Beauty and Color in Human-Computer Interaction

An Implicit Measurement

***Bachelor Thesis** by*

Lennart Overkamp

UNIVERSITEIT TWENTE.

Beauty and Color in Human-Computer Interaction

An Implicit Measurement

***Bachelor Thesis** by*

Lennart Overkamp
s1088017
Bachelor Psychology
Faculty Behavioral Sciences

Supervisors:
Dr. Martin Schmettow
Inga Schwabe

Enschede, 17 June 2013

UNIVERSITEIT TWENTE.

Contents

Abstract

This study aimed to find out whether the Stroop Priming Task could be a suitable alternative to explicit methods for measuring first impression beauty judgments of website users, as well as to shed more light on the relationships between color, visual complexity, prototypicality and the judged beauty of websites. The conducted Stroop Priming Task contained target words semantically related to beauty, neutrality and ugliness, and screenshots of website homepages, varying in visual complexity, prototypicality and color, as priming stimuli. Results indicate that the used Stroop Priming Task was not able to accurately measure the judged beauty of websites in website users' first impressions. Also, the results indicate that the absence or presence of color in websites does not significantly influence the first impression beauty judgments of website users. Possible explanations are discussed, and further research is proposed.

1. Introduction

In recent years a growing number of *Human Computer Interaction (HCI)* researchers has become interested in the study of *beauty* (Bargas-Avila & Hornbaek, 2011; Tuch, Presslaber, Stöcklin, Opwis & Bargas-Avila, 2012). The topic of beauty is of course not new; for example, Arnheim (as cited in Leder, Belke, Oeberst & Ausgustin, 2004) already showed in 1954 that people generally prefer good Gestalts, and Frith (1974, as cited in Leder *et al.*, 2004) found that symmetry is preferred over non-symmetry. However, the study of beauty in the context of HCI topics is a relatively new research direction.

Beauty is often considered to be synonymous with terms such as *visual appeal*, *aesthetics* and *attractiveness* (Lavie & Tractinksy, 2004; Tracktinsky, Cokhavi, Kirschenbaum & Sharfi, 2006); Tuch *et al.*, 2012), although Lindgaard, Fernandes, Dudek & Brown (2006) argue that the synonyms aesthetics and visual appeal differ from beauty. However, they are also quick to point out that beauty is an "elusive and confusing" construct (p. 116). In this paper all four above mentioned terms are considered to be the same constructs.

Aesthetics, and therefore beauty, is considered to be an important dimension of *User Experience (UX)*, which is a term used to indicate the body of research that studies the usage quality of interactive products. Aesthetics, as well as affect, are often used as indicators for assessing the UX of these products (Bargas-Avila & Hornbaek, 2011). However, this study keeps UX out of the picture by focusing solely on the concept of aesthetics.

Lavie & Tractinsky (2004), who conducted multiple factor analyses on the perceptions of users on beauty, found two dimensions of aesthetics: classical and expressive. *Classical aesthetics* refers to the orderliness in the design of the object, while *expressive aesthetics* refers to the originality and creativity of the object's designers. Hassenzahl & Monk (2010) regarded beauty as "a predominantly affect-driven evaluative response" to the appearance of the regarded object (p. 239), thereby emphasizing the role of responses from perceivers. These evaluative responses to objects can be considered to be a part of an *aesthetic experience,* which is defined by Leder *et al.* (2004) as "a cognitive process accompanied by continuously upgrading affective states that vice versa are appraised, resulting in an (aesthetic) emotion" (p. 493).

Leder *et al.* (2004) furthermore proposed a model for aesthetic experiences. According to this *model of aesthetic experience,* which is depicted in figure 1, people go through five stages before they reach an aesthetic judgment of an object. The first two stages, the *perceptual analysis* of the object and the *implicit memory integration*, are processed implicitly and automatically, and are therefore unconscious to the individual. The last three stages, *explicit classification*, *cognitive mastering* and *evaluation*, involve deliberate processing and are therefore explicit and conscious to the individual. Apart from going through these stages, there is a continuous emotional development, emphasizing that beauty judgments are both cognitive and affective (Leder *et al.* 2004).

Of particular interest in this study are the first two implicit, unconscious stages, as these can be considered to be *first impressions* of people when they have aesthetic experiences.

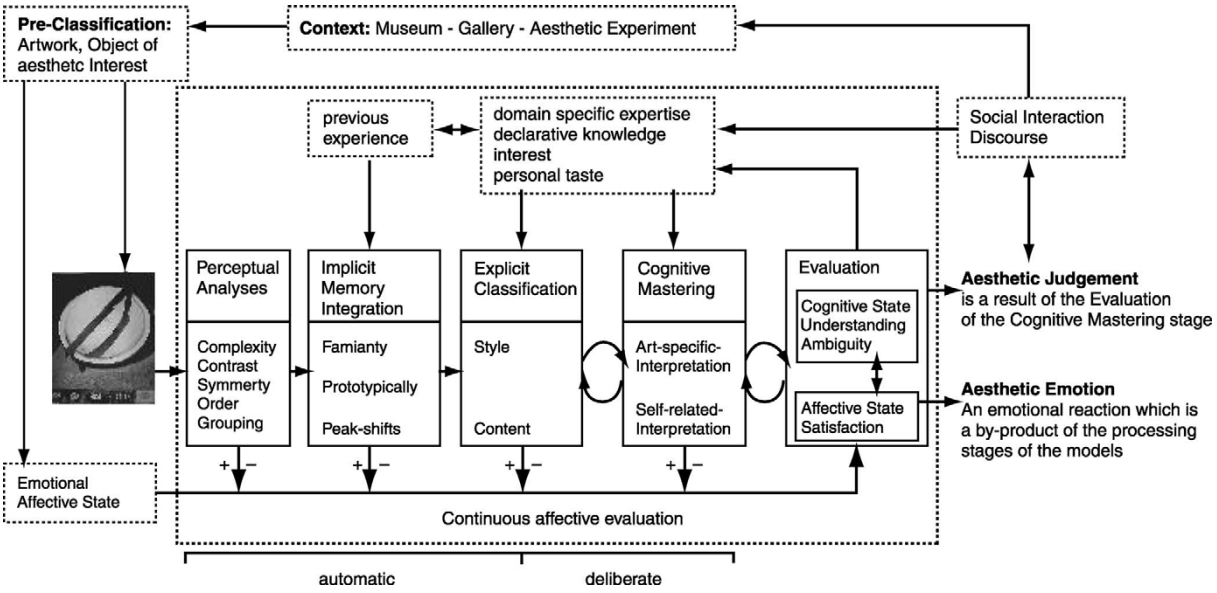

Figure 1. Model of aesthetic experience. Reprinted from Leder, H., Belke, B., Oeberst, A & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95*, 489-508. doi: 10.1348/0007126042369811

1.1 First impressions and website aesthetics

A particular part of the research on beauty that got the interest of HCI researchers is the focus on first impressions of website users, but only since relatively recently. Within this research field, researchers aim to understand what the underlying processes and consequences of first impressions of websites are. However, as will become clear in subsequent sections, there is still much unclear about these processes. This served as the main motivation for this study, in which an attempt is made to shed more light on these processes.

It is generally considered that the study of Lindgaard *et al.* (2006) started this body of research (Tuch *et al.*, 2012). In this article, Lindgaard and colleagues describe, among others, a study in which participants had to rate website homepages on visual appeal by pressing numeric keys 1 (‘very unappealing’) to 9 (‘very appealing’). The homepages appeared either 500 milliseconds or 50 milliseconds, depending on the condition the participants were assigned to. Results showed that participants form stable judgments of visual appeal of websites even within 50 milliseconds (Lindgaard *et al.*, 2006). This result was later confirmed by Tractinsky *et al.* (2006) who sought to replicate the study of Lindgaard and colleagues (2006).

This is an important finding, since first impressions can have long-term effects on the way the rest of the object's properties are evaluated. The additive value of this research field is therefore that it may enable website designers to intentionally shape their designs in such a way that it invokes positive first impressions, therefore increasing the chance that negative characteristics of the design may be overlooked in the long term (Lindgaard *et al*., 2006; Tuch *et al.*, 2012).

In this study, three website characteristics will be varied and checked for their effect on beauty judgments: visual complexity, prototypicality and color. These characteristics will be clarified in the following sections.

1.1.1 Visual complexity

Visual complexity (VC) is a concept that is difficult to define, although many attempts were made in literature. Xing & Manning (2005) did a review of definitions of complexity, and found that they generally all contained the same three components of a system: *numeric size*, *variety* and *structural rules*. While the numeric size of elements is often larger in complex systems, this component alone does not constitute complexity. And while it has been found that a *mediate* variety of system elements is often perceived as complex (as opposed to low or high variety), variety alone does not constitute complexity either. Only when the numeric size is large, the variety mediate and the system seems to contain internal rules for its structure, will the system be regarded as complex by human perceivers (Xing & Manning, 2005).

As becomes clear from the review of Xing & Manning (2005), complexity can generally be divided into two types: information complexity and cognitive complexity. *Information complexity* is the complexity from the perspective of the system *itself*. Measurement of this type of complexity is often done by performing computations on various elements of the system itself. *Cognitive complexity* is the complexity from the perspective of the *users* of the system. Measurement of this type of complexity often concerns participants who are required to assess the amount of complexity of the system in some way (Xing & Manning, 2005).

Visual complexity, then, can be regarded as synonymous with cognitive complexity, since it is the complexity that is perceived (visualized) from the perspective of, in this case, website users. Therefore, in this study, the example of Tuch *et al.* (2012) is followed by not using an objective, quantifiable definition of VC, but a subjectively perceived VC.

According to Berlyne's *theory of aesthetic preference* (1974, as cited in Tuch *et al.*, 2012 and in Martindale, Moore & Borkum, 1990) VC is a good predictor of the amount of pleasure people experience while looking at objects. The theory predicts that for optimal viewer's pleasure the VC should be moderate, since VC, ambiguity and novelty are related to an inverted U-shaped arousal curve. Indeed, Berlyne found that people prefer medium levels of VC, and Geissler, Zinkhan & Watson (2006) found in their study that web pages with moderate VC facilitate communication and gain favorable responses.

Berlyne's theory, however, has been found to be flawed. Martindale and colleagues (1990), who conducted experiments to test Berlyne's theory, mostly did not find inverted U-shaped curves. Also, Pandir & Knight (2006), who applied Berlyne's theory to the evaluation of website homepages, did not find support for this theory. Rather, they found a moderate *negative* correlation between complexity and pleasure (-.644), as well as a high negative correlation between complexity and interestingness (-.867).

More recently, Tuch *et al.* (2012) investigated the role of VC, as well as *prototypicality* (*PT*, see the section below), of websites on users' first impression. In their first study, they selected screenshots of 270 websites from the internet, and gathered VC and PT ratings of all these screenshots. Subsequently, they allocated 120 screenshots to 6 categories, made up of all possible combinations of VC (low, medium and high) and PT (low and high). They then required participants to rate these 120 website screenshots on perceived beauty using the *Visual Analog Scale (VAS).* Depending on the experimental condition the participants were in, they only had either 50, 500 or 1000 milliseconds to respond. In the second study, these presentation times were lowered to 17, 33 or 50 milliseconds. The results showed that website users prefer web pages that have low as well as medium VC above web pages with high VC, but only when PT is high rather than low. Also, the results showed that when presentation times were lower than 50 milliseconds, the processing of VC happens prior to the processing of PT, lending support to the model of aesthetic experience of Leder *et al.* (2004) (Tuch *et al.*, 2012).

Contrary to Berlyne's theory, Tuch and colleagues (2012) found linear shaped relationships between VC and beauty. They suggested this is because of the complexity of websites, placing them automatically on the right end of the inverted U-shaped curve of Berlyne. They agree however, that VC is a strong predictor of aesthetic judgments.

1.1.2 Prototypicality

Prototypicality is defined by Leder *et al.* (2004) as "the amount to which an object is representative of a class of objects". In other words, an object is said to be prototypical when it represents a certain class of objects well (Tuch *et al.*, 2012). Through experience with the internet, users form *mental models* of websites (Tuch *et al.*, 2012), which contain the users' understanding and knowledge of websites, and the expectancies of how websites work and how to use them (Wickens, Lee, Liu & Gordon Becker, 2004; Roth, Schmutz, Pauwels, Bargas-Avila & Opwis, 2010). Roth *et al.* (2010) showed that website users generally hold the same mental models, since they agree on the locations of most website elements. A website with high PT, then, should feel familiar to most users, since the locations of its elements correspond well with a (correct) mental model of that website.

In the context of beauty, Sen & Lindgaard (2008, as cited in Tuch *et al.*, 2012) found a positive correlation between PT and aesthetic appeal of images of basic objects. Veryzer & Hutchinson (1998) provided participants with product drawings that varied in PT and unity, a visual aspect that connects parts of a display in a meaningful way. They found that PT and unity had positive effects on the attractiveness judgments of the participants.

However, there is not much research yet that studied the role of PT in beauty judgments of websites. Tuch *et al.* (2012), one of the few who researched this topic, found in their study a positive relationship between PT and beauty, where websites with high PT induce better first impressions than websites with low PT. This effect was found to be stronger for websites with low VC, as opposed to high VC. As mentioned above, websites with low VC and high PT were found to be the most aesthetically appealing (Tuch *et al.*, 2012).

1.1.3 Color

The third and last website characteristic that will be varied during the experiment of this study is *color (CL)*. According to Leder *et al.* (2004), color is processed in an aesthetic experience as early as in the perceptual analysis stage (i.e. first stage) of their model, while also being processed deliberately in the explicit classification stage (i.e. third stage). Hall & Hanna (2004) researched whether different combinations of font and background colors would affect, among others, judgments of aesthetics. They found that this was indeed the case, with light blue font on dark blue background being judged most aesthetically pleasing, followed by cyan font on black background.

Moshagen & Thielsch (2010) performed a literature review of visual aesthetics, from which they derived a multitude of visual aspects associated with aesthetics. One of these visual aspects was color. They required participants to rate on a scale from 1 (‘not important at all’) to 5 (‘very important) whether they found these visual aspects to be important factors of visual aesthetics. Results showed that participants judged color to be the second most important aspect (M = 4.19), only losing to simplicity (M = 4.72). Moshagen and Thielsch later used these results to develop a rating scale for perceived website beauty, which they named the *Visual Aesthetics of Website Inventory (VisAWI)*. Visual aesthetics, the first order factor of the VisAWI, was found to consist of four second order factors, of which *colorfulness* was one. Furthermore, they found colorfulness to be significantly and positively correlated with classic aesthetics (.61), expressive aesthetics (.44) and general appeal (.60) (Moshagen & Thielsch, 2010).

As becomes clear from the literature review of Moshagen & Thielsch (2010), there have been multiple studies that recognized color as an important factor in beauty judgments. However, to the knowledge of the author, no literature has been found that researched whether the mere *absence* of color (i.e. grayscale) influences the beauty judgments of website users. However, the positive correlations between colorfulness and aesthetics found by Moshagen & Thielsch (2010) indicate that lower colorfulness is associated with lower beauty judgments. This implies that grayscale websites, which can be regarded as the absolute minimum of colorfulness, will be judged less beautiful that colored websites.

One purpose of this study is to shed more light on the relationships between website characteristics and the mechanics behind first impressions of websites by looking at the three above mentioned website characteristics. Another main goal of this study is to see whether these relationships can be measured *implicitly*, as will be discussed next in this paper.

1.2 Explicit versus implicit measurement of beauty

Up to now, beauty has mostly been *explicitly* measured using rating scales such as the ones used by Lavie & Tractinsky (2004) and the VAS (Tuch *et al.*, 2012), and questionnaires such as the *AttracDiff 2* (Hassenzahl, 2004) and the VisAWI of Moshagen and Thielsch (2010).

Explicit measures do however have some disadvantages. Most notable is the response bias known as *social desirability* (Paulhus & John, 1998; Paulhus, 2002; Robinson & Neighbors, 2006). When participants, and people in general, are describing themselves through explicit self-report, they tend to exaggerate their positive characteristics or talents, while minimizing or even rejecting their negative characteristics. In this way, their actual standing on the subject is hidden or distorted, sometimes even unconscious to the person in question (Paulhus & John, 1998; Paulhus, 2002).

Other response biases of explicit measures include the tendency to choose alternatives on the left side of a scale more often than those on the right side of the scale, the tendency to avoid extreme alternatives on a continuum of a rating scale (*leniency error*), and *central tendency*, which is a bias towards the middle alternative on a rating scale (Cohen & Swerdlik, 2010; Cohen, Manion & Morrison, 2011). Also, responses to explicit measures can be influenced by the wording of items or the omission of possible answering categories (Cohen, Manion & Morrison, 2011).

Thus, explicit measurement tools require honest, conscious self-insight from participants during self-reports in order to be valid. *Implicit* measurement tools, on the other hand, do not require this self-insight, since their validity depends not on self-report, but on the *performance* of the participants (Robinson & Neighbors, 2006).

Implicit measures do of course also have some disadvantages, as summed up by Robinson & Neighbors (2006). Most notable, they often have low test-retest stability in comparison to explicit measures. However, Robinson & Neighbors also argue that "implicit processes are inherently instable" (p. 123), so low test-retest stability should not necessarily lessen the validity of implicit measures. Also, the high test-retest stability of self-reported characteristics might be caused by the fact that people have relatively permanent beliefs about themselves (Robinson & Neighbors, 2006).

Despite this disadvantage, and taking into account the above mentioned advantage of implicit measures over explicit measures, this study aims to find out if implicit measures can be a suitable replacement to explicit measures when measuring first impression beauty judgments of website users. The following section describes the implicit measurement tool which was chosen for this study, and why it should be sufficient for the purposes of this study.

1.3 Stroop Priming Task

The implicit measurement tool used in this study is the well-known *Stroop Task*, first conducted by and named after the researcher J. R. Stroop. In his famous article of 1935, Stroop described three experiments he performed to study the interference effects of *incongruent* color words (e.g. the word 'red', printed in a blue ink) (Stroop, 1935; MacLeod, 1991). In the first experiment, participants were required to read the color *names* of incongruent color words, or *target words*, as well as the same words printed in a black ink. Results of this first experiment showed that, on average, it took participants about two seconds longer to read the incongruent target words than it took to read the black printed target words (Stroop, 1935; MacLeod, 1991). The second experiment required participants not to *read* the target words, but to name the *ink color* of the target words. Again, incongruent target words were used, but the black printed target words of the first experiment were substituted for *congruent* target words (e.g. the word 'red', printed in a red ink). Results showed that the average time participants needed to name the ink colors in the incongruent condition was 74 percent longer than in the congruent condition (Stroop, 1935; MacLeod, 1991). In his third and last experiment during this study, Stroop wanted to examine the effect of practice by, among other things, presenting the same conditions as in the second experiment for eight consequent days. He found that the practice decreased the interference in naming the ink colors of incongruent target words, but that it did not eliminate it entirely (Stroop, 1935).

Over the years, the Stroop Task has been used on multiple occasions, and with multiple variations. The standard Stroop Task is the *Color-Word Interference Test (CWIT)*, which follows the procedure used in Stroop's second experiment (Stroop, 1935; MacLeod, 1991). Participants are presented with congruent as well as incongruent target words, and are required to name, either verbally or by pressing a key, the ink colors of these words as quickly as they are able to. The interference can be measured by calculating the difference between response times of the incongruent and the congruent conditions (MacLeod, 1991).

MacLeod (1991) gives an overview of notable variations of the CWIT, including the *Picture-Word Interference Task*, wherein participants are required to name pictures that contain (incongruent) target words, and an *Auditory Analog*, wherein participants are required to articulate words that are presented auditory in an (in)congruent fashion (such as the word 'low' presented in a high pitch). Other variations include the usage of different hues and pronounceability of the target words (MacLeod, 1991).

More relevant to the current study however, is the *semantic variation* of the CWIT. The study of Klein (1964) has been credited to be the first examining this subject (MacLeod, 1991). In this study, Klein required his participants to name the ink colors, red, green, yellow or blue, of target words in six different categories. These categories differed in the amount of association with colors, ranging from nonsense-syllables to the standard incongruent target words. Results showed a rising of the average inference as the amount of association increased, although differences between consequent categories not always reached statistical significance (Klein, 1964). Dalrymple-Alford (1972), who used a similar approach to study the differences between different amounts of association with colors as Klein did, observed greater response times for incongruent color *name* words than for incongruent color-*related* words.

Summarizing, these early studies suggest that the more associated with color target words are, the more interference they cause when naming their ink colors. Or, as MacLeod puts it, "*as the word's semantic association to the concept of color increases, so does its potential to interfere."* (1991, p. 173).

Interference in color-naming is however not only caused by color associations, but can be caused by association with *priming stimuli* as well. This was first observed by Warren (1972). He primed participants by presenting a small list of words from a certain category, e.g. 'oil', 'gas' and 'coal'. After each priming event, participants were required to perform a color-naming task, in which the target words were either control words, a row of Xs, a list category name, or a word from a list. The experiment was divided into three conditions. In the first condition no list was presented, and served as a control to which the other conditions could be compared. In the second condition, named 'irrelevant list', participants were in each trial presented with a list of three words, each belonging to the same category. This was followed by a target word that was *irrelevant* to the just presented list (e.g. the category name 'relatives' or the list word 'aunt'). In the third condition, 'relevant list', in each trial a list of three words was again used as priming, followed by a target word that was *relevant* to the presented list (e.g. the category name 'fuel' or the list word 'gas'). Results of this experiment showed that it took participants longer to name the color of the target words in the relevant condition compared to the irrelevant and control conditions, indicating that the semantic associations of the relevant target words with the priming stimuli interfered in the color-naming tasks (Warren, 1972).

Warren (1974) later provided more evidence for this understanding. By creating stimulus-target pairs that differed in their strength of semantic association (i.e. high, medium and low associative strength), he showed that the interference declined as the associative strength between the priming stimuli and the target words declined. Also, he provided evidence that the semantic association is *forwards*, and not *backwards*, indicating that the priming stimulus activates the association, not the target word (Warren, 1974; MacLeod, 1991).

Additionally, it has been found that whole semantically related sentences could also be used as priming stimuli in order to produce the interference effects in color-naming tasks, not only with single words (MacLeod, 1991).

More recent studies have also used this priming variant of the CWIT, which will generally be referred to in this paper as the *Stroop Priming Task* (*SPT*). For instance, Sparrow, Liu & Wegner (2011) used a SPT to explore the question whether the internet has become an external memory system to people. They presented their participants with easy or hard trivia questions, followed by a color-naming task with either general target words or target words associated with computers. The results of this experiment show that participants, on average, had larger reaction times when naming the color of the computer words as opposed to the general words, implicating that people associate computers with external memory systems, able to provide us with answers.

Another recent usage of the SPT is that of Schmettow, Noordzij & Mundt (2013), although they followed a different approach than Warren (1972; 1974) and Sparrow *et al.* (2011) by using *pictures* instead of words as priming stimuli. The researchers were interested in *geekism*, defined by them as "a predisposition that we associate with great affinity for exploring and tinkering with technological devices." (p. 2040), which they claim to be an important part of user experiences of products (as well as hedonism and usability). Like in the current study, Schmettow *et al.* (2013) deliberately opted for an implicit method rather than the more common (explicit) self-report measures. In order to find out whether words associated with geekism (as well as hedonism and usability) would produce interference in color-naming tasks, they provided their participants with a SPT that consisted of 15 black-and-white pictures of technological devices, and 90 target words that were associated with either geekism, hedonism or usability. Results indicated that participants with a higher predisposition of geekism did indeed show longer response latencies on color-naming tasks with geekism target words, which led Schmettow and colleagues (2013) to conclude that these participants show stronger associations with geekism words as opposed to other participants.

The study of Schmettow *et al.* (2013) clearly indicates that priming pictures in a SPT can be used as well as priming words to test whether semantic associations exist between the priming stimuli and the target words. The current study follows this line of thinking in the context of first impressions on beauty judgments of website users. A SPT will be used that contains shortly appearing screenshots of website homepages as priming stimuli and target words associated with either beauty, ugliness or neutrality. When viewing these screenshots in the short period they appear, participants should form implicit judgments whether they find the screenshots aesthetically appealing or not. Consequently, in the color-naming tasks, participants, now primed to think of beauty, should show longer response latencies when naming the color of target words they associate with beauty as opposed to neutral target words. In this way, an implicit measurement of first impression beauty judgments should be realized.

1.4 Research questions and hypotheses

To summarize, the aim of this study was twofold. First, this study aimed to find out whether a SPT could be a suitable alternative to the more common explicit measures in HCI research on website beauty judgments. Second, the aim was to shed more light on the relationships between color and VC, PT and beauty judgments. An experiment was conducted to study these relationships, using the SPT for the implicit measurement.

The *main research question* used in this study was therefore as follows: "How can the relationships between color, visual complexity, prototypicality and the judged beauty of websites in website users' first impressions be implicitly measured?" Naturally, this research question can be divided into two sub questions, each corresponding to one of the two aims of this study.

The *first sub question* was as follows: "Which method can be used to measure the judged beauty of websites in website users' first impressions implicitly?" As mentioned, in this study a SPT was used to find out whether it is a suitable implicit measurement method in this context. It was hypothesized that the SPT will be able to accurately measure the first impression beauty judgments by showing semantic associations between words associated with beauty and website screenshots judged to be beautiful. Websites that are judged to be more beautiful should, as compared to neutral target words, lead to longer response latencies for 'beauty' target words and shorter response latencies for 'ugliness' target words.

And vice versa for websites that are judged to be ugly. In the following chapter the details about the procedure of the experiment will be described.

The *second sub question* was as follows: "What are the relationships between color, visual complexity, prototypicality and the judged beauty of websites in website users' first impressions?" As mentioned, through the experiment with a SPT it is intended to shed more light on these effects. In line with the earlier mentioned theoretical background (Moshagen & Thielsch, 2010), it is hypothesized that colored websites, as compared to websites in grayscale, will be judged to be most beautiful, because of the positive correlations between colorfulness and aesthetics.

2. Method

2.1 Participants

A total of 33 participants took part in the experiment, of which 14 were male and 19 were female. Their age ranged from 17 to 63, with a mean of 31.36 and a standard deviation of 15.47. All participants were Dutch, except for 6 German individuals, who were however proficient in the Dutch language. No deficiencies in recognizing colors or problems related to perceiving colors during the experiment were reported by any of the participants.

Informed consent was obtained for all participants through a short introduction by the author, and the signing by the participants of a form that contained a description of their rights and the global procedure of the experiment. The forms of the (two) 17-year-old participants were signed by one of their parents as well.

2.2 Materials

The materials used during the experiment can roughly be divided into three parts: the target words, the priming stimuli and the SPT.

2.2.1 Target words

72 Dutch words were selected by the author as target words for usage during the experiment. These words were categorized into association with one of three concepts: beauty, ugliness and neutrality. After the selection by the author, all words were put into a random order and presented to another Psychology student for a second, independent classification. The inter-rater reliability was with a Cohen's Kappa of .91 rather high.

The beauty category contained words such as ‘prachtig’ (gorgeous), ‘attractief’ (attractive) and ‘aansprekend’ (appealing), while the ugliness category contained words such as ‘pover’ (poor), ‘onaantrekkelijk’ (unattractive) and ‘onaanzienlijk’ (unsightly). The third category, neutrality, contained words such as ‘bord’ (plate), ‘lopen’ (to walk) and ‘zand’ (sand). Appendix I contains an overview of all the used Dutch target words and their meaning in English.

2.2.2 Priming stimuli

The priming stimuli were screenshots from website homepages, borrowed from Tuch *et al.* (2012). The borrowed dataset consisted of 119 screenshots of homepages, each rated by participants (Tuch *et al.*, 2012) across two dimensions: VC (low, medium or high) and PT (low or high). To allow for better discrimination between VC levels, the screenshots rated with a medium VC level were omitted from the dataset, resulting in a total of 80 screenshots. Since 72 target words were selected, and one-on-one combinations between target words and priming stimuli were used, only that same amount of screenshots was needed for the experiment. The eight screenshots with the highest standard deviations were therefore also omitted from the dataset.

The resulting 72 website screenshots were then divided into eight categories, across the dimensions VC, PT and CL. Table 1 gives an overview of the characteristics of these categories. ‘Low’ VC ratings ranged from 2.67 to 3.77 and ‘High’ VC ratings ranged from 4.46 to 5.93, while ‘Low’ PT ratings ranged from 2.32 to 4.54 and ‘High’ PT ratings ranged from 4.55 to 5.55 (see Tuch *et al.* (2012) for more details on these ratings). The screenshots in the ‘Colored’ categories had the default colors of Tuch *et al.* (2012), while the screenshots in the ‘Grayscale’ categories were manually set to grayscale by the author using the program Microsoft Office Picture Manager. Overall, all 72 priming stimuli were screenshots of 72 different websites.

Table 1. *Priming Stimuli Categories' Characteristics.*

Category name	VC level	PT level	CL	Stimuli amount
HHH	High	High	Colored	9
HHL	High	High	Grayscale	9
HLH	High	Low	Colored	9
HLL	High	Low	Grayscale	9
LHH	Low	High	Colored	9
LHL	Low	High	Grayscale	9
LLH	Low	Low	Colored	9
LLL	Low	Low	Grayscale	9

2.2.3 Stroop Priming Task

A custom SPT, developed with the 1.76.00 version of the open source software *PsychoPy* (2013), was used during the experiment. In each task, the 72 priming stimuli were divided into six blocks of twelve trials. Each of the 72 target words were randomly assigned to one of the six blocks. Per participant, the order of appearing of the six blocks, and the order of appearing of the target words and the priming stimuli *within* a block (and therefore their resulting combinations), were randomized.

2.3 Procedure

The experiments were carried out on a computer with PsychoPy (2013) installed. During the experiments, each participant was seated alone in a small room containing this computer. After welcoming the participant, the author gave a short introduction before starting the experiment. During this introduction the goal of the experiment was described as a means to find out how fast people are in the recognition of colors. This was followed by a brief description of the general procedure, after which the participants were pointed at their rights and they could sign the informed consent form. After answering any possible questions of the participant, the author then started the experiment by running the SPT with PsychoPy.

To start off the experiment, instructions describing the general procedure and the details of the required response were shown on screen. After reading these instructions, participants had to complete 20 test trials to get familiar with the procedure and the response keys. Each trial had a grayscale picture of fruit as priming stimulus and a non-existing word as target word. Each priming stimuli was presented for 3 seconds. The target words randomly appeared in either red, green or blue, and the participants could respond, within an interval of 4 seconds, with either the 'left' key for red, the 'down' key for green or the 'right' key for blue. After each response, immediate feedback about the accuracy and speed of the response was shown on the screen.

After the test trials, another instruction screen was shown, which repeated the most important instructions and notified the participants that the main experiment was about to start. It also notified them that the trials would be interrupted by a few short breaks.

The main experiment used the setup described in section 2.2.3. In between each of the six blocks a break was inserted during which a grayscale picture of a beach was presented for 30 seconds. After each break, the participants could continue with the next block by pressing any key. Each priming stimulus was presented for 3 seconds, after which the participants could respond, within an interval of 3 seconds, with either the 'left', 'down' or 'right' key for 'red', 'green' or 'blue' respectively. No immediate feedback was given during the main experiment. After finishing the experiment, the participants were thanked for their participation, and any questions they posed were answered by the author.

2.4 Analysis

The data of each experiment was automatically saved as a Microsoft Excel file (*.csv*), which was then manually converted with the open source software *R* (R Core Team, 2012) into a *.sav*-format. Further analysis was then performed with *IBM SPSS Statistics 21* (IBM, 2013).

First, incorrect responses were omitted from the dataset, followed by a z-score conversion of the VC and PT ratings. After this, a linear *Generalized Estimating Equations (GEE)* statistical procedure was used, with the response times (RT) as dependent variable, CL and the target word categories as factors, and VC and PT as covariates. The participants were set as the subject variable, with an exchangeable working correlation matrix. The used model tested the main effects of VC, PT, CL and the target word categories, and the interaction effects of the target word categories with VC, PT and CL. Using the output of the GEE, graphs were made to make potential interaction effects visual.

3. Results

Overall, a total of 2304 response times were collected during the experiment. Unfortunately, the data of one of the participants was lost during a save, resulting in a total of 32 participants. In 107 instances, a participant pressed a wrong key (*mean* = 3.34, *SD* = 2.06), and these responses were therefore omitted from the data for analysis, resulting in a total of 2197 response times.

Figures 2 and 3 show the differences in response times between low and high VC and PT respectively, divided into the three target word categories beauty (B), ugliness (U) and neutrality (N). Figure 4 shows the differences in response times between the trials with grayscale priming stimuli and the trials with colored priming stimuli, again divided into the three target word categories.

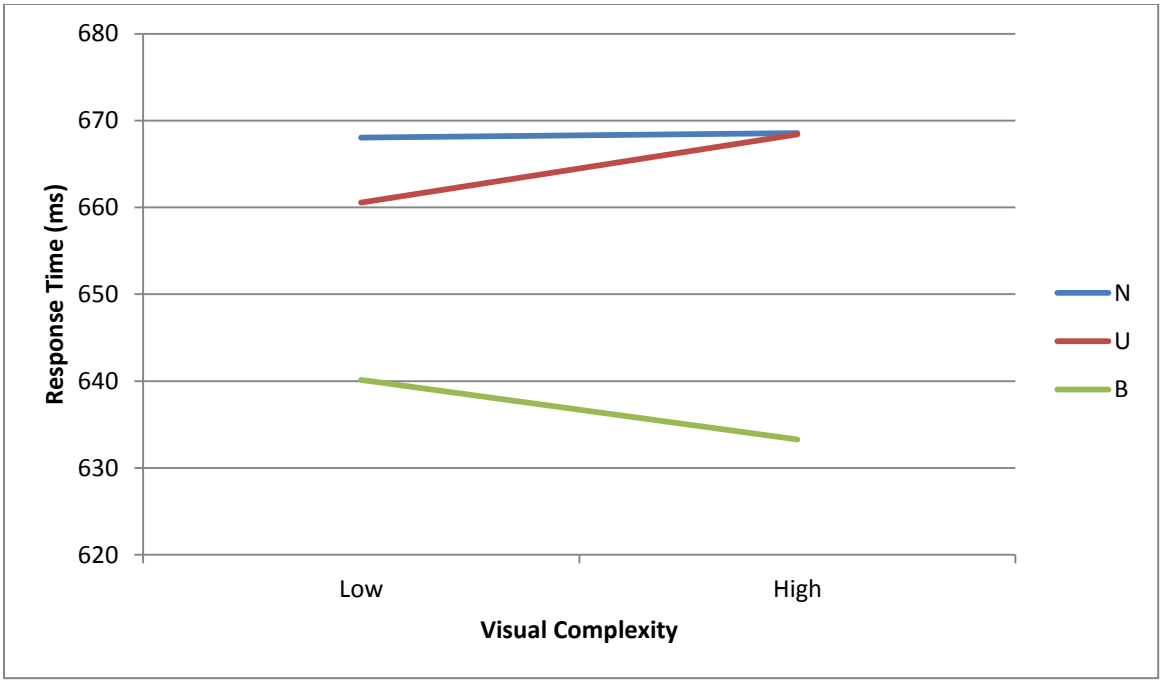

Figure 2. Response Times to Beauty, Ugliness and Neutrality Target Words for Low and High Visual Complexity

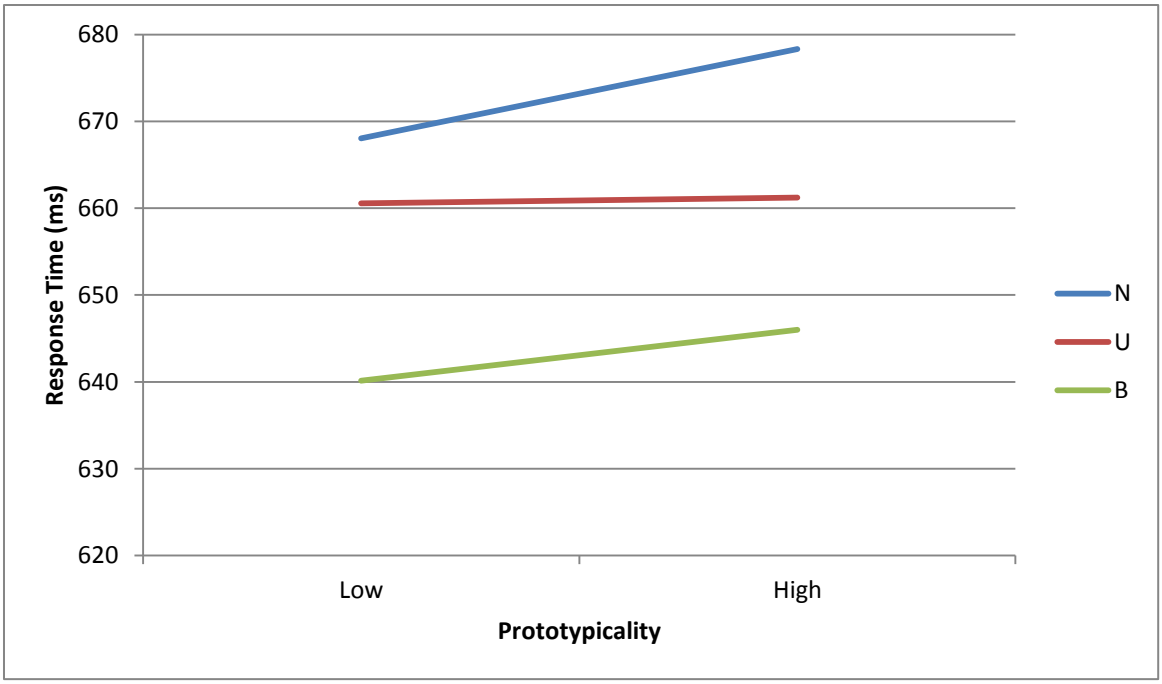

Figure 3. Response Times to Beauty, Ugliness and Neutrality Target Words for Low and High Prototypicality

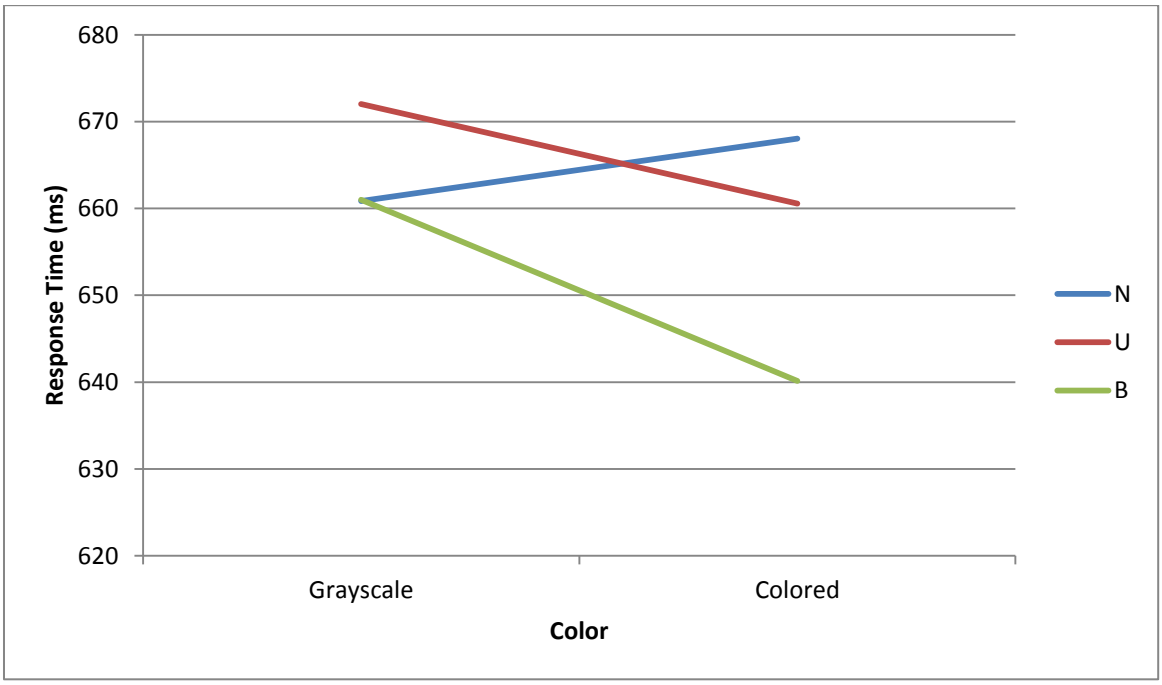

Figure 4. Response Times to Beauty, Ugliness and Neutrality Target Words for Grayscale and Colored Priming Stimuli

Overall, no significant differences were found between the three target word categories (X^2 (2) = 2.329, p = .312), nor between the grayscale and colored priming stimuli (X^2 (1) = 1.212, p = .271). There were also no significant differences between the low and high levels of VC (X^2 (1) = .013, p = .909) and PT (X^2 (1) = 1.858, p = .173). Lastly, no significant interaction effects were found between the target word categories and either VC (X^2 (2) = 3.510, p = .173), PT (X^2 (2) = .635, p = .728) or CL (X^2 (2) = 3.217, p = .200).

4. Conclusions & Discussion

This study aimed to find out whether the Stroop Priming Task could be a suitable alternative to explicit methods for measuring first impression beauty judgments, as well as to shed more light on the relationships between (the absence of) color, visual complexity, prototypicality and the judged beauty of websites. A SPT was used to test these relationships, with target words semantically related to beauty, neutrality and ugliness, and screenshots of website homepages, varying in VC, PT and CL, as priming stimuli. Generally, from the results of this study, it can be concluded that the used SPT was not able to accurately measure the judged beauty of websites in website users' first impressions. Also, the results indicate that the absence or presence of color in websites does not significantly influence the first impression beauty judgments of website users.

4.1 Stroop Priming Task

The first hypothesis of this study was that the SPT would be able to accurately measure the first impression beauty judgments by showing semantic associations between words associated with beauty and website screenshots judged to be beautiful. If this hypothesis were to be true, websites that are judged to be beautiful should have longer response times during the SPT for 'beauty' target words and shorter response times for 'ugliness' target words, as compared to neutral target words. Websites that are judged to be ugly, on the other hand, should have longer response times for 'ugliness' target words and shorter response times for 'beauty' target words.

Unfortunately, the results do not support this hypothesis. The results showed that the response times of the three target word categories did not differ significantly. Naturally, this means that the websites judged to be beautiful did not have significantly longer response times for 'beauty' target words than for 'ugliness' target words, and websites judged to be ugly did not have significantly longer response times for 'ugliness' target words than for 'beauty' target words. Furthermore, this means that the response times for 'beauty' and 'ugliness' target words did not significantly differ from the response times for the neutral target words, which were considered to be the baseline to which the 'beauty' and 'ugliness' response times could be compared. This implies that the SPT, in the form used in this study, is not a valid measure for measuring the first impression beauty judgments of website users.

The question remains, however, *why* the expected results of the SPT were not found. One possible explanation may lie in the low test-retest stability of implicit measures (Robinson & Neighbors, 2006). It is quite arguable that the results found in this study will not be found in another study that uses the same experimental setup. This implies that multiple replications of this study are needed before conclusions based on sufficient data can be drawn. Furthermore, as Robinson and Neighbors (2006) pointed out, the low test-retest stability is due to unstable implicit processes in the individual. Therefore, a sufficient amount of participants should lower the variance in response times and show a regression towards the true mean of the population. However, a quick statistical analysis using only the first 16 participants instead of all 32, showed an *increase* rather than decrease in difference between target word categories ($X^2 = 3.553$, $p = .169$), making this explanation doubtful.

Another possible explanation is that people do not process the beauty of websites semantically, but by other means instead. In other words, when someone views a website and in his first impression forms an aesthetic judgment about it, they might not form a semantically relevant *concept* of beauty. This explanation has some support from Leder *et al.* (2004), who argued that in order to understand an object semantically, people need to look at its *content*. According to their model of aesthetic experiences, content is processed in the third stage. Since this is a *deliberate* stage, as opposed to implicit, it makes sense that in the SPT the content of the website screenshots was not processed in the short duration they appeared. Thus, while people form implicit beauty judgments of websites in their first impressions, it seems they do not form semantic understandings of the concept of beauty due to the fact that the website's contents are processed *after* their first impressions. This might explain why the SPT, which makes use of semantic associations, was not able to measure beauty judgments in the participant's first impressions.

As can be seen in figures 2, 3 and 4, the response times for neutral target words tended to be higher than 'beauty' target words *as well as* 'ugliness' target words, rather than being in the middle of the three categories, as was expected. One possible explanation for this might be that the neutral target words were found to be more *familiar* by the participants than the 'beauty' or 'ugliness' target words. The neutral target words were mostly names of very common objects, such as a plate, a pencil or a chair, and should therefore be very familiar to the participants. The 'beauty' and 'ugliness' target words, however, were mostly synonyms. It is possible that a large portion of these synonyms are rarely used, and therefore are less familiar to the participants. In effect, it is possible that some target words of 'neutrality' were, due to being more familiar, actually *less* neutral than some target words of 'beauty' and 'ugliness'. This limitation might have led to the activation of certain concepts that interfered with the color-naming tasks of the 'neutrality' target words, leading to higher response times than was intended.

The fact that the SPT in this study was not a valid measure for first impression beauty judgments of website users does not mean it is an invalid tool in general. Indeed, Schmettow *et al.* (2013), who essentially used the same SPT setup as the current study, successfully showed that target words that were semantically associated with prior shown priming stimuli had longer response latencies than those that were *not* semantically associated.

However, while the setup of SPT was the same in both studies, the target word categories were altogether different. Schmettow *et al.* (2013) used target words for three very distinct concepts ('hedonism', 'usability' and 'geekism'), while the current study made use of three different concepts, of which two ('beauty' and 'ugliness') are related in the sense that they are antonyms. It is very well possible that this distinctness in the target word categories of Schmettow *et al.* (2013), which seems to be less in the current study, allowed for more distinct differences in response times in the SPT.

Apart from the target words, Schmettow *et al.* (2013) also used different priming stimuli in their experiment. They primed their participants with simple pictures of technology, which contained only an image of a technical device, without text or other elements. The screenshots of website homepages that were used in the current study, however, all contained a multitude of elements such as text and images. These screenshots were therefore to a large extent more visual complex than the pictures used by Schmettow *et al.* (2013). This was also recognized by Tuch *et al.* (2012), who argued that websites are in itself complex stimuli and are therefore automatically placed on the right end of Berlyne's inverted U-shaped curve (1974, as cited in Tuch *et al.*, 2012).

Since Tuch *et al.* (2012) found that websites with high VC generally receive lower beauty judgments, this might explain why response times for 'beauty' target words were generally the lowest.

4.2 Color, visual complexity and prototypicality

For the second goal of this study, it was hypothesized that colored websites would be judged to be more beautiful by participants in their first impressions than websites in grayscale. If this hypothesis were to be true, colored websites should have longer response times during the SPT for 'beauty' target words and shorter response times for 'ugliness' target words, as compared to 'neutrality' target words. Websites in grayscale should show reverse results, with longer response times for 'ugliness' target words and shorter response times for 'beauty' target words.

However, the results do not support this second hypothesis either. They indicate that the absence or presence of color in websites do not cause participants to judge these websites differently on attractiveness in their first impressions, since no significant differences were found between the response latencies of colored websites and the response latencies of websites in grayscale, and no significant interaction effect was observed between CL and the three target word categories. Furthermore, although a small interaction trend between CL and the target word categories can be observed in figure 4, the trend is opposite to expectation: colored websites have *lower* response latencies for 'beauty' and 'ugliness' target words than websites in grayscale.

A possible explanation for the absence of significant difference between colored websites and websites in grayscale could be related to the way people process these websites when judging their attractiveness. It is conceivable that when people judge a grayscale object's attractiveness, they do not take color into account in this judgment, since color, then, is not part of the visual aspects of this object. Should the object be colored, then color *is* part of the visual aspects of this object, meaning that people would take color into account in their beauty judgment. However, this is merely a hypothesis, as no scientific literature has been found to support this notion.

Apart from the results regarding the color of the websites, the results of the website characteristics visual complexity and prototypicality were not conform results from earlier studies, as described in sections 1.1.1 and 1.1.2, either. Most notable were the results of Tuch *et al.* (2012), who found that people tend to give higher perceived beauty ratings to websites when their VC is low (or medium) rather than high, and also when their PT is high rather than low. Also, they found that these two combined (websites with simultaneously low VC and high PT) were given the highest perceived beauty ratings. None of these findings by Tuch *et al.* (2012) were however confirmed by the present study, since no significant differences were observed between different levels of VC and PT, and no significant interactions were found between the three target word categories and either VC or PT.

The graphs of figures 2 and 3 are not conform expectations either. They tend to show the longest response times for 'neutrality' target words, the shortest response times for 'beauty' target words, and the response times for 'ugliness' target words in between those two. Only a small interaction trend between VC and the three target word categories can be observed in figure 1, as the response times for 'ugliness' target words were somewhat higher for websites with high VC than with low VC. While this is conform the finding of Tuch *et al.* (2012) that websites with high VC are judged lower on beauty, the interaction trend was, as mentioned, not significant.

A possible explanation for the lack of significant difference in response times between the two VC levels might be found in the earlier mentioned complexity of websites. Since all websites are complex stimuli, and websites with high VC have lower perceived beauty ratings (Tuch *et al.*, 2012), it is conceivable that it might be hard for participants to distinguish between different levels of VC, leading to little differences in response times.

4.3 Further research

As mentioned in section 4.1, it is arguable that people do not form semantic understandings of the concept of beauty in their first impressions of websites, implicating that the SPT cannot be used for measuring first impression beauty judgments in *general*. However, a few different explanations were offered as well, which need to be tested before such a conclusion is being drawn.

Further research on this topic should therefore try to use the SPT setup described in this paper, with a few modifications to overcome the present limitations. First, the SPT should contain a longer block of test trials to better train participants to associate the 'left', 'right' and 'down' keys with red, green and blue respectively. This is due to the fact that a multitude of participants reported problems with remembering which key belonged to which color, even during the main experiment. Also, to the opinion of the author, quite a large proportion of the response times were to be omitted due to a hit of the wrong key. Second, a thorough validation of target words should be conducted prior to the experiment itself, to avoid familiarity effects and to make sure the different target word categories are distinct concepts. Third, the SPT should contain priming stimuli that are less visual complex to avoid lower beauty judgments beforehand. Possibly, it should be considered to *not* use website screenshots at all as priming stimuli, since they are all inherently complex (Tuch *et al.*, 2012), but to use basic images such as the ones in Schmettow *et al.* (2013) instead. In this way, it can be tested whether first impression beauty judgments can be measured with a SPT at all. Should this be the case, attempts can then be made to switch back to more complex stimuli such as websites. Lastly, it may be considered to omit VC and PT as variables, and to only use grayscale and colored priming stimuli as experimental conditions, in order to get a 'clean' measure of the effect of color on beauty judgments.

Last but not least, it will be interesting to find out whether the stated hypothesis in section 4.2 is true or not. Namely, if it is true whether people only use color as a factor for their beauty judgments when the object to be judged *is* colored, and do not take color into account when this object is in grayscale.

Acknowledgements

The author would like to thank those who made this study possible, including Martin Schmettow for his support, Alexandre Tuch for lending the website screenshots dataset, and all participants for taking part in the experiment.

References

Bargas-Avila, J. A. & Hornbaek, K. (2011). Old Wine in New Bottles or Novel Challenges? A Crititcal Analysis of Empirical Studies of User Experience. *CHI '11. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 2689-2698. doi: 10.1145/1978942.1979336

Cohen, L., Manion, L. & Morrison, K. (2011). *Research Methods in Education: Seventh Edition.* 2 Park Square, Milton Park, Abingdon, Oxon: Routledge.

Cohen, R. J. & Swerdlik, M. E. (2010). *Psychological Testing and Assessment: Seventh Edition*. 1221 Avenue of the Americas, New York: McGraw-Hill.

Geissler, G. L., Zinkhan, G. M. & Watson, R. T. (2006). The Influence of Home Page Complexity on Consumer Attention, Attitudes, and Purchase Intent. *Journal of Advertising, 35 (2),* 69-80. doi: 10.1080/00913367.2006.10639232

Hall, R. H. & Hanna P. (2004). The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & Information Technology, 23 (3),* 183-195. doi: 10.1080/01449290410001669932

Hassenzahl, M. (2004). The Interplay of Beauty, Goodness, and Usability in Interactive Products. *Human-Computer Interaction, 19,* 319-349. doi: 10.1207/s15327051hci1904_2

Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human–Computer Interaction*, *25*(3), 235–260. doi: 10.1080/07370024201050013 9

IBM. (2013). *IBM SPSS Statistics*. Retrieved from: http://www-01.ibm.com/software/analytics/spss/products/statistics/

Klein, G. S. (1964). Semantic Power Measured Through the Interference of Words with Color-Naming. *The American Journal of Psychology, 77 (4),* 576-588. Retrieved from: http://www.jstor.org/stable/1420768

Lavie, T. & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies, 60 (3),* 269-298. doi: 10.1016/j.ijhcs.2003.09.002

Leder, H., Belke, B., Oeberst, A & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95*, 489-508. doi: 10.1348/0007126042369811

Lindgaard, G., Fernandes, G., Dudek, C. & Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology, 25 (2),* 115-126. doi: 10.1080/01449290500330448

MacLeod, C. M. (1991). Half a Century of Research on the Stroop Effect: An Integrative Review. *Psychological Bulletin, 109 (2),* 163-203. doi: 10.1037/0033-2909.109.2.163

Martindale, C., Moore, K. & Borkum, J. (1990). Aesthetic preference: Anomalous findings for Berlyne's psychobiological theory. *American Journal of Psychology, 103 (1),* 53-80. Retrieved from http://www.jstor.org/stable/1423259

Moshagen, M. & Thielsch, M. T. (2010).

Pandir, M. & Knight, J. (2006). Homepage aesthetics: The search for preference factors and the challenges of subjectivity. *Interacting with Computers, 18 (6),* 1351-1370. doi: 10.1016/j.intcom.2006.03.007

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In Braun, H. I., Jackson, D. N. & Wiley, D. E. (Ed.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Paulhus, D. L. & John, O. P. (1998). Egoistic and Moralistic Biases in Self-Perception: The Interplay of Self-Deceptive Styles With Basic Traits and Motives. *Journal of Personality, 66 (6),* 1025-1060. doi: 10.1111/1467-6494.00041

PsychoPy. (2013). Retrieved from: http://code.google.com/p/psychopy/downloads/list

R Core Team. (2012). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: http://www.R-project.org

Robinson, M. D. & Neighbors, C. (2006). Catching the mind in action: Implicit Methods in Personality Research and Assessment. In Eid, M. & Diener, E. (Ed.), *Handbook of multimethod measurement in psychology* (pp. 115-125). Washington DC, US: American Psychological Association.

Roth, S. P., Schmutz, P., Pauwels, S. L., Bargas-Avila, J. & Opwis, K. (2010). Mental models for web objects: Where do users expect to find the most frequent objects in online shops, news portals, and company web pages? *Interacting with Computers, 22*, 140-152. doi: 10.1016/j.intcom.2009.10.004

Schmettow, M., Noordzij, M. L. & Mundt, M. (2013). An Implicit Test of UX: Individuals Differ in What They Associate with Computers. *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 2039-2048. doi: 10.1145/2468356.2468722

Sparrow, B., Liu, J. & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science, 333 (6043),* 776-778. doi: 10.1126/science.1207745

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18 (6),* 643-662. doi: 10.1037/h0054651

Tractinksy, N., Cokhavi, A., Kirschenbaum, M. & Sharfi, T. (2006). Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human-Computer Studies, 64*, 1071-1083. doi: 10.1016/j.ijhcs.2006.06.009

Tuch, A. N., Presslaber, E. E., Stöcklin, M., Opwis, K., & Bargas-Avila, J. a. (2012). The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International Journal of Human-Computer Studies*, *70*(11), 794–811. doi: 10.1016/j.ijhcs.2012.06.003

Veryzer, R. W., Jr. & Hutchinson, J. W. (1998). The Influence of Unity and Prototypicality on Aesthetic Responses to New Product Designs. *Journal of Consumer Research, 24 (4),* 374-385. doi: 10.1086/209516

Warren, R. E. (1972). Stimulus Encoding and Memory. *Journal of Experimental Psychology, 94 (1),* 90-100. doi: 10.1037/h0032786

Warren, R. E. (1974). Association, Directionality, and Stimulus Encoding. *Journal of Experimental Psychology*, *102 (1),* 151-158. doi: 10.1037/h0035703

Wickens, C. D., Lee, J. D., Liu, Y. & Gordon Becker, S. E. (2004). *An Introduction to Human Factors Engineering: Second Edition.* Upper Saddle River, New Jersey: Pearson Prentice Hall.

Xing, J. & Manning, C. A. (2005). *Complexity and Automation Displays of Air Traffic Control: Literature Review and Analysis* (Final Report). Washington DC: U.S. Department of Transportation, Office of Aerospace Medicine.

Appendix I

Table 1. *'Beauty' Target Words in Dutch and their English Translations*

Target Word	**English Translation**
'Beauty'	
Schitterend	Brilliant
Aansprekend	Appealing
Stijlvol	Stylish
Gaaf	In good order
Verzorgd	Taken care of
Keurig	Neat
Aanlokkelijk	Alluring
Mooi	Beautiful
Netjes	Neatly
Prachtig	Gorgeous
Leuk	Nice
Elegantie	Elegance
Attractief	Attractive
Verleidelijk	Tempting
Smaakvol	Tasteful
Knap	Handsome
Jofel	Nice
Aantrekkelijk	Attractive
Bekoorlijkheid	Charm
Bevallig	Graceful
Fraai	Beautiful
Sierlijk	Gracefully
Schoonheid	Beauty
Uitnodigend	Inviting

Table 2. *'Neutrality' Target Words in Dutch and their English Translations*

Target Word	**English Translation**
'Neutrality'	
Bord	Plate
Glas	Glass
Lopen	To walk
Water	Water
Schaar	Scissors
Papier	Paper
Zeep	Soap
Raam	Window
Stoel	Chair
Fles	Bottle
Knop	Button
Hout	Wood
Melk	Milk
Wind	Wind
Potlood	Pencil
Snoer	Wire
Metaal	Metal
Schroef	Screw
Ijzer	Iron
Handdoek	Towel
Lezen	To read
Pen	Pen
Tafel	Table
Zand	Sand

Table 3. *'Ugliness' Target Words in Dutch and their English Translations*

Target Word	**English Translation**
'Ugliness'	
Onbekoorlijk	Unappealing
Slordig	Slovenly
Monsterlijk	Monstrous
Onaantrekkelijk	Unattractive
Pover	Poor
Sjofel	Seedy
Onooglijk	Unsightly
Karakterloos	Flabbily
Onaanzienlijk	Unsightly
Stuitend	Disgusting
Onverzorgd	Untended
Armelijk	Poorly
Haveloos	Shabby
Stijlloos	Tastelessness
Misvormd	Malformed
Schraal	Poor
Armoedig	Poor
Wanstaltig	Misshapen
Karig	Scanty
Smakeloos	Tasteless
Afzichtelijk	Unsightly
Mismaakt	Misshapen
Lelijk	Ugly
Afstotend	Repulsive