The Pretty and the Useful:
Effects of Aesthetics and Usability on Mobile Webshop Evaluation

Laura Quintana, B.Sc.

Department of Psychology

University of Basel

Submitted in October 2017

Supervisors:

Florian Brühlmann M.Sc.

Prof. Dr. Klaus Opwis

Author Note

Laura Quintana Gomez

Schlüsselgasse 4, 4102 Binningen

Immatriculation number: 11-056-959

l.quintanagomez@unibas.ch

Declaration of scientific integrity:

The author hereby declares that she has read and fully adhered the Code for Good Practice in Research of the University of Basel.

Abstract

Previous research has examined the overall evaluation of websites on desktop computers with correlations of aesthetics, usability and content at different points in time. However, these correlations do not allow detection of causal effects and findings are limited to one type of device. Therefore, this thesis examined the influence of aesthetics and usability on user evaluations of mobile webshops at two points in time experimentally, while content was held constant. Data of 93 participants (18 to 54 years, 55% males) was collected in a laboratory experiment with a $2 \times 2 \times 2$ mixed design. They rated one of four mobile webshop versions after one second exposure time and after longer interaction with respect to perceived aesthetics, usability, content, affect, first and overall impression as well as the intention to revisit and recommend the webshop. Findings showed that aesthetic's halo effect on overall evaluative ratings weakened after interactive exposure. This effect was more pronounced when usability was low. Content and affect were also influenced by the aesthetics of the webshop. As on desktop devices, ratings showed similar stability between the two evaluation times, but a differing evaluation pattern emerged with usability playing a bigger role on mobile. This study contributes to identifying similarities and differences across devices. The manipulations used in this study suggest design implications and thus offer a commercial use for practitioners in mobile-commerce.

The Pretty and the Useful:
Effects of Aesthetics and Usability on Mobile Webshop Evaluation

Introduction

The average global smartphone penetration (i.e. the amount of sales compared to the total theoretical market of a product) reached 66% in the beginning of 2017 and is growing continuously (Kemp, 24.01.2017). Y.-K. Lee, Chang, Lin, and Cheng (2014) stated that smartphones moved from being cutting-edge communication gadgets to necessities in people's lives. As a consequence, they present opportunities and challenges, such as increased access, portability, personal space, opportunistic interaction, and reduced complexity (Billi et al., 2010). These opportunities lead people to use mobile devices for shopping purposes more often (Wang, Malthouse, & Krishnamurthi, 2015). Electronic commerce (e-commerce) via mobile, also called m-commerce, has grown rapidly in recent years together with the popularity of smartphone usage (Chang, Williams, & Hurlburt, 2014). Wang et al. (2015) found that, especially for low-spending customers, order rate and size (e.g., amount in dollars) increased with m-commerce accustomization.

Although m-commerce is growing, only little research has been conducted in this field. But the rapid evolution of m-commerce requires practitioners to continually update their knowledge about it to gain a competitive advantage (Seth, 2014), making knowledge about the evaluation of mobile webshops a key point of interest. Evaluation refers to the perception and integration process of constructs building up a website's high-level impressions (e.g., first and overall impression). On desktop, aesthetics, usability, and content have been identified as core constructs (Cober, Brown, Levy, Cober, & Keeping, 2003; Schenkman & Jönsson, 2000; Tarasewich, Daniel, & Griffin, 2001), and the impact of each construct on overall evaluations of different types of websites has been studied by Thielsch, Blotenberg, and Jaron (2013). In a correlational path model, they displayed how highly said constructs correlated with first and overall impression as well as intention to revisit and recommend a website. While aesthetics had the biggest influence at first and overall impression, and usability played a significant but subdued role, content was identified as the most important construct for intentions to revisit and recommend. Mobile website evaluation has been neglected in previous research.

Comparing the desktop model to mobile devices with findings from an experiment allows the measure of causal effects on overall evaluations and brings insights regarding device specificity. For example, Chittaro (2006) found that for information visualisation it is not possible to transfer knowledge from desktop to mobile due to technical limitations. Although mobile devices are well equipped with functions similar to desktop computers, the interaction differs due to reduced screen size, context of use and interaction modality (Sohn, Seegebarth, & Moritz, 2017), making research across devices important.

The aim of this thesis is to investigate the evaluation of four versions of a mobile webshop selling books with aesthetics and usability experimentally manipulated. While content was held constant, affect was measured to investigate the evaluation process more integratively. The influence of the manipulations on overall evaluative and construct-specific measures at different points in time bring new insights that contribute to research in several ways. First, constructs playing a similar or diverging role for website evaluation on desktop compared to mobile are identified to investigate device specificity of findings. Second, manipulating aesthetics and usability generates knowledge about causal effects on construct-specific and overall evaluative ratings. Third, the stability of evaluation is investigated by the repeated measure of overall ratings, which has not been done so far on mobile devices. Finally, the manipulations applied in this study give rise to guidelines for practitioners in m-commerce in terms of design of colors, logos, and information architecture.

Theoretical Background

Over 1.61 billion individuals are using the Internet for shopping purposes, resulting in a penetration rate of 22% of the global population which has internet access (Kemp, 24.01.2017). As Benou and Bitos (2008) pointed out, the increase of smartphone users resulted in the extension and development of e-commerce to m-commerce. For example in South Korea, 55% of purchases in e-commerce come from mobile (Kemp, 24.01.2017). Consequently, it has become crucial to understand the perception and integration process of constructs building up user evaluation of mobile webshops. Several models differentiate three phases of website interaction: initial exposure and impression formation, evaluation and use

phase, and intentional outcomes (Metzger, 2007; Ou & Sia, 2010). Information processing starts within milliseconds of visiting a website and impression building begins (Thielsch & Hirschfeld, 2012; Tuch, Presslaber, Stöcklin, Opwis, & Bargas-Avila, 2012). These first impressions are based on bottom-up processes of visual perception, described for example in the model of aesthetic processing proposed by Leder, Belke, Oeberst, and Augustin (2004). In contrast, deliberate impressions are top-down driven and result from cognitive processes and reasoning during website evaluation (Leder et al., 2004).

In the following, the constructs playing a key role in the evaluation process are defined and findings for mobile are reviewed separately. Subsequently, the model of website evaluation on desktop by Thielsch et al. (2013) is presented. Finally, the study conducted in this work is introduced.

Important constructs of evaluation

Content. Content is defined as 'a set of interactive or non-interactive objects containing information represented by text, image, video, sound or other types of media on a web user interface' (ISO, 2006, p. 3). Huizingh (2000) provided a framework distinguishing web content from design. The former refers to information features or services offered, whereas the latter defines how content is made available for users. Content can be subjectively experienced and was found to be the most important construct for website evaluation and success, being the main reason why websites are looked at in the first place (Palmer, 2002; Thielsch et al., 2013). Content suitable for e-commerce has been studied by Porat and Tractinsky (2012). They collected ratings for thirteen product domains on the basis of the requirement to touch the product before purchase, as well as perceived cost of the product and likelihood of buying it online. Books were rated with a low need to touch before buying, as a low-priced product and having a high probability of being bought online, making it a good product to test webshop evaluations (Porat & Tractinsky, 2012).

Usability. Usability is defined as the 'extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use' (ISO, 1998, p. 2). Thus, it is shaped by the interaction of people,

tools and problems (Naur, 1965). Previous research has focused on the key question of how to work with and enhance the usability of interactive systems (Hornbæk, 2006). Results include guidelines to improve usability (Smith & Mosier, 1986) and different different methods of measuring it (Frøkjær, Hertzum, & Hornbæk, 2000; ISO, 1998; Nielsen & Levy, 1994). A differentiation between objective and subjective usability measures is made (Hornbæk, 2006). Objective usability measures for example the time needed to complete a task, clicks needed and navigational efforts. In contrast, subjective usability measures for instance the learnability of a program or ease of use. Usability manipulations can affect both types of measures simultaneously. A change in the information architecture (IA) can lead to different navigational patterns, measured objectively, and impact perceived wayfinding as well as ease of use, measured subjectively. On desktop, Tuch, Roth, Hornbæk, Opwis, and Bargas-Avila (2012) used different labels manipulating the information scent of a webshop's IA, affecting navigation and search patterns to the products objectively and subjectively. This method can also be applied on other devices (e.g., mobile).

Aesthetics. Aesthetics research in Human-Computer Interaction (HCI) increased over the past few years with growing interest in user experience (UX) with interactive technology (Hassenzahl, Diefenbach, & Göritz, 2010; Partala & Saari, 2015; Tuch, Trusell, & Hornbæk, 2013). In this context, Moshagen and Thielsch (2010) define aesthetic perception as 'an immediate pleasurable subjective experience that is directed towards an object' (p. 3), integrating the subjective aspect. Perceived aesthetics is important for website evaluation, especially for first impression (Jennings, 2000; Lindgaard, Dudek, Sen, Sumegi, & Noonan, 2011; Tuch, Presslaber, et al., 2012), but also for overall impression (Thielsch et al., 2013; Van Schaik, Hassenzahl, & Ling, 2012), perceived information quality (Hartmann, Sutcliffe, & Angeli, 2008), and perceived usability (Sonderegger & Sauer, 2010; Tractinsky, Katz, & Ikar, 2000; Tuch, Roth, et al., 2012). Based on website appearance, users draw inferences onto other website characteristics important for judgements of the UX (Van Schaik et al., 2012). In addition, several studies reported a high stability of perceived aesthetics ratings after short exposure time (Lindgaard, Fernandes, Dudek, & Brown, 2006; Thielsch & Hirschfeld, 2010), which prompted the search for design features that increase the website's aesthetics. As a

result, various studies identified critical properties on desktop, such as colors (Reinecke et al., 2013), visual complexity (Michailidou, Harper, & Bechhofer, 2008) and prototypicality (Tuch, Presslaber, et al., 2012). Yet it remains unclear if and how these findings extend to mobile devices.

Affect. Differentiations between affect, emotion and mood are important and oftentimes neglected in HCI research (Boehner, DePaula, Dourish, & Sengers, 2007). Fredrickson (2000) defined affect as consciously accessible feelings, being a component of the subjective experience, and also being present in many affective phenomena, including physical sensations, attitudes, moods, and even traits. Boehner et al. (2007) argued that affect should be considered part of UX, because it occurs in interaction. To produce desired affective states increasing purchase probability, the design of shopping environments was introduced in marketing under the term "atmospherics" (Kotler, 1973). Examples in the physical space are layout, design and employee appearance (Baker, Grewal, and Levy, 1992), or musical and olfactory stimuli (Spangenberg, Grohmann, & Sprott, 2005). Creating store atmospherics online is much more difficult, because e-retailers have to invest more in visual design to compensate the medium's shortcomings, such as the absence of physical space (Levin, Levin, & Heath, 2003; Porat & Tractinsky, 2012). Still, if successfully implemented, relations between the perception of design elements on desktop, affective states, and consumer behavior remain. For instance, Mummalaneni (2005) found that mediated by pleasure, atmospherics had positively influenced the number of items purchased.

The role of time of evaluation. According its definition, usability only starts to play a role when 'a specified user' interacts with technology (ISO, 1998), building a clear contrast to aesthetics, where processing occurs instantly (Leder et al., 2004). Consequently, the point in time when evaluation of a website takes place plays a key role for dependencies between the constructs influencing evaluative ratings. Effects of varying exposure times resulted in plentiful research since the study of Lindgaard et al. (2006). They found correlations of visual appeal ratings up to $r = .96$ after 50 ms website exposure and later ratings without time constraint. The high stability of first impressions was supported by other research (Olivola & Todorov, 2010; Tractinsky, Cokhavi, Kirschenbaum, & Sharfi, 2006). Several explanations

were suggested for these findings. One refers to the halo effect, which occurs if one construct leaves a positive impression (e.g., a beautifully designed website) and influences overall evaluations and other constructs (Lindgaard et al., 2006). This psychological mechanism of deductive reasoning is based on visual cues (Rosenzweig, 2014). Thus perceived aesthetics may affect the attitude of people towards a website and its inherent qualities, such as usability. Another explanation provided was that people tend to stay consistent in their ratings to reduce cognitive load when interacting with a website or a software (Hollender, Hofmann, Deneke, & Schmitz, 2010). This phenomenon is attributed to the confirmation bias, where evidence in support of the initial impression is looked for and conflicting information ignored (Nickerson, 1998). Hence, a positive first impression of a website may lead to downplay of potential issues encountered later: bad usability may be generously overlooked (Campbell & Pisterman, 1996). However, mixed results were reported regarding the stability of aesthetic appraisals after website interaction and usability begins to influence ratings (a list of correlative and experimental studies available in (Tuch, Roth, et al., 2012, p. 4 and 6). The few experimental studies conducted found that high aesthetics positively influenced usability ratings (Ben-Bassat, Meyer, & Tractinsky, 2006; S. Lee & Koubek, 2012; Tractinsky et al., 2000), but also that high usability positively influenced aesthetics ratings (Ben-Bassat et al., 2006; S. Lee & Koubek, 2012). Experimental insights with mobile devices might help to clarify these mixed findings from desktop.

Mobile-specific findings

This study focused on one clearly distinguishable type of content, additionally specified by the device: m-commerce. The interest in m-commerce as online distribution channel has increased continuously over the last decade and gains relevance with further development and wider circulation of mobile devices and online businesses (Gross, 2015). Research conducted in m-commerce so far was more focused on single constructs, instead of overall evaluations and subsequent purchase behavior. Hartmann et al. (2008) focused on mobile in one study and found that customized content was preferred over non-customized content and subjectively perceived as better. For usability, the PACMAD (People At the Centre of Mobile Application

Development) model has been developed based on reviewing other mobile usability models, bringing together significant attributes in order to form a more integrative model (Harrison, Flood, & Duce, 2013). Miniukovich and De Angeli (2014) found that first impressions of perceived aesthetics persist on mobile interfaces, although it was suggested that exposure time may be longer. In aesthetics scores, 36% of variation were explained using six metrics: color depth, dominant colors, visual clutter, symmetry, figure-ground contrast and edge congestion. The relationship between design elements, affect and trust has been investigated (Li & Yeh, 2010), but the role of affect in mobile webshop overall evaluations is undetermined.

The evaluation model for desktop websites

Content, aesthetics and usability's influences on website evaluation have been examined by Thielsch et al. (2013) in a series of three studies on desktop. In the first study, 330 web users were asked about constructs they subjectively perceive as most relevant to evaluate a website's quality. Content was stated to be first, followed by usability and aesthetics. The second and third study (together N = 812) tested 46 websites with nine different content domains, including e-commerce. Variables measured after short and after interactive exposure were content, aesthetics, usability, as well as first and overall impression. The intention to revisit and recommend the website was measured once, after interactive exposure without time constraint. As a result, a correlational path model showed that all three constructs correlated significantly with first and overall impression, aesthetics having the highest correlation. Moreover, content correlated highest with the intention to revisit and recommend a website. Figure 1 depicts the correlational path model from the third study.

While this model improves the understanding of correlations between the key constructs and overall evaluations, several things remain ambiguous. First, it is unclear how long the initial exposure time was, making it difficult to estimate the deliberate processing time which the participant had to form an impression. Second, the intention to revisit and recommend the website was only measured after interactive exposure, not allowing the comparison of high-level ratings' stability over time. Third, the correlational path model implies causal directionality, but the paths are based on theoretical reasoning and not on experimental

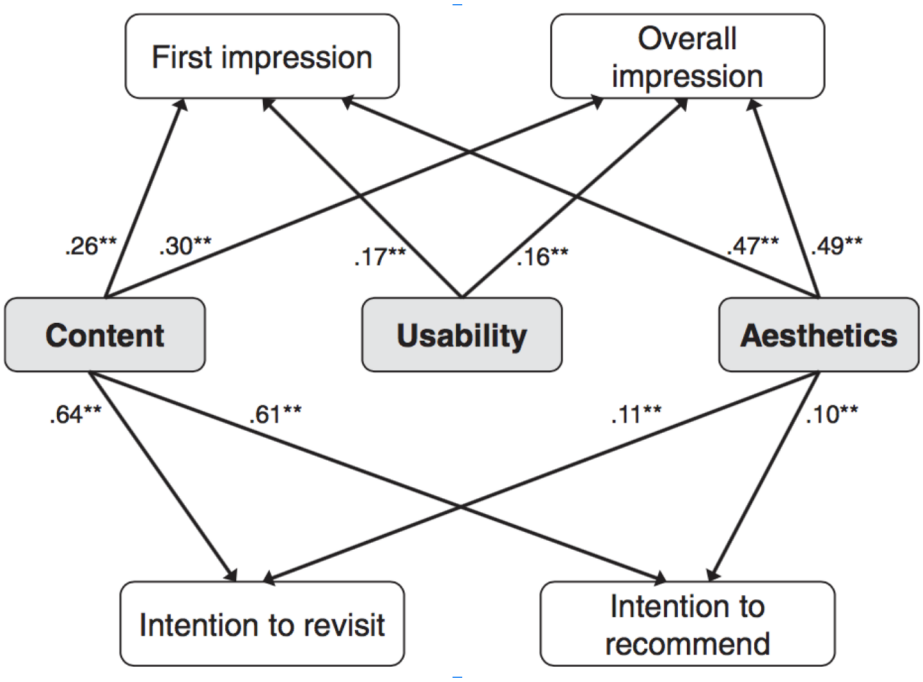

Figure 1. Path model of the third study presented by Thielsch et al. (2013) (p. 95).

manipulations. The emerging influential pattern thus remains unclear since causation cannot be determined in correlative studies (Aldrich, 1995). Fourth, within the model in Figure 1, some correlations are missing. Striking is the missing path between usability and intentions to revisit and recommend, as well as correlations between first and overall impression. Although some correlations within the overall evaluative constructs were reported to be significant in a separate table (Thielsch et al., 2013, p. 95), they were not further discussed by the authors, leaving a gap in the model's explanatory power. Fifth, other constructs, such as affect (Porat & Tractinsky, 2012), have shown to influence the evaluation of websites on desktop, but were not measured. These shortcomings were addressed in the present study, as detailed in the following section.

The present study

This study focuses on construct-specific and overall evaluations of mobile webshops. It builds on the model of Thielsch et al. (2013) and aims to extend it by several means. To begin with, the study is conducted in the field of m-commerce. Even though desktop devices dominated for nearly three decades, mobile devices are increasing in usage (Kemp, 24.01.2017). In order to gain insights on the applicability of research conducted on desktop so

far, it is important to learn about global factors across devices and factors being device-specific. In addition, an experimental setup is used, manipulating aesthetics and usability, to investigate their influence and detect causal effects on the evaluation pattern, while content is held constant. This method aims to answer the first research question: How do aesthetics and usability influence construct-specific and overall evaluations of mobile webshops after short exposure time and after interactive exposure with no time constraint?

Furthermore, exposure time for first impression is standardized and set to 500 ms. Based on this and Thielsch's model the second research question is stated as follows: How stable is the evaluation of a mobile webshop over time in regards to perceived content, aesthetics and usability?

Finally, adjusted measures also capture affect and the overall evaluation by measuring all dependent variables after first (T1) and overall impression (T2) to test the judgment stability. To the author's knowledge no study has measured overall evaluations and hevaioral intentions on mobile so far. (Zhang & Li, 2004) stated that, on desktop, these ratings should be as stable as the ratings of aesthetic appeal. This results in the third and last research question: Do overall evaluative and behavioral intention ratings regarding a mobile webshop change over time?

In this study, a three-factorial mixed design was chosen with high/low usability and high/low aesthetics as between-subject variables and time of measurement as within-subject variable to answer the research questions. A mobile webshop selling books was developed and used with four conditions: high aesthetics and high usability (HAHU), high aesthetics and low usability (HALU), low aesthetics and high usability (LAHU) and low aesthetics and low usability (LALU).

Pre-study

Four mobile book webshops with manipulated aesthetics and usability were created for the experimental conditions. The "Saleor" webshop developed by Mirumee[1] served as template while the content was acquired from different online bookshops including cover,

[1] for more information on the webshop template visit: http://getsaleor.com

content description and a short biography of the author. Effects of identity and image were excluded by creating a new webshop instead of using an existing one that might elicit unwanted associations (Tractinsky & Lowengart, 2007). A pre-study and a Latent Semantic Analysis were conducted to ensure successful manipulation of the four webshops with the following hypotheses postulated:

1. The information scent for high usability conditions is significantly higher than for low usability conditions.

2. High aesthetics conditions have significantly higher mean ratings of aesthetic appraisals than low aesthetic conditions.

Usability manipulation

Manipulating usability and aesthetics systematically and independently is difficult because confounding effects arise easily (Tuch, Roth, et al., 2012). To leave the aesthetic appraisal uncompromised, usability was manipulated via the IA, changing navigation menu labels and assignment of items to categories, as previously done in a study conducted by Tuch, Roth, et al. (2012). To estimate how easily content can be found with the good and bad IA, the information scent for task keywords was evaluated using the Latent Semantic Analysis (LSA) procedure[2]. LSA extracts contextual-usage meaning of words via statistical computations that were trained on large corpuses of text. Thereby, all word contexts in which a word is used or not, provide a set of mutual constraints determining the similarity and meaning of words. For each word pair that is compared, the cosine value is calculated and values close to 1 indicate a high semantic relation (Landauer, Foltz, & Laham, 1998). Because the LSA is not available in German (native language of sample) it was used with English translations. Cosine values of labels of each navigation level with the task keyword were averaged and compared for the good and the bad IA to indicate the quality of the information scent across all three navigation levels. Figure 2 shows the LSA values for the task solving path through the different navigation levels and the average across all labels for the bad and the good IA.

[2]For more information and the analysis tool please visit: http://lsa.colorado.edu

Target Item	Good IA				Bad IA			
	1st level	2nd level	3rd level	averaged path scent	1st level	2nd level	3rd level	averaged path scent
thriller book	book	fiction	thriller		favourites	newest	fiction	
	1,00	0,50	1,00	**0,83**	0	0,14	0,5	**0,21**
love story eBook	eBook	fiction	romance		newest	favourites	eBook	
	1,00	0,64	0,46	**0,70**	0,12	0,11	1,00	**0,41**
"garden desire" book	book	past time	garden		bestseller	monthly favourite	past time	
	1,00	0,31	1,00	**0,77**	0,07	0,07	0,31	**0,15**

Figure 2. LSA of the information scent for the bad and the good IA in the mobile webshops.

Aesthetics manipulation

To manipulate aesthetics, the focus was set on design properties creating an appearance, meaning background color, color composition and logo (Bloch, Brunel, & Arnold, 2003; Lowry, Wilson, & Haig, 2014). Factors such as proportion, shape and size were held constant since they could interfere with usability by increasing visual complexity, which has shown to decrease user satisfaction in mobile webshops (Sohn et al., 2017). For the high aesthetics condition, the following three colors previously identified as preferred were chosen: white, light grey and dark grey (Bonnardel, Piolat, & Le Bigot, 2011; Cyr, Head, & Larios, 2010; Schloss & Palmer, 2011; Seckler, Opwis, & Tuch, 2015). The logo was designed to confirm a good and clean impression increasing credibility (Lowry et al., 2014) with a modern and readable font in a dark grey that contrasts well with the background. In the low aesthetic condition, intense colors that were previously identified as less preferred (Bonnardel et al., 2011; Seckler et al., 2015) were chosen: green, turquoise and black. The logo was designed to contrast unpleasantly with the green and turquoise having a purple color and an old fashioned font providing a negative and ugly impression. To ensure good readability, a contrast check between font and background was conducted. High aesthetics conditions had a contrast of 10.6:1, while low conditions had one of 11.6:1, both clearly exceeding the threshold of 4.5:1 set by the web content accessibility guidelines to pass as readable (Consortium et al., 2008). Screenshots of the four webshops are displayed in Figure 3 (for the webshop links refer to Appendix A).

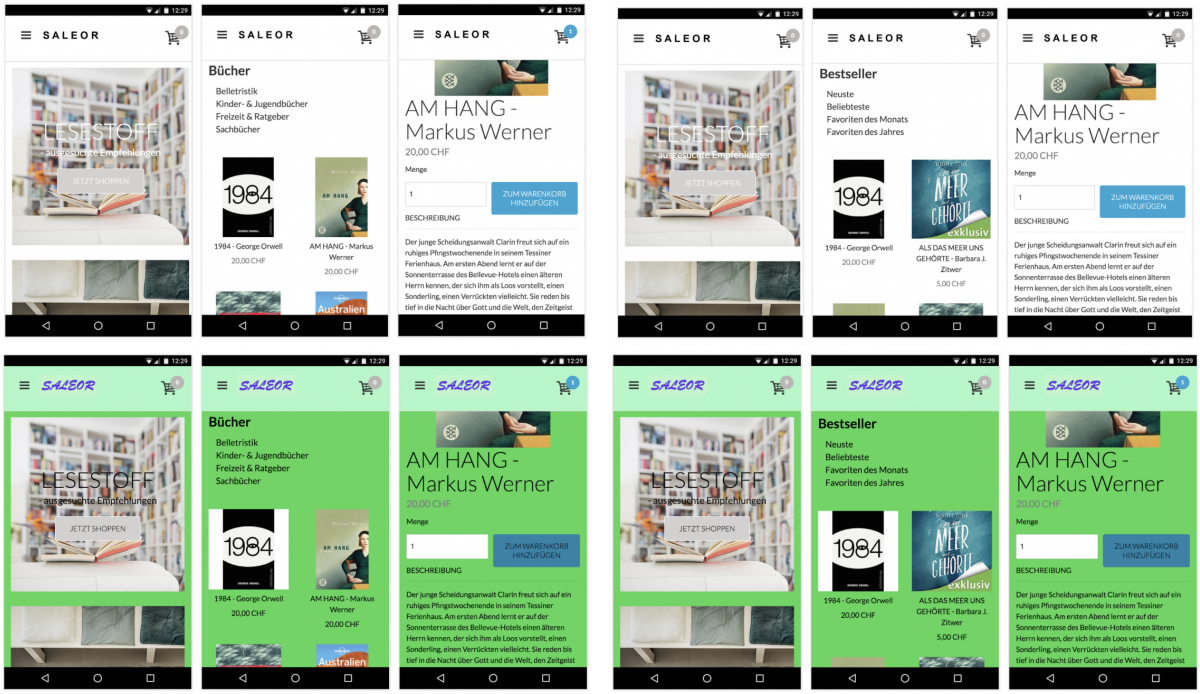

Figure 3. Screenshots of the four mobile webshop conditions (upper left: HAHU, upper right: HALU, lower left: LAHU, lower right: LALU).

Procedure and sample

The pre-study took about five minutes to complete and was conducted on EFS Survey. Subjects consented to study participation and data evaluation before being randomly assigned to one of the four conditions and receiving introductory information. Subsequently, three mobile webshop screenshots aligned in a row (as in Figure 3) were displayed without time constraint and participants rated the webshop regarding aesthetics, usability, overall impression and intention to revisit and recommend the mobile webshop (scales are described in the method section of the main study). Finally, demographic information was collected before the study compensation code worth 70 cents (US Dollars) was handed out. The results are based on a dataset collected from German speaking Crowdflower[3] users in May 2017. All participants had to be 18 years or older and pass an attention check, which otherwise led to direct exclusion from the survey. A total of 134 observations were made, but to attain high data quality the following exclusion criteria were applied: incomplete responses (N = 10), less than 150 seconds to complete the study (N = 6) and more than 23 of 25 equal responses (N = 6),

[3]Information on the crowdsourcing platform are found here: https://www.crowdflower.com

leaving a final sample of 112 participants. Divided by condition, this results in 28 participants for HAHU, 30 for HALU, 26 for LAHU and 28 for LALU. The age range of the participants reached from 18 to 82 (M = 39.9, SD = 14) with N = 32 female (36%) and N = 80 male (64%).

Results

Table 1 depicts mean ratings for perceived aesthetics and usability per condition are given. A Welch two sample t-test was applied to compare the mean ratings between high and low conditions and to measure their mean effect size (refer to Appendix B for links to the dataset and R script). Results of perceived aesthetics showed a significant difference between the high and low aesthetics condition mean ratings $t(109) = 2.63$, $p = 0.005$, with high aesthetics having higher ratings than low aesthetics and Cohen's $d = 0.50$. For perceived usability, no significant difference between high and low usability condition mean ratings were found $t(110) = 0.07$, $p = .474$, and Cohen's $d = 0.01$.

Table 1

Descriptive statistics of aesthetics and usability ratings.

Perceived aesthetics	High Aesthetics *M* (*SD*)	Low Aesthetics *M* (*SD*)
High Usability	4.88 (1.21)	4.16 (1.4)
Low Usability	4.98 (1.21)	4.49 (1)
Both groups:	4.93 (1.21)	4.33 (1.2)
Perceived usability	**High Usability** ***M* (*SD*)**	**Low Usability** ***M* (*SD*)**
High Aesthetics	5.14 (1)	5.09 (1.29)
Low Aesthetics	4.74 (1.48)	4.76 (1.29)
Both groups:	4.94 (1.24)	4.93 (1.29)

Implications for the main study

Results of the LSA and the pre-study show that the manipulations of aesthetics and usability were successful. The usability ratings did not show a significant difference in the

pre-study ratings, but a halo effect (Rosenzweig, 2014) of the aesthetics manipulation resulted in slightly higher mean ratings for the two high aesthetics conditions. However, this does not imply that the manipulation was not successful since an interaction with the webshop is necessary to perceive and evaluate its usability properly (ISO, 1998). In conclusion, the first and second hypotheses are supported. In the following section, information on the main study measures, procedure and participants sample are given.

Method

Measures

The dependent variables used were chosen because they showed good validations and for comparability as they were previously used by Thielsch et al. (2013). Answers were given on an agreement likert scale from 1 (*do not agree at all*) to 7 (*completely agree*) if not stated otherwise.

Perceived usability. The scale developed by Flavián, Guinalíu, and Gurrea (2006) and translated to German by Thielsch (2008) was used to measure perceived usability. The scale shows high internal consistency with $\alpha = .95/.96/.96$ [4] and consists of seven items.

Objective usability. Objective usability was measured controlling the average clicks and time needed to complete the three tasks via the website analytics tool Piwik[5]. Unfortunately, the website had problems tracking the sessions separately when using the same smartphone and clustered several participants into one tracking session, which was not dividable. Consequently, the output only contained 70 rows of data instead of 124 and they all varied greatly in duration and clicks. Thus, the data could not be further taken into account. However, task scores could be used as measure of performance. By solving three tasks during interaction with the webshop without time constraint, participants could achieve up to three points.

Perceived aesthetics. To measure perceived aesthetics the Visual Aesthetics of Websites Inventory developed by Moshagen and Thielsch (2010) was used, consisting of 18

[4] Internal consistency values are presented as: α = original validation / Thielsch et al. (2013) / this thesis.

[5] For information on the website analytics tool visit: https://piwik.org/

items and four facets (simplicity, diversity, colorfulness and craftsmanship). The values of internal consistency were again high with $\alpha = .94/.94/.91$.

Content. Content was measured with a scale developed by (Thielsch & Hirschfeld, 2016, under review), containing 12 items in four facets (comprehensibility, appeal, informational content and credibility). Internal consistency was high in this study with $\alpha = .94$. The scale is a revised version from a scale developed by Thielsch (2008) with nine items, where internal consistencies were also high with $\alpha = .85$ and $\alpha = .88$ from Thielsch et al. (2013).

Affect. The self-assessment manikin developed by Bradley and Lang (1994) measured affect. The instruction "Please rate how you feel at the moment." was freely translated to German. The answers were given on a nine-point graphical scale that consists of one valence (positive - negative), one arousal (high - low) and one dominance (big - small) rating.

First and overall impression. First and overall impression were measured by giving a single grade from the Swiss school grades system, ranging from 1 (worst) to 6 (best). Thielsch et al. (2013) used the same scale but in reverse (1 being the best and 6 the worst), according to the German school grades system.

Intention to revisit. Thielsch et al. (2013) used a scale that consists of three items to measure the intention to revisit. Internal consistency can only be given for the study described in this thesis with $\alpha = .88$.

Intention to recommend. A frequently used measure in the commercial sector to predict business growth is the intention to recommend, which is assessed with the Net Promoter Score (NPS) (Reichheld & Markey, 2011). It is based on a single item asking 'How likely is it that you would recommend this website to a friend or colleague?' and is answered on scale ranging from 0 to 10.

Demographics. Demographics which were asked from participants concerned their gender, age, job situation, webdesign knowledge, phone brand as well as online purchase quantity and frequency.

Procedure

The study was carried out in the lab of the psychological department of the University of Basel on a Windows PC. The study ran on desktop in fullscreen mode via EFS Survey and a Samsung Galaxy S5 mini smartphone was used for the mobile webshop interaction. Participants were instructed to read and sign a consent form before starting with the study. First, a welcome text with some general information was shown before participants had to rate their affect as a baseline. After random assignment to one of the four webshop conditions, a short video with three screenshots of the mobile webshop aligned in a row (as in Figure 3) was shown for one second. Then, participants indicated their affective state again and answered the first battery of questionnaires to measure first impressions. Thereafter, three tasks had to be completed in the mobile webshop on the Galaxy S5 smartphone to ensure everyone had the same device and resolution. Tasks were chosen with care to provide context (Van Schaik & Ling, 2009). However, no performance-oriented pressure was placed on participants as this would possibly reduce the perceived value of beauty, artificially decreasing perceived aesthetics ratings Ben-Bassat et al. (2006). The tasks were the following: 'You would like to gift a friend with a thriller book. Please place a thriller book of your choice in the shopping basket.', 'Please place an eBook about a love story of your choice in the shopping basket.' and 'Find the following hobby guidebook that was cured favourite of the month and place it in the shopping basket: "Garden Desire" by Helena Attlee'. After task completion, affect and the battery of questionnaires were rated again in reference to the overall impression collected from the interaction time with the mobile shop. Finally, demographic information was collected before a thank you gift of five Swiss francs or one signature for psychology students was handed out and the participant was dismissed. Figure 4 visually depicts the procedure.

Sample

In total, 124 participants were recruited in June and July 2017 via social media, the author's private environment and university-internal recruitment platforms. Of the total sample, two participants were excluded due to technical difficulties. Another 29 were excluded due to working or educational experience in webdesign (N = 11) or related domains

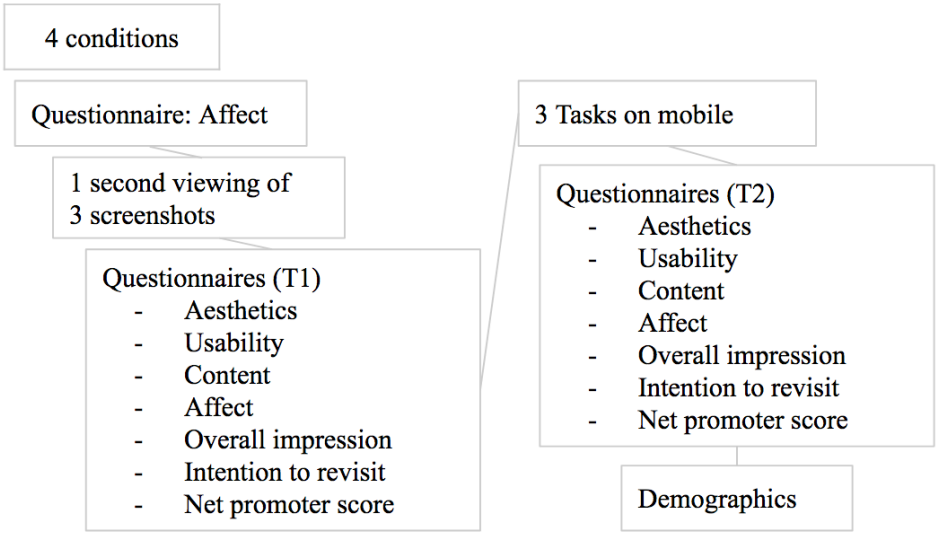

Figure 4. The procedure of the main study.

(N = 18), after calculating the main analysis twice, once with and once without experts. Results of the analysis differed significantly when experienced participants were included with additional main effect and interactions becoming significant. In addition, they showed a significant counterbalance in the distribution across the different conditions, confirming the necessity of exclusion. Consequently, the final sample consisted of 93 observations (HAHU n = 29, HALU n = 21, LAHU n = 19, LALU n = 24). Participants all consented to study participation and data usage. Their ages ranged from 18 to 54 years (M = 25.89, SD = 5.92), with N = 51 being male (55%) and N = 42 female (45%). Most participants were students (N = 63) and/or employed (N = 36). Task scores, phone brands, online shopping frequency and quantity distributed evenly across all four conditions according to the Chi-square tests (refer to Appendix B for links to the dataset and R script of the main study).

Results

The analysis was conducted with the software R Studio. For all statistical tests an alpha level of .05 was used. To analyse the dependent variables a mixed analysis of variance type III (ANOVA) was calculated with a $2 \times 2 \times 2$ (high/low aesthetics - high/low usability - points in time T1/T2) repeated measures design. Preliminary tests of normal distribution (Shapiro-Wilk test) and homogeneity of variance (Levene test) were conducted. While Levene tests showed

normal homoscedasticity for all dependent variables at both points in time, the Shapiro-Wilk test revealed some significant deviations from normal distribution. ANOVAs were found to be quite robust against violations of normal distribution assumptions as long as homogeneity of variance was given (Schminder, Ziegler, Danay, Beyer, & Bühner, 2010), so the author decided to proceed with the analysis. Mean ratings are depicted in Table 2 and results are presented separately for each dependent variable.

Table 2

Descriptive statistics of the dependent variables divided by time and condition.

Dependent variable	HAHU	HALU	LAHU	LALU
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Aesthetics T1	4.80 (0.98)	4.72 (0.94)	3.37 (1.09)	3.40 (1)
Aesthetics T2	5.37 (0.86)	4.79 (0.88)	3.72 (1.3)	3.70 (1.09)
Usability T1	4.38 (1.1)	4.37 (0.78)	3.49 (1.4)	3.53 (1.33)
Usability T2	4.91 (1.31)	3.20 (1.41)	4.59 (1.34)	2.79 (1.43)
Task score (T2)	2.48 (0.63)	2.10 (1.00)	2.58 (0.69)	2.00 (1.02)
Intention to revisit T1	4.18 (1.12)	4.51 (1.27)	2.67 (1)	2.82 (1.29)
Intention to revisit T2	4.42 (1.57)	2.91 (1.66)	3.13 (1.68)	2.59 (1.72)
Intention to recommend T1	6.07 (2.06)	6.23 (2.7)	3.65 (1.74)	3.16 (2.05)
Intention to recommend T2	6.53 (2.32)	4.49 (2.84)	4.36 (2.81)	3.39 (2.54)
First impression (T1)	3.90 (1.47)	3.81 (1.6)	2.55 (1.3)	2.84 (1.23)
Overall impression (T2)	3.43 (1.84)	2.61 (1.44)	3.03 (1.53)	2.45 (1.56)

Perceived aesthetics. For perceived aesthetics, a significant main effect for the aesthetics factor was found, $F(1, 89) = 58.84, p < .001, \eta^2 = .40$, indicating that the manipulation actually influenced the ratings, with high aesthetics conditions having higher ratings than low aesthetics conditions. Also, results revealed a significant main effect for time, $F(1, 89) = 7.82, p = .006, \eta^2 = .09$, with ratings being higher at T2 for all conditions. No significant main effect for usability arose, $F(1, 89) = 0.86, p = .369, \eta^2 = .01$, and no

significant interactions[6]. The line diagram in Figure 5 shows the rise of aesthetics ratings for all conditions after interaction.

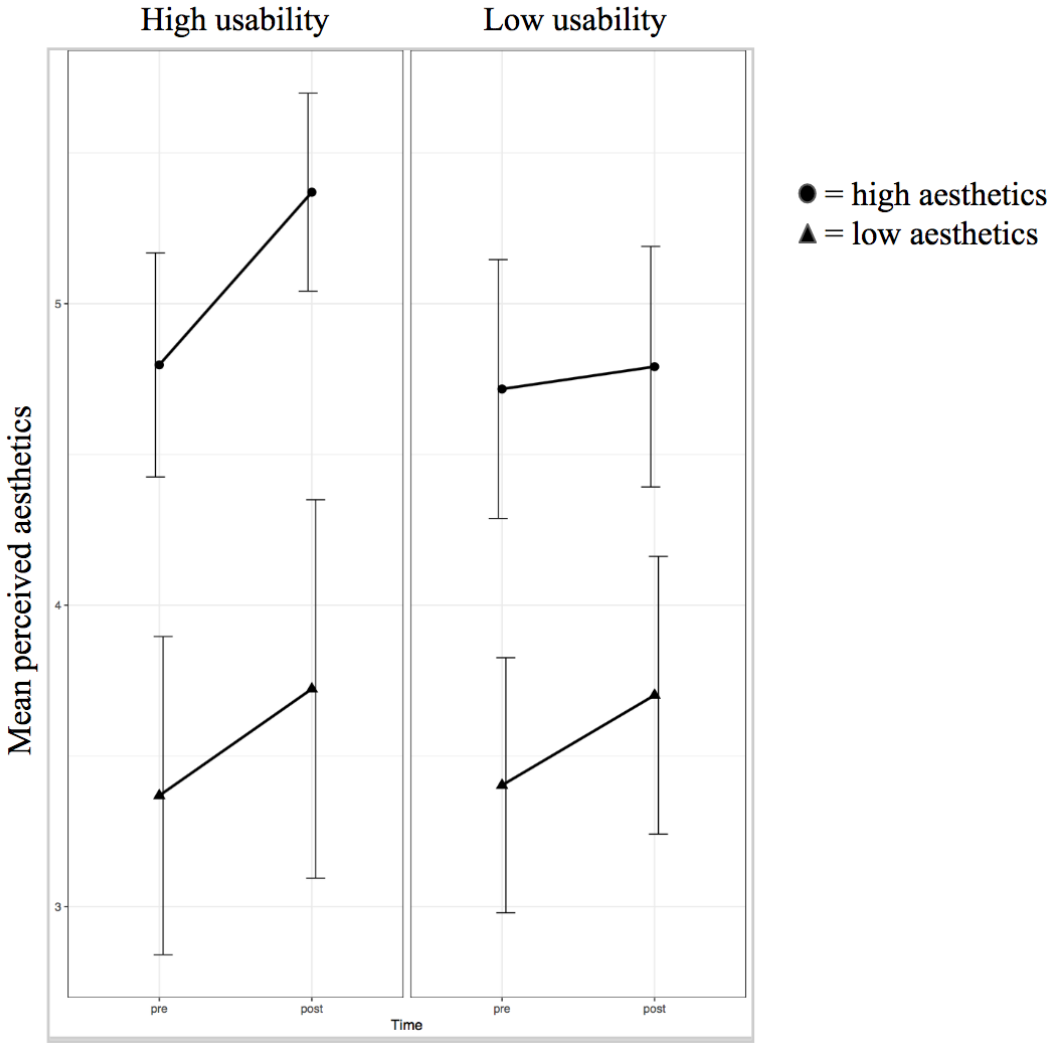

Figure 5. Perceived aesthetics ratings for each level of aesthetics, usability and time. Error bars of the 95% confidence interval are visible.

Perceived usability. A significant main effect for usability was found in the perceived usability ratings, $F(1, 89) = 17.86, p < .001, \eta^2 = .17$, with ratings being higher for high usability than for low usability conditions at T2. The significant main effect for aesthetics, $F(1, 89) = 8.93, p = .004, \eta^2 = .09$, showed high aesthetics conditions had significantly higher ratings in perceived usability than low aesthetics conditions. The significant interaction

[6]Non-significant results from interactions are not included in the main text when no significance was expected. They can be extracted from the dataset and R script provided in Appendix B.

between usability and time, $F(1, 89) = 26.89, p < .001, \eta^2 = .23$, confirmed the successful manipulation of usability by ratings increasing for high usability conditions and decreasing for low usability conditions over time (see Figure 6). The aesthetics and time interaction as well as the three-way interaction were not significant. A Welch two sample *t*-test showed that the difference between high and low usability conditions in task scores was significant with $t(75) = 2.70$, $p = .004$, confirming that the usability manipulation also influenced objective usability by lowering the performance of the participants in the low usability conditions.

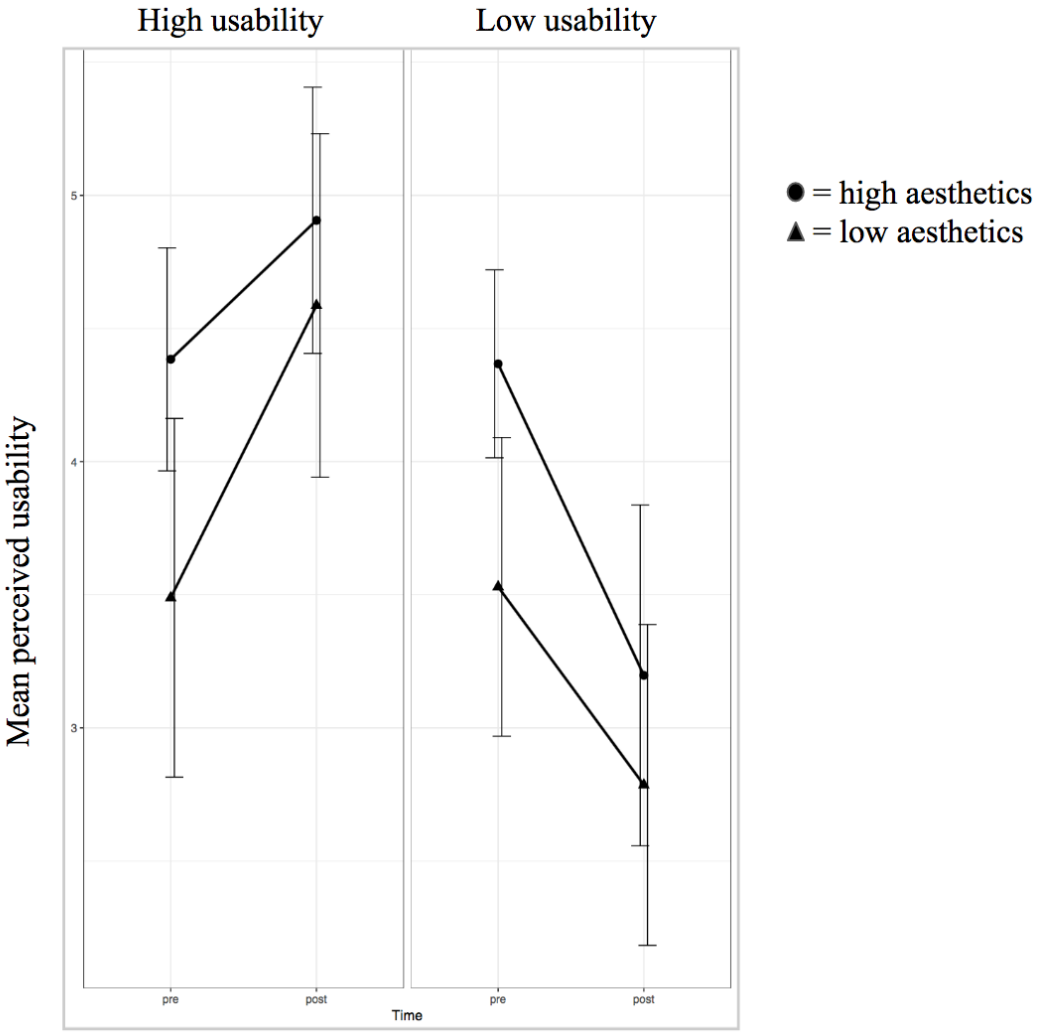

Figure 6. Perceived usability ratings for each level of aesthetics, usability and time. Error bars of the 95% confidence interval are visible.

First and overall impression. Once more, aesthetics had a significant main effect on impression ratings, $F(1, 89) = 4.32, p = .04, \eta^2 = .05$, because these were higher for high aesthetics conditions, except for HALU at T2. Usability showed no significant main effect,

$F(1, 89) = 1.21, p = .274, \eta^2 = .01$, but results showed significant interactions for aesthetics and time, $F(1, 89) = 6.63, p = .01, \eta^2 = .07$, and usability and time, $F(1, 89) = 4.54, p = .04, \eta^2 = .05$ due to the drops in ratings, except for LAHU displaying a positive influence of high usability at T2 (see Figure 7). The three-way interaction of aesthetics, usability and time was not significant.

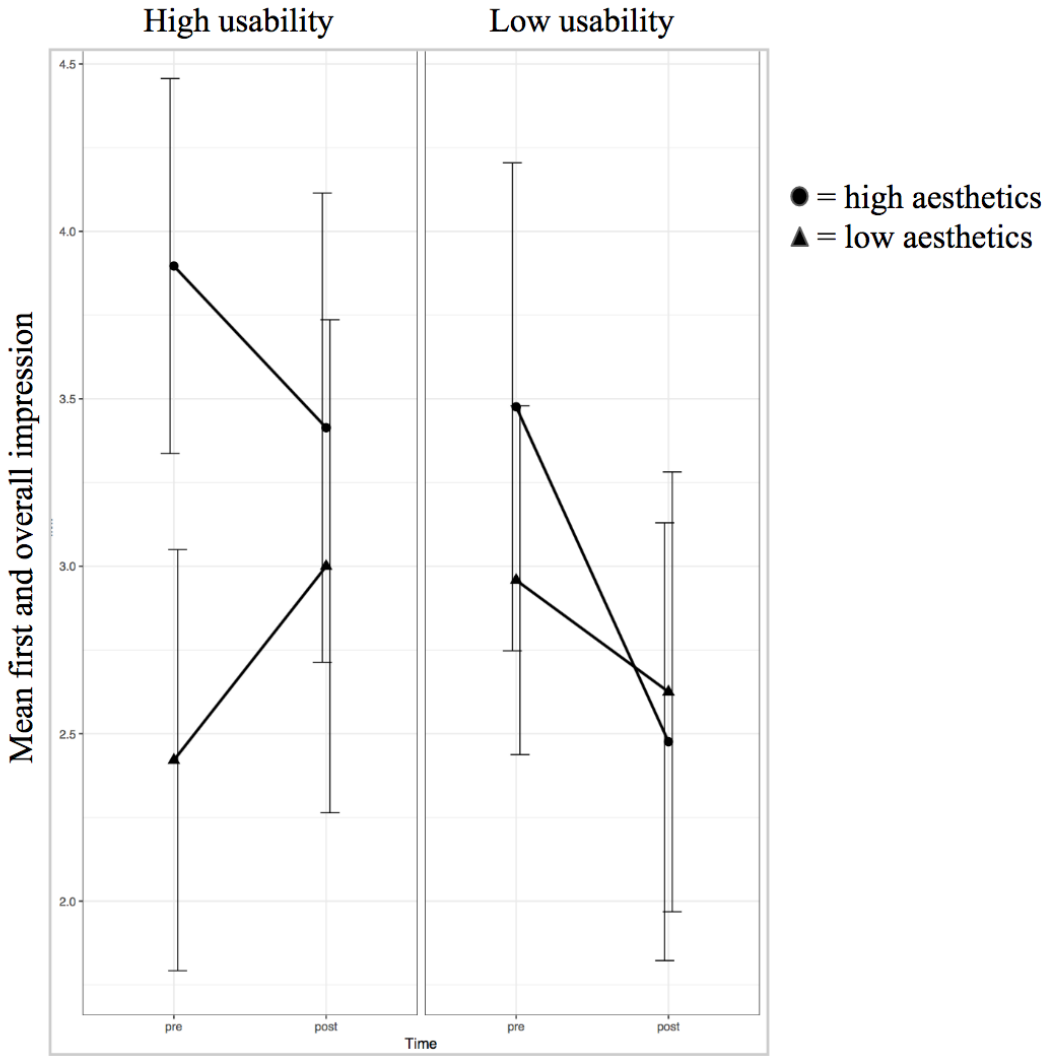

Figure 7. First and overall impression ratings for each level of aesthetics, usability and time. Error bars of the 95% confidence interval are visible.

Intention to revisit. Aesthetics showed a significant main effect, $F(1, 89) = 18.02, p < .001, \eta^2 = .17$, due to high aesthetics conditions having higher ratings, except HALU at T2, while usability showed no significant main effect, $F(1, 89) = 1.99, p = .167, \eta^2 = .02$. The rise in ratings for high usability conditions and drop in ratings

for HALU to revisit the shop (see Figure 8) led to significant interactions for aesthetics and time, $F(1, 89) = 9.96, p = .002, \eta^2 = .10$, and usability and time, $F(1, 89) = 15.78, p < .001, \eta^2 = .15$. The three-way interaction of aesthetics, usability and time was again not significant.

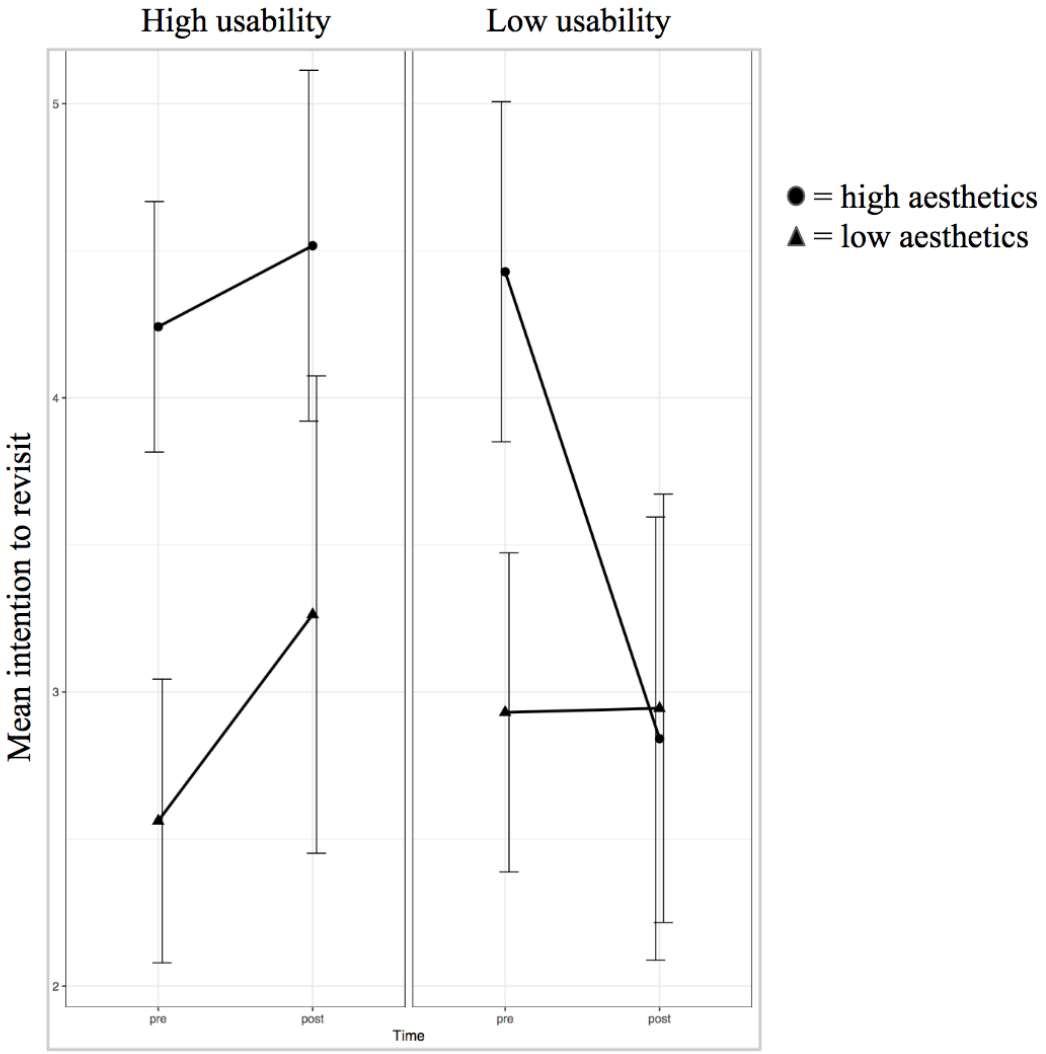

Figure 8. Intention to revisit ratings for each level of aesthetics, usability and time. Error bars of the 90% confidence interval are visible.

Intention to recommend. Aesthetics showed a significant main effect, $F(1, 89) = 21.68, p < .001, \eta^2 = .20$ with high aesthetics conditions again displaying higher NPS ratings, except HALU at T2, while usability marginally did not, $F(1, 89) = 3.75, p = .052, \eta^2 = .04$. Two significant interactions were found for aesthetics and time, $F(1, 89) = 9.42, p = .003, \eta^2 = .10$, and for usability and time, $F(1, 89)$

$= 8.21, p = .005, \eta^2 = .08$, due to HALU's drop contrasting with the other conditions' rise in ratings (see Figure 9). Once more, three-way interaction of aesthetics, usability and time was not significant.

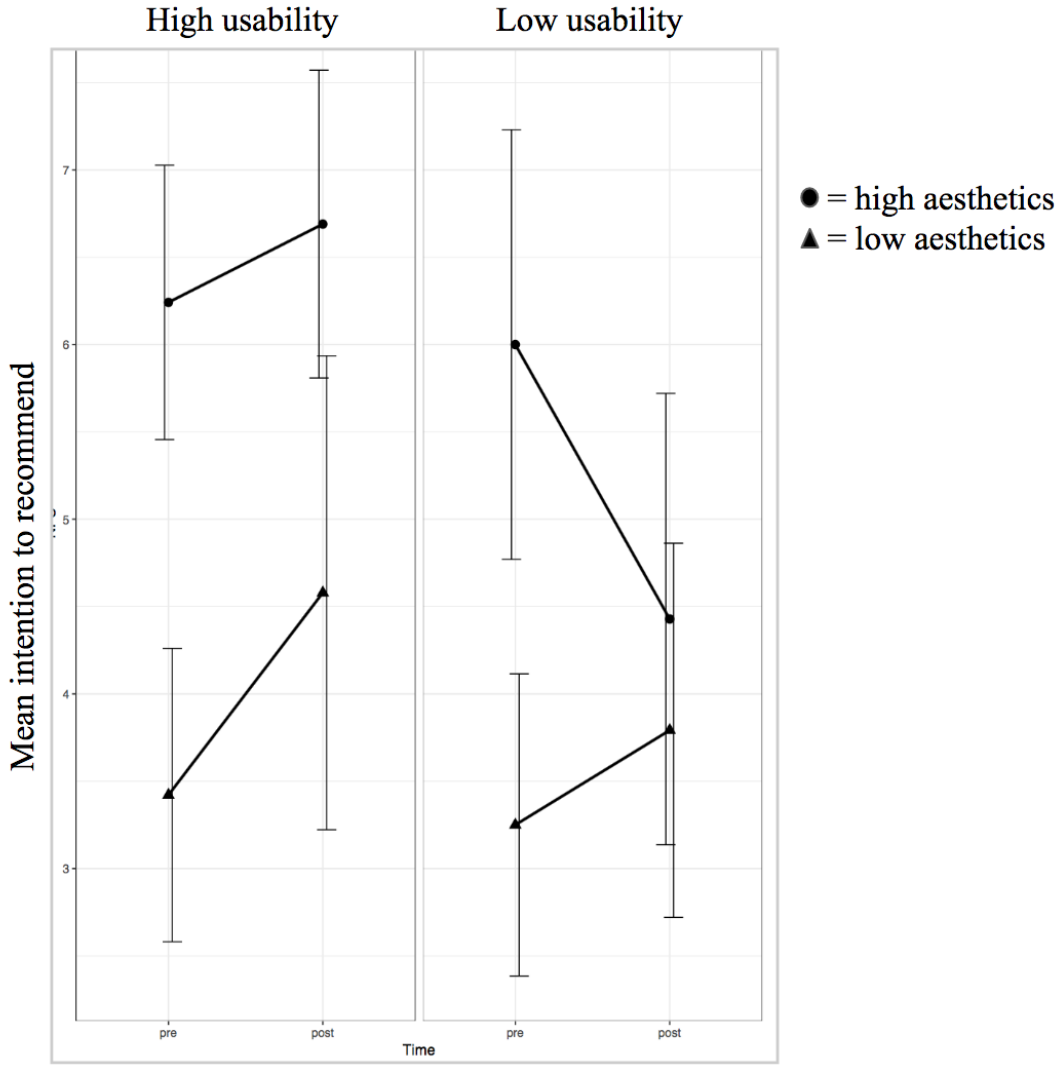

Figure 9. Intention to recommend ratings for each level of aesthetics, usability and time. Error bars of the 95% confidence interval are visible.

Correlations

Pearson's product-moment correlations within constructs between the two points in time were calculated to estimate the stability of ratings. Values of *r* are displayed in Table 3 or else stated in the main text. Perceived aesthetics ($p < .001$), perceived usability ($p = .04$), perceived content ($p < .001$) and first and overall impression ratings ($p < .001$) correlated significantly.

Also between T1 and T2, ratings of intention to revisit were significantly correlated with $r(91) = .44, p < .001$, and ratings of recommendation intentions with $r(91) = .62, p < .001$. The following correlations between constructs were significant at $p < .001$. Perceived aesthetics at T1 correlated with perceived usability, $r(91) = .53$ and content, $r(91) = .76$, as well as first impression, $r(91) = .60$. At T2, perceived aesthetics correlated with perceived usability, $r(91) = .55$, content, $r(91) = .73$, and overall impression, $r(91) = .35$. Perceived usability at first measure also correlated with content, $r(91) = .70$, and first impression, $r(91) = .36$. And at T2, perceived usability correlated with content, $r(91) = .59$, and overall impression, $r(91) = .37$. With $r(91) = .43$, content at T1 and first impression also correlated as well as content at T2 and overall impression with $r(91) = .41$.

Table 3

Correlations between aesthetics, usability and content after short exposure (T1) and interactive exposure (T2), as well as first and overall impression.

Measure	1	2	3	4	5	6	7	8
1. Aesthetics T1	–							
2. Aesthetics T2	0.6	–						
3. Usability T1	0.53	0.4	–					
4. Usability T2	0.21	0.55	0.21	–				
5. Content T1	0.76	0.55	0.7	0.27	–			
6. Content T2	0.35	0.73	0.31	0.59	0.43	–		
7. First Impression	0.6	0.39	0.36	0.16	0.43	0.23	–	
8. Overall Impression	0.09	0.35	0.02	0.37	0.11	0.41	0.44	–

Explorative analysis

The mean ratings for content and affect dimensions are displayed in Table 4.

Content. The main effect for aesthetics was significant, $F(1, 89) = 27.57, p < .001, \eta^2 = .24$, showing that perceived content is also influenced with high aesthetics conditions being rated higher at T1 and T2. Again, usability showed no significant

Table 4

Descriptive statistics of the dependent variables divided by time and condition.

Dependent variable	HAHU *M* (*SD*)	HALU *M* (*SD*)	LAHU *M* (*SD*)	LALU *M* (*SD*)
Content T1	4.28 (0.88)	4.29 (0.83)	3.20 (1.25)	3.17 (0.82)
Content T2	5.13 (0.88)	4.74 (1.15)	4.19 (1.23)	4.20 (1.2)
Valence T0 (baseline)	6.59 (1.62)	7.48 (1.36)	6.42 (1.61)	6.88 (1.26)
Valence T1	6.55 (1.5)	7.05 (1.2)	6.16 (1.57)	5.88 (1.65)
Valence T2	6.79 (1.61)	6.19 (1.86)	6.16 (1.26)	5.58 (2)
Arousal T0 (baseline)	6.59 (1.61)	7.48 (1.61)	6.42 (2)	6.88 (1.54)
Arousal T1	6.55 (1.48)	7.05 (1.31)	6.16 (1.81)	5.88 (1.98)
Arousal T2	6.79 (1.68)	6.19 (1.66)	6.16 (1.46)	5.58 (2.16)
Dominance T0 (baseline)	6.10 (1.84)	5.62 (1.66)	5.16 (1.98)	5.75 (1.7)
Dominance T1	5.79 (1.97)	5.38 (1.36)	4.58 (1.77)	4.54 (1.61)
Dominance T2	5.66 (1.84)	5.05 (1.69)	4.90 (1.94)	4.58 (2.04)

main effect, $F(1, 89) = 0.32, p = .572, \eta^2 = .00$. Instead, results revealed a significant main effect for time, $F(1, 89) = 43.97, p < .001, \eta^2 = .33$. This effect is visible in the rise of content ratings for all conditions (see Figure 10). No interaction was significant, since ratings developed in the same direction for all conditions.

Affect - valence, arousal and dominance. For the dimensions of affect a sphericity test was conducted to ensure variances of differences between all pairs in the within-subject conditions were equal. Where sphericity was violated, a Greenhouse-Geisser correction was applied. A significant main effect for the aesthetics factor was found for valence, $F(1, 89) = 4.62, p = .03, \eta^2 = .05$, and for dominance, $F(1, 89) = 4.46, p = .04, \eta^2 = .05$, being higher in high aesthetics conditions. Also, results revealed significant main effects for time in valence, $F(1, 89) = 9.53, p < .001, \eta^2 = .10$, arousal, $F(1, 89) = 4.07, p = .02, \eta^2 = .04$, and dominance, $F(1, 89) = 8.28, p < .001, \eta^2 = .09$. Valence and dominance dropped and arousal increased from baseline (T0) to T1 and ratings dropped again at T2, except for low aesthetics

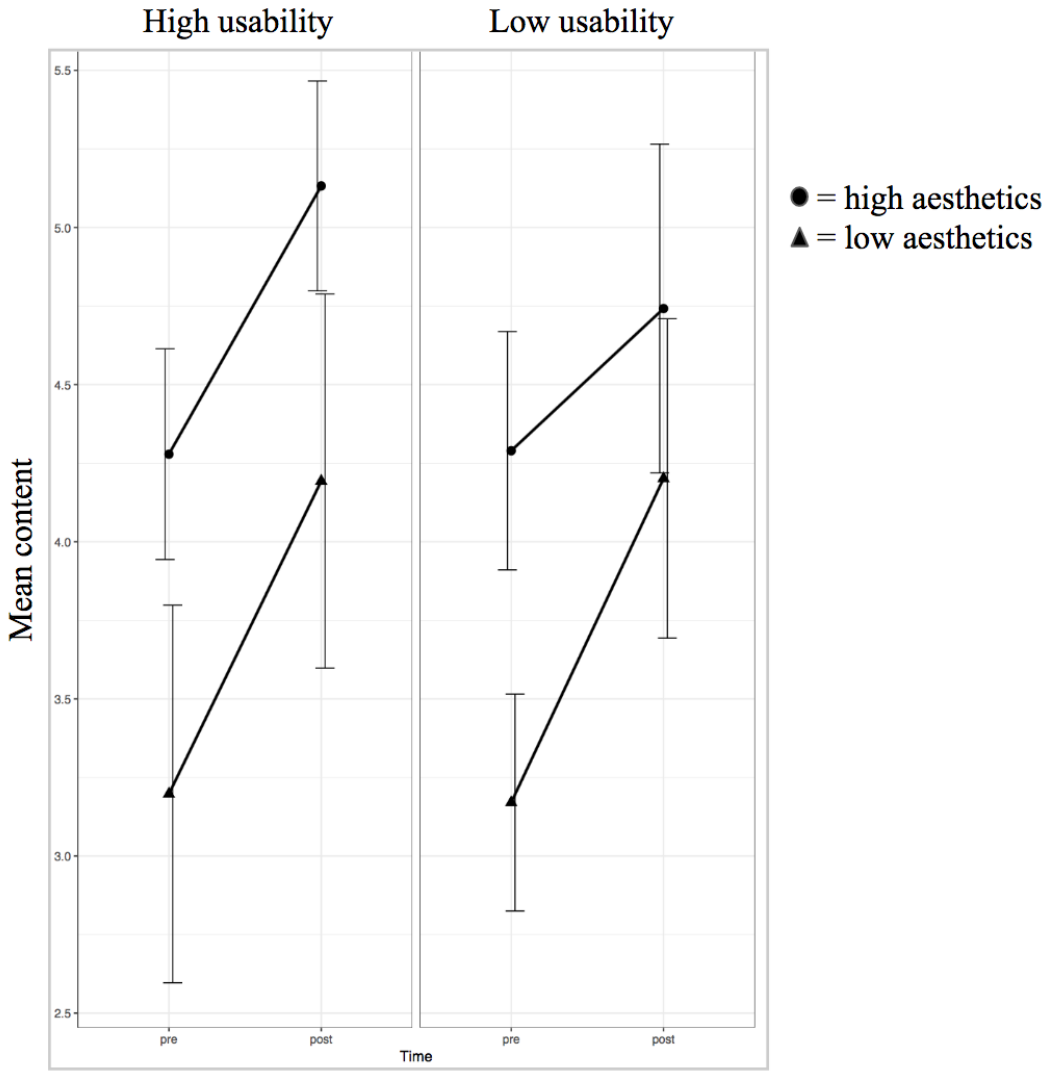

Figure 10. Content ratings for each level of aesthetics, usability and time. Error bars of the 95% confidence interval are visible.

conditions' dominance, LAHU's arousal and high usability conditions' valence that rise (see Figure 8). Valence also showed significant interaction for usability and time, $F(1, 89) = 8.49, p < .001, \eta^2 = .08$, resulting from a drop in low usability conditions while high usability conditions were stable or increased. For arousal, a significant interaction for usability and aesthetics over time arose, $F(1, 89) = 8.49, p = .04, \eta^2 = .04$, which could be caused by the rise of ratings from T0 to T1 and also for HAHU from T1 to T2, whereas the other conditions' feelings of arousal dropped from T1 to T2. Other interactions were not significant. Correlations within measures between T0 and T1 ranged from $r(91) = .74 - .79$, all significant with $p < .001$. Correlations between T0 and T1 with T2 ranged from $r(91)$

$= 0.44 - .61$, also all significant with $p < .001$. In Figure 11, the arousal dimension of affect is visualised over all three points in time to better understand the three-way interaction.

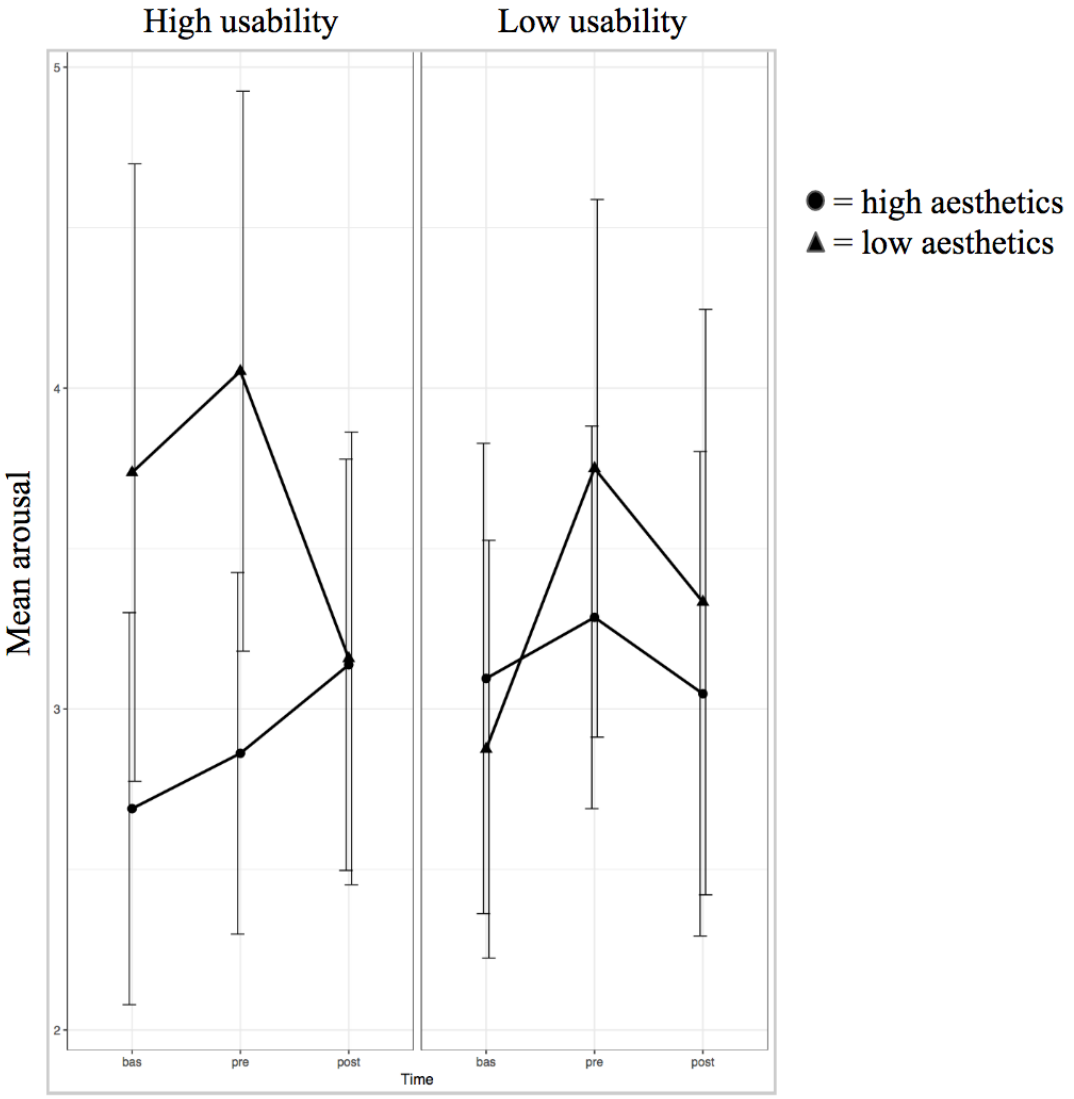

Figure 11. Arousal ratings for each level of aesthetics, usability and time. Error bars of the 95% confidence interval are visible.

Discussion

Integration of findings

M-commerce is growing in usage and thus presents an opportunity to increase order size and rate for online businesses (Wang et al., 2015). Consequently, knowledge of overall evaluations of mobile webshops is of key interest to increase success rate and profit. Also,

research profits from insights gained from the comparison of previous findings from desktop with mobile devices. Hereby, the model of Thielsch et al. (2013) is of special interest, presenting an integrative attempt to understand website evaluation with the core constructs contributing to it. Knowledge of causal effects on the one hand and stability of constructs playing a role at different points in time of evaluation on the other hand increases the value of the findings presented below. The first research question, whether aesthetics and usability influence construct-specific and general evaluations of mobile webshops at different points in time, can be answered positively. Aesthetics influenced perceived usability, but usability did not influence perceived aesthetics. Furthermore, aesthetics displayed a halo effect (Rosenzweig, 2014) for all ratings at first impression and continued to influence construct-specific ratings over time. These findings are in line with studies conducted on desktop which state that aesthetics plays a major role in first impressions (Tuch, Presslaber, et al., 2012, p.795). The second and third research question referred to the stability of a mobile webshop's evaluation over time regarding perceived content, aesthetics, usability, overall evaluative, as well as behavioral intention ratings. Between the two measuring times high correlations were found for the constructs aesthetics, usability, content and first and overall impression ratings. Perceived aesthetics and content ratings rose over time, showing an increase across conditions. High aesthetics followed by low usability showed a drastic decrease in overall evaluative ratings. If a mobile webshop does not fulfill good usability, customers do not want to revisit and recommend it, even though the first impression elicited by high aesthetics was good. Therefor, overall evaluative ratings were stable for conditions without conflicting manipulations. Most findings are in accordance with those from desktop computers, except the heightened influence of usability on mobile. This is the first study that offers empirical evidence for the comparison of evaluative patterns of webshops between desktop and mobile devices. Findings are discussed separately for each dependent variable and then compared to the model of Thielsch et al. (2013). Subsequently, limitations of the study will be addressed before recommendations for future research are given.

Aesthetics. Ratings of perceived aesthetics rose with longer exposure time within all conditions, showing that aesthetic perception intensified positively over time. The longer the

exposure, the more aesthetic the webshop is perceived. Results further suggest that this effect is independent of the usability of the webshop. Correlations to usability and content were high at both times of measure. The high correlation with first impression reduced considerably compared to overall impression, suggesting that over time, other constructs become more important. These findings are in line with research on desktop computers, where perceived aesthetics was identified as important for usability, first impressions, and, even though to a smaller extent on mobile, for overall impressions (Lindgaard et al., 2006; Sonderegger & Sauer, 2010; Thielsch & Hirschfeld, 2010; Tuch, Roth, et al., 2012; Van Schaik et al., 2012).

Usability. Ratings of perceived usability were influenced by the aesthetics manipulation at first impression. This points to a halo effect, as high aesthetics conditions displayed higher usability ratings, independent of the usability condition at both times of measure. However, once participants interacted with the mobile webshop, drastic rating changes occurred, as shown by the significant interaction of usability and time. The weight of usability's influence on overall evaluative ratings increased and correspondingly decreased for aesthetics. Objective usability in terms of task performance was also influenced by the usability manipulation as performance in low usability conditions was reduced. This underlines the importance of high usability in practice and is in line with findings from desktop (Tuch, Roth, et al., 2012).

First and overall impression. The significant main effect of aesthetics on first and overall impression reinforces indicators of a halo effect, with ratings being mostly higher for high aesthetics conditions. The significant interaction between aesthetics and time showed that interaction weakens the effect of aesthetics on the general evaluation of the webshop. Moreover, the interaction between usability and time indicates that if usability is low, interaction leads to a decrease in the overall evaluation while this effect does not occur if usability is high. In summary, these observations emphasize the importance of usability for overall evaluation of a webshop. These results reinforce desktop related findings stating that both constructs influence each other to a varying degree at different points of measure (Ben-Bassat et al., 2006; S. Lee & Koubek, 2012).

Intentions to revisit and recommend the webshop. The aesthetics halo effect was also observed at behavioral intention ratings but it disappears when interaction with a webshop of low usability occurs. High usability showed a positive influence on ratings after interaction, and even improved low ratings presumably deriving from low aesthetics.

Additional insights can be gained when considering the overall evaluative findings in combination. Primarily, conditions without conflicting manipulations only displayed highest or lowest ratings after interaction. This can be explained by high-level cognitions and reasoning taking longer to process than bottom-up perception (Leder et al., 2004), indicating that exposure times of one second may have been too short. Miniukovich and De Angeli (2014) further suggested that exposure time for first impressions have to be longer on mobile, because of the reduced screen size and subsequent display of fewer information. In addition, previous research on desktop found high stability of appeal ratings between first and overall impressions, which should highly correlate with overall evaluative ratings (Lindgaard et al., 2006; Tractinsky et al., 2006; Zhang & Li, 2004). This pattern can be confirmed in the mobile context, except when high aesthetics is followed by low usability. This implies that low usability can undermine the halo effect for overall evaluations. Consequently, if a mobile webshop does not fulfill good usability, customers do not want to revisit and recommend it, even though the first impression elicited by high aesthetics was good.

Content. Aesthetics had a significant main effect on content ratings. This also supports the halo effect, with high aesthetics conditions having higher ratings. Besides, aesthetics' influence on content has been found on desktop too (Hartmann et al., 2008). Longer exposure time is needed to cognitively process the content presented. Thus, the rise in content ratings over all conditions shows positive perception and independency of ratings from high or low usability. Correlations of content with first and overall impression remain stable over time.

Affect. Also for affect, aesthetics showed a significant main effect, with high aesthetics conditions having a more positive valence and a higher feeling of dominance. This is in line with research conducted on desktop (Mummalaneni, 2005) . Over time, high usability conditions' valence and low aesthetics conditions' dominance rose further. Affect changing over time supports it being an integral aspect of the experience (Boehner et al.,

2007). An additional finding was that low task performance could explain the drop in feelings of valence for low usability conditions after interaction. The rise of arousal between T0 and T1 can potentially be explained by an onset of stress since exposure time was perceived as very short by participants. This was also stated repeatedly in the open feedback (e.g., “The video was way too short, I could not see anything!”, P105, male, 24). Because the values at baseline differed quite strongly between the conditions, results should be interpreted with care and further investigations are recommended.

In the following, the findings of this study are compared to Thielsch’s model. Overall, correlations reported by Thielsch et al. (2013) find partial support in the influences reported in this study. In contrast to the model’s implications, the impact of aesthetics at first impression, as well as the influence of usability after interaction, are more pronounced. A potential explanation could be that due to differing exposure time, the halo effect from aesthetics was more pronounced in this study, also increasing usability’s correlation with first impression. Further investigation of exposure times is necessary, since Thielsch et al. (2013) do not provide information on it. For overall impression, the influence of usability and aesthetics was balanced for mobile. This contrasts to desktop, where usability showed much lower correlations than aesthetics. Regarding the intention to revisit and recommend a mobile webshop, usability had a significant influence on mobile. Usability increased ratings if it was high, and otherwise decreased them if it was low. This clearly contrasts with desktop, where no significant correlation was stated between usability and behavioral intentions in Thielsch’s model. Furthermore, correlations between content, aesthetics and usability differed slightly in the present study, but this could come from the specific content domain chosen, while Thielsch’s model presents correlations across content domains. Thus, the results advocate for high stability between construct-specific measures independent of device and time of measure.

From the findings of this work some immediate implications for practitioners in the m-commerce sector emerge:

- A high aesthetic appeal of mobile webshops has a strong, positive influence on the webshop's first and overall impression regarding aesthetic appeal and content as well as the first impression of usability and intentions to revisit and recommend.

- White, grey and blue colors and a clean, simple logo enhance the aesthetic appeal and rating of mobile (and desktop) webshops, which in turn positively influences other constructs which are critical to a good impression via the halo effect.

- Usability seems to be an important factor to uphold a positive first impression once interaction took place and can have a very strong, negative effect on overall ratings if not considered.

- Clear navigation labels and a well structured categorization of the products available increase the perceived usability.

Limitations

Certain limitations regarding the applied methods of the study conducted in this thesis have to be taken into account. The first limitation is the missing search function. Several participants complained about it in the open feedback (e.g., "A webshop without search function does not belong in the 21st century.", P28, male, 23). The feature was not implemented to ensure all participants would use the manipulated navigation. But seeing that so many participants desired a search function, it could have been perceived as usability flaw. A further limiting aspect is that participants complained about not having enough time to properly look at the webshop during the initial short exposure. Some stated to have relied on the few signals they could extract, which was perceived as stressful and confusing. Also, the initial exposure might have suffered from increased complexity since three mobile screenshots were shown simultaneously in a row on desktop instead of seeing them on a mobile device itself. This was done to ensure that enough stimulus material was seen during first exposure due to the reduced size of mobile screenshots, and to control the exposure time, since there was no technological solution available to ensure a precise and standardized exposure time on the mobile device. An additional limitation is that the ecological validity of the mobile

webshop might have been compromised due to standardized product features (e.g., fixed prices for product categories) in order to avoid bias from differing content. Also, all participants used the identical smartphone, a measure aimed at controlling resolution and navigation possibilities. Furthermore, motivational aspects to shop online for a book via mobile could not be taken into consideration in the laboratory setup with the fake webshops, but this aspect might also influence the user's evaluative pattern. The manipulations used in the study offer limited results, because only two levels per factor were used. Unfortunately, the tracking of most objective usability measures failed due to technical issues in the study. They would have generated additional insights into the different effects between perceived and objective usability. Also, they would offer performance metrics other than task score (e.g., time needed), to measure behavioral effects of manipulations. Finally, for the main study analysis, experts were excluded due to differing results as well as counterbalanced distribution across conditions. Bonnardel et al. (2011) pointed out that an experienced sample analyses a user interface differently from normal users (Bonnardel et al., 2011), but it is not clear in which way this might occur for mobile webshops. These limitations indicate possible directions of future research, as discussed in the next section.

Future research

The findings obtained from the study conducted in this work lead to implications for the experimental setup of studies which are concerned with evaluations of mobile webshops and related fields. Differing first impression exposure times can influence given ratings, depending on how much can be perceived during the exposure. Experimenting with different exposure times can yield further insights for mobile, as it did for desktop devices, regarding parameters such as rating stability (Tuch, Presslaber, et al., 2012). With evolving software for mobile and improved connectivity, the first impression could be measured directly on mobile, increasing ecological validity. Concerning manipulations, the study could profit from a systematic content variation to detect domain-specific differences in evaluation patterns, as was attempted on desktop in the third study of Thielsch et al. (2013), although in that case no differences were reported. Likewise, from multi-level manipulations of aesthetics and usability, as well as

from the use of other design elements influencing the constructs, conclusions about the specificity of findings, also depending on their strength, could be drawn. This has already been suggested by (Tuch, Roth, et al., 2012, p. 7): “With boundary conditions we describe the possibility that different degrees of usability and aesthetics manipulation may affect the aesthetics-usability relation differently.”. Tracking objective usability could provide further insights into performance and differences between objective and subjectively perceived usability. Another beneficial addition to the methodological setup would be measures of actual buying behavior and validated dependent measures adjusted for mobile devices. Using an existing webshop or varying more elements in a created one (e.g. price or ratings) can increase ecological validity. Implementing more variations and features as well as using personal mobile devices might increase ecological validity and create a more realistic feel for the shopping experience. Finally, while for the analysis of this study only the sum scores of the self-reported measures were of interest, future research could also delve into the single facets of aesthetic appeal and content ratings to investigate the findings on a finer level. In addition, diverging results from an expert subsample could be rechecked with more expert participants. Generally speaking, the field of HCI would benefit from more in-depth mobile studies to enable cross-device comparisons and to reveal similarities and differences between devices. With such additional findings, recommendations for practitioners could be collected to increase the value of research findings for commercial purposes.

Conclusion

Due to the popularity of smartphone usage, m-commerce has grown rapidly over the past years (Chang et al., 2014). This makes knowledge of the perception of constructs of which the evaluation of mobile webshops consist a key point of interest. On desktop, a model developed by Thielsch et al. (2013) showed the correlations of content, aesthetics and usability with overall evaluative ratings after short exposure time and longer interaction. To extend this model, this thesis investigated causal effects of aesthetics and usability on mobile webshop evaluations over time to further determine stability of ratings. These objectives were achieved with an experimental setup and dependent measures applied twice, after short

exposure time and after longer interaction. Data yielded from 93 participants showed that the evaluation pattern found by Thielsch et al. (2013) is only partially applicable in the context of m-commerce. Aesthetics showed a halo effect on all dependent ratings after short exposure. This effect lessened for overall evaluative ratings once participants interacted with the webshop. Perceived aesthetics and content ratings continued to increase over time and usability had a stronger influence on behavioral intentions for mobile than for desktop, especially when it was low. In order to be revisited and recommended, a mobile webshop needs to fulfill good usability, even if the first impression elicited by high aesthetics was good. Without conflicting manipulations, the ratings submitted by the participants of the study were quite stable and correlating highly between the two times of measure. Affect changed over time, indicating that it is an integral aspect of the evaluative experience. The aforementioned findings highlight the importance of investigating data across different devices to enable judgement whether or not overarching or device-specific effects arise. The manipulations applied in this study yield design implications for m-commerce practitioners, to utilize the halo effect detected from high aesthetics after short exposure time, as well as the maintenance of this positive impression through high usability.

References

Aldrich, J. (1995). Correlations genuine and spurious in pearson and yule. *Statistical science*, 364–376.

Ben-Bassat, T., Meyer, J., & Tractinsky, N. (2006). Economic and subjective measures of the perceived value of aesthetics and usability. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *13*(2), 210–234.

Benou, P., & Bitos, V. (2008). Developing mobile commerce applications. *Journal of electronic commerce in organizations*, *6*(1), 74–88.

Billi, M., Burzagli, L., Catarci, T., Santucci, G., Bertini, E., Gabbanini, F., & Palchetti, E. (2010). A unified methodology for the evaluation of accessibility and usability of mobile applications. *Universal Access in the Information Society*, *9*(4), 337–356.

Bloch, P. H., Brunel, F. F., & Arnold, T. J. (2003). Individual differences in the centrality of visual product aesthetics: Concept and measurement. *Journal of consumer research*, *29*(4), 551–565.

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies*, *65*(4), 275–291.

Bonnardel, N., Piolat, A., & Le Bigot, L. (2011). The impact of colour on website appeal and users' cognitive processes. *Displays*, *32*(2), 69–80.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49–59.

Campbell, A., & Pisterman, S. (1996). A fitting approach to interactive service design: The importance of emotional needs. *Design Management Review*, *7*(4), 10–14.

Chang, J. M., Williams, J., & Hurlburt, G. (2014). Mobile commerce. *IT Professional*, *16*(3), 4–5.

Chittaro, L. (2006). Visualizing information on mobile devices. *Computer*, *39*(3), 40–45.

Cober, R. T., Brown, D. J., Levy, P. E., Cober, A. B., & Keeping, L. M. (2003). Organizational web sites: Web site content and style as determinants of organizational attraction. *International Journal of Selection and Assessment*, *11*(2-3), 158–169.

Consortium, W. W. W., et al. (2008). Web content accessibility guidelines (wcag) 2.0.

Cyr, D., Head, M., & Larios, H. (2010). Colour appeal in website design within and across cultures: A multi-method evaluation. *International journal of human-computer studies*, *68*(1), 1–21.

Flavián, C., Guinalíu, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management*, *43*(1), 1–14.

Fredrickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion*, *14*(4), 577–606.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 345–352).

Gross, M. (2015). Mobile shopping: a classification framework and literature review. *International Journal of Retail & Distribution Management*, *43*(3), 221–241.

Harrison, R., Flood, D., & Duce, D. (2013). Usability of mobile applications: literature review and rationale for a new usability model. *Journal of Interaction Science*, *1*(1), 1.

Hartmann, J., Sutcliffe, A., & Angeli, A. D. (2008). Towards a theory of user judgment of aesthetics and user interface quality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *15*(4), 15.

Hassenzahl, M., Diefenbach, S., & Göritz, A. (2010). Needs, affect, and interactive products–facets of user experience. *Interacting with computers*, *22*(5), 353–362.

Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior*, *26*(6), 1278–1288.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies*, *64*(2), 79–102.

Huizingh, E. K. (2000). The content and design of web sites: an empirical study. *Information & Management*, *37*(3), 123–134.

ISO, W. (1998). 9241-11. ergonomic requirements for office work with visual display terminals (vdts). *The international organization for standardization*, *45*.

ISO, W. (2006). 9241-151. ergonomics of human-system interaction — part 151: Guidance on world wide web interfaces. *The international organization for standardization*.

Jennings, M. (2000). Theory and models for creating engaging and immersive ecommerce websites. In *Proceedings of the 2000 acm sigcpr conference on computer personnel research* (pp. 77–85).

Kemp, S. (24.01.2017). *Digital in 2017: Global overview.* Retrieved from `https://wearesocial.com/special-reports/digital-in-2017-global-overview`

Kotler, P. (1973). Atmospherics as a marketing tool. *Journal of retailing*, *49*(4), 48–64.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

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

Lee, S., & Koubek, R. J. (2012). Users' perceptions of usability and aesthetics as criteria of pre-and post-use preferences. *European Journal of Industrial Engineering*, *6*(1), 87–117.

Lee, Y.-K., Chang, C.-T., Lin, Y., & Cheng, Z.-H. (2014). The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress. *Computers in Human Behavior*, *31*, 373–383.

Levin, A. M., Levin, I. R., & Heath, C. E. (2003). Product category dependent consumer preferences for online and offline shopping features and their influence on multi-channel retail alliances. *J. Electron. Commerce Res.*, *4*(3), 85–93.

Li, Y.-M., & Yeh, Y.-S. (2010). Increasing trust in mobile commerce through design aesthetics. *Computers in Human Behavior*, *26*(4), 673–684.

Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., & Noonan, P. (2011). An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *18*(1), 1.

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.

Lowry, P. B., Wilson, D. W., & Haig, W. L. (2014). A picture is worth a thousand words: Source credibility theory applied to logo and website design for heightened credibility and consumer trust. *International Journal of Human-Computer Interaction*, *30*(1), 63–93.

Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the Association for Information Science and Technology*, *58*(13), 2078–2091.

Michailidou, E., Harper, S., & Bechhofer, S. (2008). Visual complexity and aesthetic perception of web pages. In *Proceedings of the 26th annual acm international conference on design of communication* (pp. 215–224).

Miniukovich, A., & De Angeli, A. (2014). Visual impressions of mobile app interfaces. In *Proceedings of the 8th nordic conference on human-computer interaction: Fun, fast, foundational* (pp. 31–40).

Moshagen, M., & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies*, *68*(10), 689–709.

Mummalaneni, V. (2005). An empirical investigation of web site characteristics, consumer emotional states and on-line shopping behaviors. *Journal of Business Research*, *58*(4), 526–532.

Naur, P. (1965). The place of programming in a world of problems, tools, and people. In *Proceedings of the ifip congress* (Vol. 65, pp. 195–199).

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175.

Nielsen, J., & Levy, J. (1994). Measuring usability: preference vs. performance. *Communications of the ACM*, *37*(4), 66–75.

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*(2), 315–324.

Ou, C. X., & Sia, C. L. (2010). Consumer trust and distrust: An issue of website design. *International Journal of Human-Computer Studies*, *68*(12), 913–934.

Palmer, J. W. (2002). Web site usability, design, and performance metrics. *Information systems research*, *13*(2), 151–167.

Partala, T., & Saari, T. (2015). Understanding the most influential user experiences in successful and unsuccessful technology adoptions. *Computers in Human Behavior*, *53*, 381–395.

Porat, T., & Tractinsky, N. (2012). It's a pleasure buying here: The effects of web-store design on consumers' emotions and attitudes. *Human–Computer Interaction*, *27*(3), 235–276.

Reichheld, F. F., & Markey, R. (2011). *The ultimate question 2.0: How net promoter companies thrive in a customer-driven world*. Harvard Business Press.

Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., & Gajos, K. Z. (2013). Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2049–2058).

Rosenzweig, P. (2014). *The halo effect:... and the eight other business delusions that deceive managers*. Simon and Schuster.

Schenkman, B. N., & Jönsson, F. U. (2000). Aesthetics and preferences of web pages. *Behaviour & Information Technology*, *19*(5), 367–377.

Schloss, K. B., & Palmer, S. E. (2011). Aesthetic response to color combinations: preference, harmony, and similarity. *Attention, Perception, & Psychophysics*, *73*(2), 551–571.

Schminder, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? reinvestigating the robustness of anova against violations of the normal distribution. *European Research Journal of Methods for the Behavioral and Social Sciences*, *6*, 147–151.

Seckler, M., Opwis, K., & Tuch, A. N. (2015). Linking objective design factors with subjective aesthetics: an experimental study on how structure and color of websites affect the facets of users' visual aesthetic perception. *Computers in Human Behavior*, *49*, 375–389.

Seth, E. (2014). Mobile commerce: A broader perspective. *IT Professional*, *16*(3), 61–65.

Smith, S. L., & Mosier, J. N. (1986). *Guidelines for designing user interface software*. Mitre Corporation Bedford, MA.

Sohn, S., Seegebarth, B., & Moritz, M. (2017). The impact of perceived visual complexity of mobile online shops on user's satisfaction. *Psychology & Marketing*, *34*(2), 195–214.

Sonderegger, A., & Sauer, J. (2010). The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied ergonomics*, *41*(3), 403–410.

Spangenberg, E. R., Grohmann, B., & Sprott, D. E. (2005). It's beginning to smell (and sound) a lot like christmas: the interactive effects of ambient scent and music in a retail setting. *Journal of business research*, *58*(11), 1583–1589.

Tarasewich, P., Daniel, H. Z., & Griffin, H. E. (2001). Aesthetics and web site design. *Quarterly Journal of Electronic Commerce*, *2*, 67–82.

Thielsch, M. T. (2008). Ästhetik von websites. *Wahrnehmung von Ästhetik und deren Beziehung zu Inhalt, Usability und Persönlichkeitsmerkmalen. Münster: MV Wissenschaft*.

Thielsch, M. T., Blotenberg, I., & Jaron, R. (2013). User evaluation of websites: From first impression to recommendation. *Interacting with Computers*, *26*(1), 89–102.

Thielsch, M. T., & Hirschfeld, G. (2010). High and low spatial frequencies in website evaluations. *Ergonomics*, *53*(8), 972–978.

Thielsch, M. T., & Hirschfeld, G. (2012). Spatial frequencies in aesthetic website evaluations–explaining how ultra-rapid evaluations are formed. *Ergonomics*, *55*(7), 731–742.

Tractinsky, 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*(11), 1071–1083.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with computers*, *13*(2), 127–145.

Tractinsky, N., & Lowengart, O. (2007). Web-store aesthetics in e-retailing: A conceptual

framework and some theoretical implications. *Academy of Marketing Science Review*, *2007*, 1.

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.

Tuch, A. N., Roth, S. P., Hornbæk, K., Opwis, K., & Bargas-Avila, J. A. (2012). Is beautiful really usable? toward understanding the relation between usability, aesthetics, and affect in hci. *Computers in Human Behavior*, *28*(5), 1596–1607.

Tuch, A. N., Trusell, R., & Hornbæk, K. (2013). Analyzing users' narratives to understand experience with interactive products. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2079–2088).

Van Schaik, P., Hassenzahl, M., & Ling, J. (2012). User-experience from an inference perspective. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *19*(2), 11.

Van Schaik, P., & Ling, J. (2009). The role of context in perceptions of the aesthetics of web pages over time. *International Journal of Human-Computer Studies*, *67*(1), 79–89.

Wang, R. J.-H., Malthouse, E. C., & Krishnamurthi, L. (2015). On the go: How mobile shopping affects customer purchase behavior. *Journal of Retailing*, *91*(2), 217–234.

Zhang, P., & Li, N. (2004). Love at first sight or sustained effect? the role of perceived affective quality on users' cognitive reactions to information technology. *ICIS 2004 Proceedings*, 22.

Appendix A

Links to the German Book Webshop Versions

High aesthetics - high usability version

https://dev.psycho.unibas.ch/?c=1&d=1

High aesthetics - low usability version

https://dev.psycho.unibas.ch/?c=2&d=1

Low aesthetics - high usability version

https://dev.psycho.unibas.ch/?c=1&d=2

Low aesthetics - low usability version

https://dev.psycho.unibas.ch/?c=2&d=2

Appendix B

Links to the R Script and Dataset of the Main Study

Dataset:

https://drive.google.com/open?id=0B3xxmrC3uE6ncFhPbkhwa0pCbnc

R Script (worked with R Version 1.0.136):

https://drive.google.com/open?id=0B3xxmrC3uE6nd1o1ek1nMFE4ZVU