end date 07 - 07 - 2023

INTRODUCTION **

Please describe, the context of your project, and address the main stakeholders (interests) within this context in a concise yet complete manner. Who are involved, what do they value and how do they currently operate within the given context? What are the main opportunities and limitations you are currently aware of (cultural- and social norms, resources (time, money,...), technology, ...).

Developing a process for annotation of images on the basis of aesthetics with the purpose of training generative models:

The age of AI is here. Biases are a big problem for this technology (Google AI, n.d.). The mislabeling of data for machine learning models is one of the causes of this problem, which will be the focus of this project. The project will specify on the LAION-5B dataset, and develop a more valuable process to evaluate its aesthetics. Aesthetics is a rich and complex concept, but it is important to keep the labeling simple. To achieve this goal, methods of crowd-sourced labeling will be tested to take into account the unified model of aesthetics developed by Berghman & Hekkert (2017).

With 10 million daily users, Stable Diffusion, the image generating-system by Stability AI has a considerable reach (Mehta, 2022). Stable Diffusion is trained on three sub-sets of the LAION-5B data set. These sub-sets are pairs of images and their alt-text, derived from the internet. The three LAION sub-sets are laion2B-en, laion-high-resolution, and laion-aesthetics v2 5+. The evaluation of the 'aesthetics' of the pictures was done by asking participants the following question: "How much do you like this image on a scale from 1 to 10?" (Schuhmann, 2022). The reason the current evaluation was conducted this way is most likely due to its simplicity. This oversimplified question lacks the richness to adequately describe aesthetics. For a data set that is deployed on such a large scale and has a touch point with millions of people a day, it is bad that aesthetics are so superficially reflected. This is where a significant improvement can be made, and that will be the goal of this project.

There are several models regarding aesthetics that are universal but relate differently in context. One promising model to apply here is Berghman & Hekkert's (2017) unified model of aesthetics, because it is a profound theory based around three dichotomous facets. Berghman & Hekkert's model relies not on a superficial notion of beauty (or "likeability," as the current dataset reflects), but on deeper aesthetic value. As the dataset becomes richer based on aesthetics, it may be a good step in the right direction to reduce biases. These researchers proved the following universal impacts on aesthetics: connectedness - autonomy, typicality - novelty, unity - variety (see figure 2 in this project brief).

Connectedness - autonomy: describes how humans feel the need to be part of a group, whilst preserving a sense of individuality.

Typicality - novelty: describes our need for the familiar in contrast with our evolutionary want for new.

Unity - variety: describes the craving to feel a sense of one whole, where things fit together, and our interest in contrasting elements within this whole

Part of the design challenge will be to communicate these abstract and complex principles to the participants.

The goal of this research project is to develop a method for annotating images based on a nuanced and scientific model of aesthetics. A challenge here will be to maintain simplicity, the strength of the current analysis method. It is therefore important that the new process is still based on simplicity, while accounting

space available for images / figures on next page

Personal Project Brief - IDE Master Graduation

introduction (continued): space for images

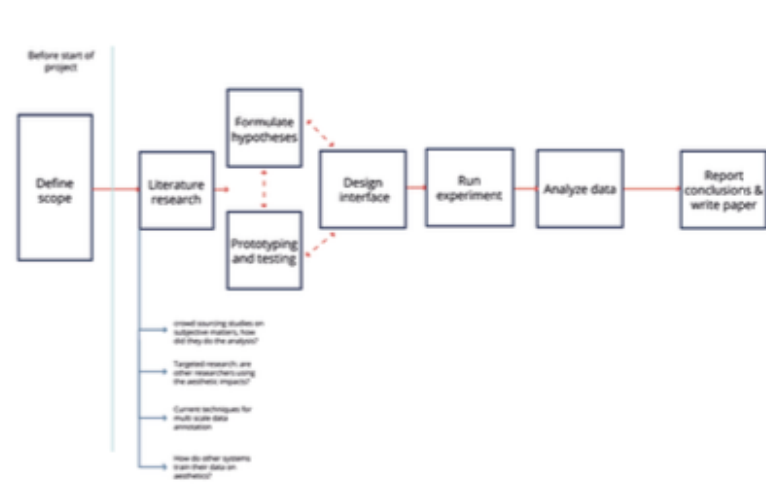


image / figure 1: Overview of the project



image / figure 2: Example of an image annotation tool by Lomas et al. (2023)

Personal Project Brief - IDE Master Graduation**PROBLEM DEFINITION ****

Limit and define the scope and solution space of your project to one that is manageable within one Master Graduation Project of 30 EC (= 20 full time weeks or 100 working days) and clearly indicate what issue(s) should be addressed in this project.

The goal of the project is to create and design a process for the annotation of images on the basis of aesthetics with the purpose of training generative models.

Included in the scope:

- an exploration of existing models to evaluate aesthetics
- assessing these models and making a choice in which one is most suitable for this purpose
- examining how the selected model can be most effectively used for the annotation of images
- Analysing and reporting the results

Excluded from the scope:

- expanding knowledge what aesthetics is
- applying the findings of the research in a generative AI model

ASSIGNMENT **

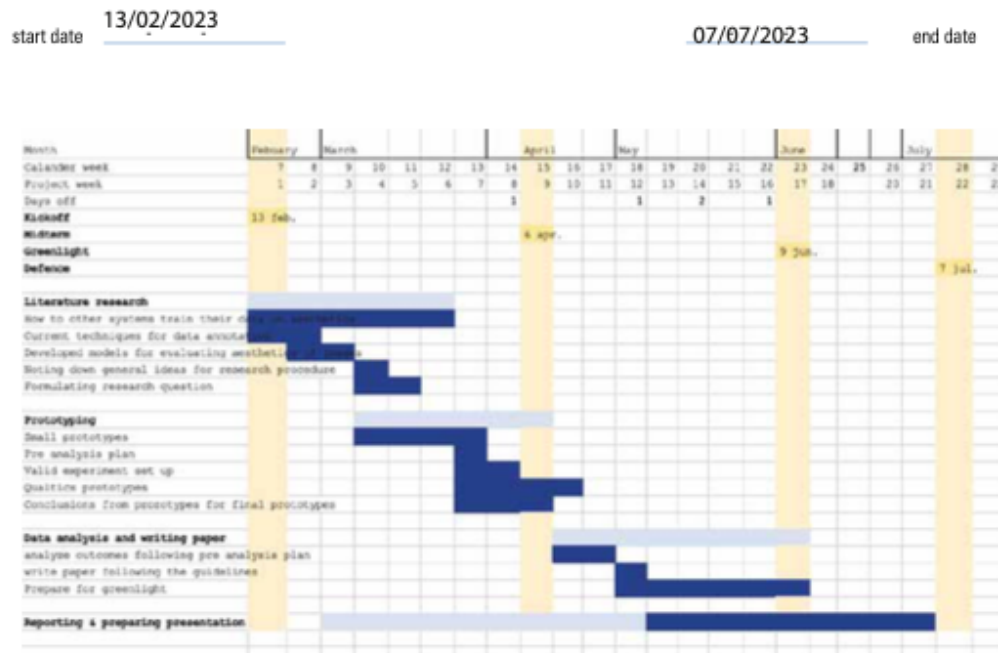
State in 2 or 3 sentences what you are going to research, design, create and / or generate, that will solve (part of) the issue(s) pointed out in "problem definition". Then illustrate this assignment by indicating what kind of solution you expect and / or aim to deliver, for instance: a product, a product-service combination, a strategy illustrated through product or product-service combination ideas, In case of a Specialisation and/or Annotation, make sure the assignment reflects this/these.

We will explore which aesthetics evaluation method is most relevant on data selection for the LAION data set. In addition, we will look at the best way to conduct this evaluation.

The desired outcome of this project is a data annotation methodology that allows crowd-workers to images on multi-scale aesthetic principles. The methodology will also be studied and evaluated through online experiments.

PLANNING AND APPROACH **

Include a Gantt Chart (replace the example below - more examples can be found in Manual 2) that shows the different phases of your project, deliverables you have in mind, meetings, and how you plan to spend your time. Please note that all activities should fit within the given net time of 30 EC = 20 full time weeks or 100 working days, and your planning should include a kick-off meeting, mid-term meeting, green light meeting and graduation ceremony. Illustrate your Gantt Chart by, for instance, explaining your approach, and please indicate periods of part-time activities and/or periods of not spending time on your graduation project, if any, for instance because of holidays or parallel activities.



The GANTT chart could not export with the specific dates in view, because of the duration of the project. The specific dates of the project can be found in figure 2.

The important dates of the project:

13 February: kick-off meeting

6 April: midterm meeting

9 June: green light

7 July: defense

As you can see I scheduled two weeks extra for the project (on top of the 100 days). I did this because there are quite some holidays in the 3th and 4th quarter of the academic year, and to give myself some space to take a few days off if the project allows it.

Personal Project Brief - IDE Master Graduation

MOTIVATION AND PERSONAL AMBITIONS

Explain why you set up this project, what competences you want to prove and learn. For example: acquired competences from your MSc programme, the elective semester, extra-curricular activities (etc.) and point out the competences you have yet developed. Optionally, describe which personal learning ambitions you explicitly want to address in this project, on top of the learning objectives of the Graduation Project, such as: in depth knowledge a on specific subject, broadening your competences or experimenting with a specific tool and/or methodology, Stick to no more than five ambitions.

I did my undergraduate and graduate studies at this faculty. I learned many design competencies here, but lack research skills. I find this a pity, because in my master (Design for Interaction) I found out that I really enjoy doing research, and I would like to do a PhD after graduation. Hence, I consciously chose to work on a research project, which does not result in design outcomes.

I would like to get better at setting up a solid research and data analysis. Next to this I would also like to improve my scientific writing skills. While this was a driver for seeking a project like this, it is not the immediate reason I am writing this proposal now.

What is happening in our world right now I find super interesting. I expect products like Stable Diffusion and GPT to drastically change the way we work. Artificial intelligence has been around for a while, but these tools make the connection between people who don't actually know anything about coding or computers and this technology. Last Christmas, my mother (a copywriter) told me that she had started experimenting with prompts in GPT, in order to get snippets of text prescribed. To set up her WiFi, she still calls me or my brother, but GPT allows her to use machine learning for her work. I find this very fascinating and think it is important that user-centered designers play a big role in shaping these 'translators'.

FINAL COMMENTS

In case your project brief needs final comments, please add any information you think is relevant.