Int. J. Human–Computer Studies 165 (2022) 102837

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

# Visual aesthetics and user experience: A multiple-session experiment

Juergen Sauer [a,*], Andreas Sonderegger [a,b]

[a] Department of Psychology, University of Fribourg, Rue de Faucigny 2, Fribourg 1700, Switzerland
[b] Business School, Institute for New Work, Bern University of Applied Sciences, Bern, Switzerland

ARTICLE INFO

*Keywords:*
User experience
Usability
Aesthetics
Sensory modality

ABSTRACT

The article reports a longitudinal lab experiment, in which the influence of product aesthetics and inherent product usability was examined over a period of 7 weeks. Using a 2 × 2 × 7 mixed design, visual aesthetics (high vs. low) and usability (high vs. low) were manipulated as between-subjects variables whereas exposure time was used as a repeated-measures variable. One hundred and ten participants took part in the study, during which they carried out typical tasks of operating a fully automated coffee machine. We measured user experience by using the following outcome variables: perceived usability, perceived attractiveness, performance, affect, workload and perceived coffee quality (gustatory aesthetics). We found no effect of visual aesthetics on user experience (including perceived usability as the chief outcome variable), which is in contrast to a considerable number of previous studies. The absence of such an effect might be associated with influencing factors that have not yet been given sufficient attention (e.g., user identification with product, sensory dominance, characteristics of specific products).

## 1. Introduction

For a considerable number of years, the influence of product aesthetics on perceived usability has been the subject of research (e.g., Moshagen et al., 2009; Mahlke and Thüring, 2007). This pattern of influence has been termed the 'what-is-beautiful-is-good'-effect (Tractinsky et al., 2000). More recently, the opposite type of influence has also been the subject of research (though much less prominently), which refers to the influence of product usability on perceived aesthetics (e.g., Tuch et al., 2012; Hamborg et al., 2014). This has been termed the 'what-is-good-is-beautiful'-effect accordingly (Tuch et al., 2012).

Much of this work has been driven by concerns that the outcomes of usability tests may be affected by factors that are entirely unrelated to product usability. Although users typically interact with a device over a longer usage period, remarkably few studies have examined user-device interaction over an extended duration. To our knowledge, only few studies have adopted a longitudinal approach in that field (Kujala and Miron-Shatz, 2013; Sonderegger et al., 2012). However, despite some methodological concerns that were associated with the completion of single-session studies (e.g., Kujala and Miron-Shatz, 2013), little research has subsequently heeded the call for more longitudinal studies. The present study contributes to overcoming this dearth of research.

### 1.1. Product aesthetics and user experience

While product aesthetics as an objective component refers to the degree to which a product triggers pleasure from sensory perception (Hekkert and Leder, 2008), the term 'user experience' as a subjective component refers to the degree to which users react to a product or artefact in terms of their actions, sensations, considerations and feelings (Wright et al., 2003). Previous work has primarily focused on perceived usability as the primary outcome variable of user experience (often without considering perceived usability as an element of user experience). It is acknowledged that the definitions of user experience and usability (and their relationship to each other) are not unanimously agreed upon in the research community (see Sauer et al., 2020).

When reviewing the research literature, it emerges that the 'what-is-beautiful-is-good'-effect is widely reported. A range of studies has already examined the influence of product aesthetics on perceived usability, reporting a positive relationship in correlation studies (e.g., Kurosu and Kashimura, 1995; Tractinsky et al., 2000; Hartmann et al., 2007; Schenkman and Jönsson, 2000). A considerable number of experimental studies were also conducted, being able to establish a cause-effect relationship between aesthetics and perceived usability (e.g., Brady and Phillips, 2003; Nakarada-Kordich and Lobb, 2005; Ben-Bassat et al., 2006; Moshagen et al., 2009; Sonderegger and Sauer 2010;

* Corresponding author.
*E-mail address:* juergen.sauer@unifr.ch (J. Sauer).

https://doi.org/10.1016/j.ijhcs.2022.102837
Received 19 November 2020; Received in revised form 27 February 2022; Accepted 6 April 2022
Available online 9 April 2022

Sauer and Sonderegger, 2009). Some of the experimental studies found this effect only prior to user-device interaction, with the effect disappearing after users gained some experience with the product (Minge and Thüring, 2018). Interestingly, the 'what-is-beautiful-is-good'-effect may not only apply to perceived usability but also to objective user performance, with a recent meta-analysis confirming a small significant effect of aesthetics (Thielsch et al., 2019).

The 'what-is-beautiful-is-good'-effect was found across different technical devices. So far, research was able to demonstrate this effect in several artefacts, including mobile phones (e.g., Sonderegger and Sauer 2010; MP3-players (Mahlke and Thüring, 2007), web pages (Brady and Phillips, 2003) and electronic phonebooks (Ben-Bassat et al., 2006). This suggests that the effects are observed for tangible as well as for non-tangible products. In addition to its generalisability across artefacts, the effect was also found in many different cultures, including Japan (Kurosu and Kashimura, 1995), Israel (Tractinsky et al., 2000; Ben--Bassat et al., 2006), Germany (e.g., Moshagen et al., 2009; Mahlke and Thüring, 2007), and Switzerland (e.g., Sonderegger and Sauer, 2010). Despite the widespread occurrence of the 'what-is-beautiful-is-good'-effect, there are a smaller number of studies in the literature, in which the effect did not occur (Tuch et al., 2012; Hamborg et al., 2014; Van Schaik and Ling, 2009). Again, there appears to be no difference between tangible and non-tangible products, with both types being represented in the studies cited.

To understand the psychological mechanism underlying the 'what-is-beautiful-is-good'-effect, the halo effect has often been used as a theoretical explanation (e.g., Sutcliffe and Namoune, 2008; Tractinsky et al., 2000). The halo effect assumes that salient characteristics of an object (or a person) have an influence on how less salient characteristics are perceived. Product aesthetics is a salient characteristic since it is perceived very quickly compared to usability characteristics, which require a certain degree of interaction with the product before a judgement can be made. Other work has referred to the processing fluency theory (Reber et al., 2004; Thielsch and Hirschfeld, 2010) to provide an explanation for this pattern of findings. However, recent evidence has discounted the possibility of processing fluency theory playing a role here (Bölte et al., 2017). While there may be some debate about the psychological mechanism underlying the 'what--is-beautiful-is-good'-effect, there is little doubt about the mechanism being highly robust, at least if it is examined over a short period (e.g., single testing session). This methodological qualification is of high importance because if a device is tested over a longer period, the positive effect of an aesthetically appealing product on perceived usability may vanish over time (Sonderegger et al., 2012). A recent study showed that even in a single testing session, such an effect may occur if the repeated use of a device is modelled (Lee and Ha, 2019). This work showed that the negative influences of an aesthetically unappealing device on perceived usability begins to weaken when the device is repeatedly used.

### 1.2. Inherent product usability and perceived aesthetics

The inverse relationship has also been examined in a smaller number of studies. Sometimes referred to as the 'what-is-good-is-beautiful'-effect (e.g., Tuch et al., 2012), this relationship describes the effect product usability might have on perceived aesthetics. The term 'product usability' is sometimes also referred to as 'inherent usability' (e.g., Fu et al., 2002) to emphasise that it refers to an inherent feature of the device rather than the user's perception of product qualities. Some research revealed a positive effect of product usability on perceived aesthetics of the product (Tuch et al., 2012; Hamborg et al., 2014; Minge and Thüring, 2018). However, other work did not find such a relationship (Sonderegger et al., 2012). Given the rather small number of studies that empirically examined the effect, it is difficult to state with some confidence the reasons for the inconsistent findings in the literature. Considering that the study that did not find the 'what-is-good-is-beautiful'-effect was a longitudinal study (whereas the three studies that did find the effect were not), it might suggest a methodological bias of the results, in a similar way as for the 'what-is-beautiful-is-good'-effect.

### 1.3. Visual and non-visual aesthetics

The discussions surrounding the 'what-is-beautiful-is-good'-effect have been complemented by recent research that addressed the question of the moderating influence of the sensory modality. The vast majority of studies cited above examined visual aesthetics rather than aesthetics that is perceived by any of the other sensory modalities (i.e. audition, smell, taste and touch). The definition of product aesthetics as an objective component of the artefact (see Hekkert and Leder, 2008) applies not only to visual aesthetics but also to non-visual aesthetics. For example, Postrel (2003) refers to tone and texture as non-visual elements in addition to colour and form. Put differently, both types of product aesthetics (or beauty) are 'objectified, because people experience beauty as something that lies in an object, rather than exclusively being the result of a positive sensation of the body' (Moshagen and Thielsch, 2010; p. 689). The positive (or negative) sensation of the body would refer to the human response to these product features, representing the degree to which a human believes that the device is aesthetically pleasing (van der Heijden, 2003), be it visual or non-visual in nature. In this article, we will refer to the concepts of 'product aesthetics' (e.g., see Hekkert and Leder, 2008) and 'perceived attractiveness' to make the important distinction between objective design aspects of the product and the human response to it.

The distinction between visual and non-visual aesthetics refers to the sensory modality that is used to perceive the object (visual vs. audition, smell, taste or touch). An important difference between visual and non-visual aesthetics concerns the high speed with which the visual attractiveness of an artefact may be judged (e.g., Lindgaard et al., 2006; Thielsch and Hirschfeld, 2012), which is in contrast to the longer perception processes of non-visual stimuli. Furthermore, the difference between visual and non-visual aesthetics may result in different patterns of findings. For example, empirical work examining different forms of non-visual aesthetics was unable to replicate the 'what--is-beautiful-is-good'-effect in a series of studies manipulating haptic and auditory product perception (Sonderegger and Sauer, 2015). Visual perception may also influence non-visual processes in that non-visual aesthetics could also be evaluated when looking at a picture. For example, based on previous experience a user might evaluate the metal surface of a device to be haptically very pleasant without having actually experienced the sensation of this particular device (though the evaluation of non-visual features by merely looking at the device may be deceptive).

When judging the aesthetic appeal of a device, the distinction between visual and non-visual aesthetics cannot always be that clearly made. For most products, more than one sensory modality is involved. Schifferstein (2006) provides a list of products, which were rated by users with regard to the importance of each of the five modalities. For example, for coffeemakers vision and audition were considered of high importance, followed by touch and smell (both of intermediate importance), and taste (low importance). For telephone operation, the visual modality was considered of similar importance as the haptic and auditory modality while smell and taste were only of very low importance. One may expect that for modern smartphones the visual modality will gain in relative importance in relation to other modalities considering that Schifferstein investigated traditional telephones. Overall, the importance users attach to each sensory modality is expected to influence the way the aesthetics of this product has an effect on other variables.

### 1.4. Multi-session versus single-session approach

Despite some concerns associated about the use of single-session studies as a methodological approach (e.g., Kujala and Miron-Shatz, 2013), this approach is still predominant in this research strand, with little work having heeded the call for more longitudinal studies. To our knowledge, only very few studies have adopted such a multi-session approach. In one study, users tested a mobile phone over two weeks to determine how user experience and behaviour change during that time (Sonderegger et al., 2012). In a second study, a mobile phone was tested even over a period of five months (Kujala and Miron-Shatz, 2013). Interestingly, in a further study adopting a single-session approach but with repeated exposures to a website (Lee and Ha, 2019), similar effects were observed as in the longitudinal studies in that positive effects of aesthetics on the outcomes of usability tests have vanished over time.

The argument for the multi-session approach can also be made from a theoretical point of view because of the user-product relationship continuously evolving. Karapanos et al. (2009) argued that the nature of user-product interaction changes as a function of three phases (orientation, incorporation and identification). In the orientation phase, the user may experience excitement about the new product but also some frustration because some product features are not easily usable. In the incorporation phase, a more long-term perspective is adopted, with the device's usefulness becoming a predominant factor for device evaluation. In the identification phase, the device is accepted as part of the user's life, becoming part of the user's identity. While the validity of this three-phase model needs to be determined still, it points out the important issue of on-going changes in the user-product relationship, which may also affect the emergence of the two effects in question (i.e. 'what-is-beautiful-is-good'-effect and the 'what-is-good-is-\beautiful '-effect). All these points raise the question of whether a methodological artefact is at the root of the problem, resulting in the risk of overestimating the size of the 'what-is-beautiful-is-good'-effect, and possibly the 'what-is-good-is-beautiful'-effect.

### 1.5. The present study

The goal of this longitudinal experiment was to determine the influence of aesthetic appeal and inherent usability of a product on user experience, captured during the completion of a multi-session usability test. It used a different device than preceding longitudinal studies (coffee machine rather than mobile phone) to determine whether similar effects are observed across different devices.

Like the previous longitudinal studies, it was aimed at examining the research questions under more realistic usage conditions involving an extended usage phase that goes beyond a one-off usability test. This was achieved by asking participants to pay regular visits to the lab (i.e. once a week over a period of seven weeks) to complete a set of typical user tasks. The experiment protocol was aimed at obtaining a comprehensive measure of user behaviour and user reactions when operating the device. We used standardised tasks for performance testing and a range of self-report measures as variables (e.g., usability, aesthetics, workload, and affect). Being based on the concept of user experience, these self-report measures go beyond traditional measures used in usability testing (e.g. satisfaction). The self-report measures were largely based on established scales but we also included self-developed scales comprising only a very small number of items with a view to limiting the length of each experimental session to 15 min and hence maintaining participant motivation over the full duration of the study.

Based on the literature review, several hypotheses were put forward. Hypothesis 1A postulated that higher aesthetic appeal of the device was expected to result in higher levels of perceived usability than if a less aesthetically appealing device was used. This assumption was based on a number of studies having examined the 'what-is-beautiful-is-good'-effect (e.g., Tractinsky et al., 2000; Moshagen et al., 2009). Hypothesis 1B assumed that the effect of aesthetics on perceived usability would become smaller in magnitude with increasing exposure time. Such changes over time have been observed in previous studies (Lee and Ha, 2019; Sonderegger et al., 2012). Hypothesis 2A predicted that poorer inherent usability would lead to lower usability ratings, to lower attractiveness ratings, and to higher levels of negative emotion. Such effects would correspond to the 'what-is-good-is-beautiful'-effect, which was empirically supported by a number of studies (e.g., Tuch et al., 2012; Hamborg et al., 2014). Finally, Hypothesis 2B postulated that these effects were expected to become smaller with increasing exposure time. This assumption is based on the findings of a previous longitudinal study, which failed to establish this effect after several testing sessions (Sonderegger et al., 2012).

## 2. Method

### 2.1. Participants

For this study, we recruited 110 participants (88% male) from the University of Fribourg. Their ages ranged from 18 to 44 years ($M$ = 21.98, $SD$ = 4.09). Most of them were undergraduate psychology students (91%), the others were members of staff from the same department.

Participants did not receive any financial compensation for taking part in the study. However, they were offered a take-away coffee after each testing session in the laboratory. The psychology students received additional course credits for their participation. Ownership of a fully automated coffee machine amongst participants was rather low (12.7%). None of the participants was in possession of the particular model used in the study.

### 2.2. Experimental design

A 2 × 2 × 7 mixed design with the following independent variables was implemented: product aesthetics (high/low), inherent usability (high/low) and exposure time (measured over 7 weeks). Product aesthetics and inherent usability were manipulated as between-subjects variables and exposure time was a within-subjects variable.

### 2.3. Measures and instruments

*Perceived attractiveness.* Two items measuring perceived visual attractiveness were developed by the authors for this study. The first enquired about the aesthetics of the general design ('The design of the coffee machine is aesthetically appealing'), and the second more specifically about the screen of the machine ('The design of the display is aesthetically appealing'). For both items, 20-point Likert scales were used, ranging from 1 (strongly disagree) to 20 (strongly agree). We calculated Spearman-Brown coefficients as an indicator of internal consistency for the two-item scale (Eisinga et al., 2013). Coefficients ranged between $\rho' = 0.61$ and $\rho' = 0.75$, which represented rather low reliabilities.

*Perceived usability.* The well-established system usability scale (SUS; Brooke, 1996) was administered to measure user assessments of product usability. Comprising ten items, each item makes use of a 5-point Likert scale. A sample item is: 'I thought the system was easy to use'. The internal consistency of the scale is reported to be very good (i.e. Cronbach's $\alpha > 0.90$; Bangor et al., 2008; Brooke, 2013). A comparison of the SUS scores with usability outcomes suggested that it was a valid instrument (Peres et al., 2013). In the present experiment, internal consistency ranged from $\alpha = 0.86$ to $\alpha = 0.89$ across the seven weeks.

*Perceived coffee quality.* To rate the quality of the coffee (as a measure of gustatory attractiveness), a 9-point Likert scale questionnaire was employed, which is commonly used by a manufacturer of coffee machines (Jura™) for their in-house coffee testing, measuring coffee characteristics on 9 different items. The scale was modified for the purpose of this study, with four of the nine items being retained

(intensity (weak - high), acidity (little - much), bitterness (little - much), and overall rating (poor - very good). The internal consistency of the scale was found to be rather low across the seven measurement points (Cronbach's $\alpha_{min} = 0.04$; $\alpha_{max} = 0.39$). This may not be surprising given that different users prefer different facets of coffee quality (e.g., some prefer high levels of acidity while others prefer high levels of bitterness or intensity). Therefore, we focussed on the overall rating and did not consider the other facets of perceived coffee quality for statistical analyses.

*Perceived workload.* Perceived workload was assessed by using the NASA-TLX (Hart and Staveland, 1988). Employing a 20-point Likert scale (1 = very low, 20 = very high) users rated the 6 items of this scale (e.g., 'How mentally demanding was the task?'). As a well-established instrument for workload measurement, the NASA-TLX has acceptable psychometric properties (Cronbach's $\alpha > 0.80$; Xiao et al., 2005), though the internal consistency was found to be considerably lower in the present experiment ($\alpha_{min} = 0.34$; $\alpha_{max} = 0.61$).

*Emotional distress.* Affect was assessed by using the distress dimension of the Short Stress State Questionnaire (SSSQ; Helton, 2004, Helton and Naswall, 2015). It assesses negative affective states in the form of emotional distress by using eight adjectives (e.g., upset, impatient, angry, irritated), which are to be rated by test participants. As response format, we used a 5-point Likert scale, ranging from 'not at all' to 'extremely'. Previous research found the psychometric properties of the scale to be satisfactory (Cronbach's α ranging from 0.87 to 0.89; Helton and Naswall, 2015). In the present experiment, it varied between α = 0.78 and α = 0.87.

*User performance.* User performance was measured in two ways. Task completion time was measured in seconds. Interaction efficiency (indicating deviations from the optimal user-device dialogue) was computed by subtracting the minimum number of user inputs required to complete the task from the actual number of user inputs made.

### 2.4. Materials

Two fully automated coffee machines (Jura™ IMPRESSA Z9) were used in the experiment (see versions 'Original' and 'Chaos' in Fig. 1). To avoid possible influences of product aesthetics, both machines were covered with a matt-black sticker foil. Due to the high price of the product (CHF 2750; about € 2500) and its limited edition, none of the participants was in possession of such a coffee machine.

'100% Premium Arabica' coffee beans (Jacobs™ Médaille d'Or) were chosen for the experiment. This choice was based on a rating of five experts assessing this coffee out of a range of alternative options. The coffee beans used in the experiment were required to be appropriate for small coffees (e.g., Espresso or Ristretto) as well as large ones (e.g., Latte Macchiato or Americano) because of the complementary coffee to be taken away after completing the experiment.

Transparent glass cups (70 ml) were used for the experiment. Plastic cups, milk and sugar were provided for the complementary coffee to be taken away after task completion.

Two HP™ computers and Dell™ monitors were used for the participants to fill in the questionnaires on the platform 'unipark.com', using the software 'EFS Survey'™. A Logitech™ web cam was set up next to the machine, capturing the participants' interaction with the machine (hand gestures were recorded but no facial expressions).

### 2.5. Pilot studies and experimental manipulation of coffee machine

We carried out two pilot studies in order to evaluate the appropriateness of the experimental manipulations. The manipulations involved changing the visual appearance of the coffee machine and modifying its inherent product usability.

*Product aesthetics.* The goal of the first pilot study was to determine which of possible alternatives of creating a low-aesthetics version is the best. It was conducted in the form of an online survey using pictures of the appliances. Fig. 1 presents the original coffee machine (i.e. one that we considered to have high aesthetic appeal) together with the four alternative designs, which we considered to have low aesthetic appeal (please note that we deliberately did not create overly unattractive designs to provide realistic and hence ecologically valid experimental materials to participants). All five devices were evaluated by 54 participants using three items ('I like this coffee machine', 'The design of this coffee machine is aesthetically pleasing', 'I would like to have this coffee machine'). We used the mean rating of the three items for the analysis. Analysis of variance showed that the overall aesthetics ratings were significantly different from each other ($F(4, 212) = 41.3$, $p < .001$, $\eta^2 = 0.44$), with the means and standard deviations for each device being presented in Fig. 1. It emerged that the device with the covering 'chaos' was judged to be poorest with regard to its aesthetic appeal and was therefore chosen to be used in the main study, in which it was subjected to a comparative evaluation with the original design (note that the brand name of the coffee machine in the original design was covered up).

*Inherent product usability.* The goal of the second pilot study was to examine the effectiveness of the manipulation of product usability. The inherent usability manipulation was modelled on the principles used in a previous study (Sonderegger et al., 2012), in which the usability of mobile phones was manipulated by changing the display colours and brightness (i.e. reducing readability) and removing text labels of icons (i.e. reducing comprehensibility). In the present lab-based study, readability was reduced by setting the display level of brightness to the

| Original | Chaos | Streetart | Flowers | Labyrinth |
|---|---|---|---|---|
|  |  |  |  |  |
| *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* |
| 4.72 (2.34) | 2.36 (2.93) | 2.81 (3.28) | 2.52 (2.63) | 3.06 (3.72) |

**Fig. 1.** Versions of the coffee machine, with rating scores of aesthetic appeal obtained in the pilot study ($N = 54$). The versions 'original' (high aesthetic appeal) and 'chaos' (low aesthetic appeal) were chosen for the main study.

maximum of 100%, resulting in a machine with poor contrast. Comprehensibility of the controls was reduced by replacing the text labels with numbers (e.g., 'Espresso' displayed next to the espresso icon was replaced by the number '6′). Both measures made it harder for users in the low-usability condition to solve their tasks. Fig. 2 shows a photo of the usability manipulation. Ten participants were asked to complete a subset of the 21 tasks to be carried out in the main study. A one-way analysis of variance confirmed a significant difference of the inherent usability manipulation, $F$ (1, 9) = 37.99, $p < .001$. It showed that participants in the low-usability condition used more interactions ($M = 124$, $SD = 8.7$) than in the high-usability condition ($M = 83$, $SD = 12.5$). Due to technical problems with the online questionnaire measuring perceived usability, the subjective data could not be analysed.

### 2.6. User tasks

Over a period of seven weeks, participants visited the laboratory once a week and were asked to complete three tasks (i.e. they completed 21 different tasks in total). They were given the instructions about each task in writing. The selection of tasks was guided by the idea that a broad range of the most important functions of the coffee machine should be covered. The task sequence was the same for all participants. Examples of the tasks completed include: (a) 'Switch on the machine and follow the instructions on the display', (b) 'Prepare an expresso using the rotating switch, which you will drink later on', (c) 'Select from the menu "maintenance status" the option "rinsing coffee machine", and switch off the machine after the water has stopped running', and (d) 'Check whether the machine needs rinsing'.

### 2.7. Procedure

The experiment took place in two usability laboratories at the University of Fribourg, which were similar in set-up and size. Each laboratory was equipped with a coffee machine. Participants were randomly assigned to one of the four conditions. Prior to visiting the laboratory for the first time, they were asked to complete an online survey, in which they were asked to choose a 30-min time slot between 8 am and noon, during which time they would be available for the following seven weeks. This fixed time slot facilitated the organisation of the experiment and helped reduce the influence of time-of-day effects on the results.


**Fig. 2.** Experimental manipulation of high usability (top) and low usability (bottom).

Although each time slot was of 30-min duration, the experimental session only lasted between 10 and 15 min, allowing for some temporal flexibility if participants arrived late or technical problems were to delay the experiment. In case participants were not able to make it at the fixed time slot, an alternative time slot was offered to them as close as possible to the one originally fixed.

One of the three experimenters welcomed the participant in either French or German language. Thanking them for their participation, we explained that the purpose of this study was to see how users interacted with new technical devices over a period of seven weeks. Each experimenter tested participants in all experimental conditions to reduce the impact of an experimenter effect.

In order to reduce the impact of differences between experimenters, the oral instructions were kept short and supplemented by instructions in writing. The participants were informed that the study was about the use of coffee machines and coffee taste, involving a weekly visit to the lab to operate the machine and to taste a coffee afterwards. If they agreed to take part, they signed a form of informed consent.

An instruction sheet was placed next to the coffee machine, which could be consulted by participants at any time during all experimental sessions. It served as a short manual explaining the main functions of the machine. Participants were instructed that they should carry out the task as if they were to make a coffee at home. The participant was informed that the experimenter would remain in the laboratory, taking some notes and being available for questions should the need arise. If the participant was unable to resolve the task, the experimenter would help, guiding them through the task to find the solution. The idea behind providing guidance to the participant was to ensure an equal exposure to the different functions of the machine. Otherwise, it might affect performance during the completion of subsequent tasks. If the participant had no further questions, she or he was asked to begin with the first task.

After having completed the tasks, participants were asked to rate the quality of the coffee. The following settings of the coffee machine were used throughout the experiment to keep tasting conditions constant: size of coffee (40 ml), coffee intensity (2 out 5), high water temperature, and no milk or sugar was added. The coffee was prepared by each participant within these constraints. Participants were asked to taste the coffee, as they would normally do. They were free to swallow the coffee or spit it out.

This was followed by the completion of several short questionnaires (i.e. SSSQ, SUS, NASA-TLX, perceived attractiveness and choice of complementary coffee). At the end of the first day of testing, participants also filled in two more questionnaire, one collecting demographic data and the other preferences and habits concerning coffee consumption. Finally, participants received a complementary coffee to take away as a small compensation for their taking part in the study. The experimenter prepared this coffee on the same machine in accordance with the participant's preferences, though it was not visible to the participant how exactly the coffee was being made.

### 2.8. Data analysis

The data were analysed by using a three-way analysis of covariance, with the following variables being employed as covariates: age, self-rated coffee expertise and self-rated experience in coffee machine use. The analysis was based on a mixed model, with product aesthetics and inherent usability being entered as between-subjects factors while exposure time was entered as a within-subject factor. If the effect was significant of exposure time, post-hoc analyses using Sidak corrections were carried out. In addition, we performed a correlational analysis on all dependant variables. We carried out z-transformations on the two

measures of performance to reduce the influences caused by differences in task difficulty between testing sessions. If the sphericity assumption for repeated measures testing was violated, the Greenhouse-Geisser correction was applied. However, whenever such a case occurred, we have rounded down the degrees of freedom to the next integer to improve readability of the results section. Assumptions of normality were violated for the following measures: perceived attractiveness, perceived coffee quality, perceived workload, emotional distress, task completion time, and interaction efficiency. The assumption of homogeneity of error variances was violated for perceived attractiveness and for some measures of interaction efficiency. Since parametric tests can be considered robust when the experimental groups are of equal size (which is the case in this study), parametric tests were conducted despite the violation of these assumptions for some measures (Glass et al., 1972). All error bars represent 95% confidence intervals of between-subjects comparisons.

## 3. Results

### 3.1. User ratings of artefact

*Perceived attractiveness.* For the aesthetic appeal of the coffee machine, the analysis of covariance revealed a significant main effect of visual product aesthetics, $F(1, 102) = 9.0, p < .01, \eta^2_{partial} = 0.081$, with aesthetics being rated higher for the high-aesthetics device than for the low-aesthetics device throughout the study (see Fig. 3). This result indicates a successful manipulation check of product aesthetics. Furthermore, the data analysis revealed that the high-usability coffee machine was rated higher in aesthetic appeal than the low-usability device, $F(1, 102) = 5.05, p < .05, \eta^2_{partial} = 0.047$. Exposure time did not have a significant influence on participant ratings of aesthetics of the coffee machine, $F(4, 458) = 1.97, p > .05, \eta^2_{partial} = 0.019$. None of the interactions was significant. Two of the three covariates had a significant influence on perceived attractiveness ($F_{ExpMachine}$ (1, 102) = 9.14, $p < .01, \eta^2_p = 0.082$; $F_{ExpCoffee}$ (1, 102) = 9.14, $p < .01, \eta^2_p = 0.076$). Experience in coffee machine use was positively correlated with perceived attractiveness (ranging from $r_{day2} = 0.03$ to $r_{day6} = 0.16$, with all correlations not reaching significance level), and self-rated coffee experience showed a negative correlation (ranging from $r_{day2} = -0.1$ to $r_{day3} = -0.21$, with two correlations reaching significance level).

*Perceived usability.* As expected, the analysis showed that the high-usability coffee machine was rated significantly higher in usability by participants than the low-usability device (see Fig. 4), $F(1, 102) = 7.27, p < .01, \eta^2_p = 0.067$. This finding confirms a successful manipulation check for product usability. There was no significant effect of product aesthetics, $F < 1$. Usability ratings were found to change during exposure time, $F(5, 542) = 5.6, p < .001, \eta^2_p = 0.052$. Post-hoc analyses revealed that all data points were statistically significant from each other ($p < 0.05$), except for weeks 1,2 and 6 (i.e. no difference between any of the three), and weeks 5 and 7 (i.e. no difference between them). No significant interactions were observed. Finally, we found one of the three covariates (i.e. experience in coffee machine use) to have a significant link with perceived usability, $F_{ExpMachine}(1, 102) = 8.08, p < .01, \eta^2_p = 0.073$, indicating positive correlations ranging from $r_{day5} = 0.1$ to $r_{day1} = 0.31$ (with three correlations being significant).

*Perceived coffee quality.* The ratings of coffee quality as a measure of gustatory attractiveness were analysed over the course of the experiment. The rating of overall coffee quality showed no significant main effect for any of the IVs and no interaction (all $F < 1.71$). The data for the overall rating are presented in Fig. 5. Two of the three covariates (i.e. age, coffee experience) showed a significant association with coffee quality. The older the participants were, the higher the quality rating was, $F(1, 102) = 5.10, p < .05, \eta^2_p = 0.048$ (correlations ranging from $r_{day5} = 0.12$ to $r_{day7} = 0.22$, with five coefficients being significant). Furthermore, the more (self-reported) expertise participants had, the higher they rated the quality of the coffee, $F(1, 102) = 4.55, p < .05, \eta^2_p = 0.043$, with correlations ranging from $r_{day6} = 0.13$ to $r_{day2} = 0.24$ (with four coefficients being significant).

### 3.2. Self-reported user state

*Perceived workload.* The data for perceived workload are presented in Fig. 6. The ratings of the NASA-TLX did not differ as a function of product aesthetics, $F(1, 102) = 1.22, p > .05, \eta^2_p = 0.012$, and inherent usability, $F < 1$. The analysis also revealed that perceived workload varied over time, $F(5, 566) = 3.15, p < .01, \eta^2_p = 0.030$. Post-hoc analyses showed that workload was perceived to be higher in the third week than in all other weeks (post-hoc test: $p < .001$), and lowest in week 6

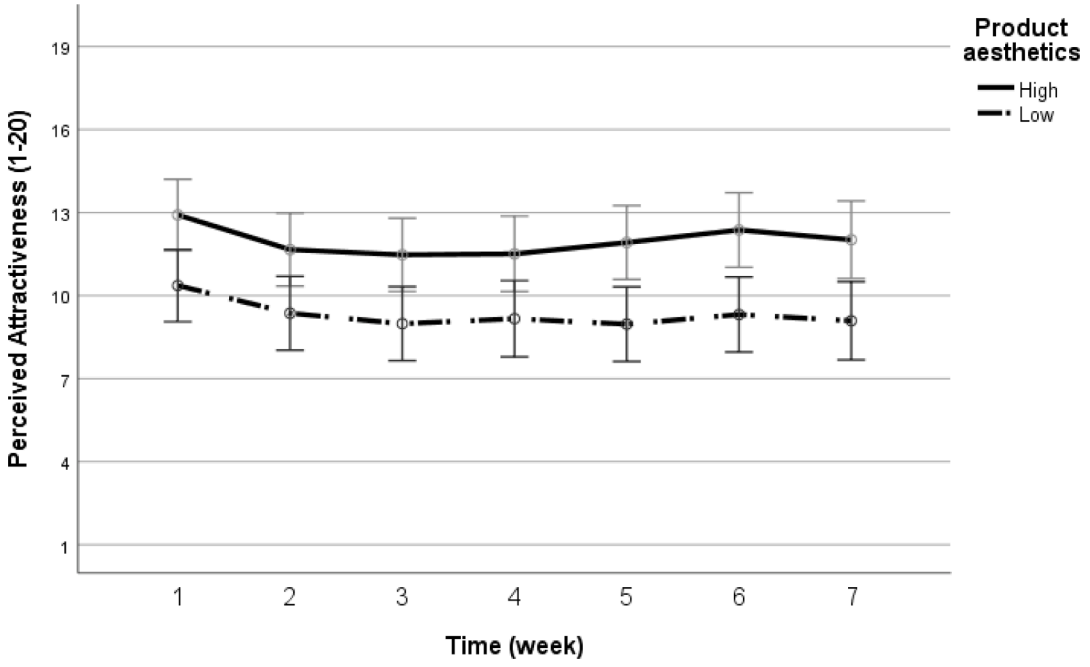


**Fig. 3.** Perceived attractiveness as a function of exposure time and product aesthetics (bars indicate 95% confidence intervals of between-subjects comparisons).

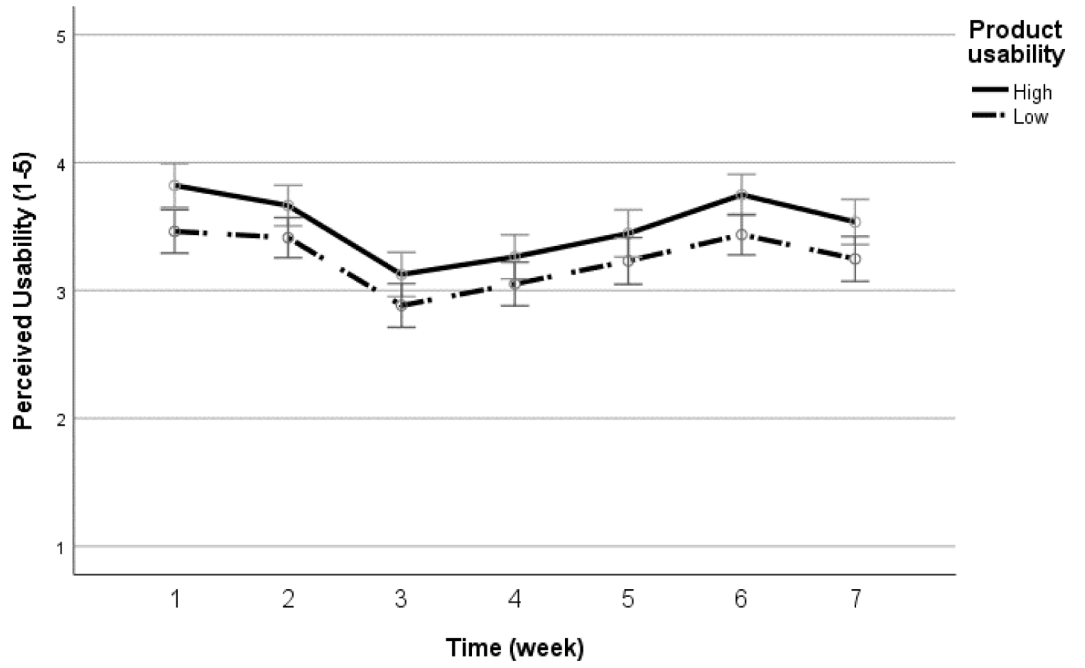


**Fig. 4.** Perceived usability as a function of exposure time and inherent usability (bars indicate 95% confidence intervals of between-subjects comparisons).

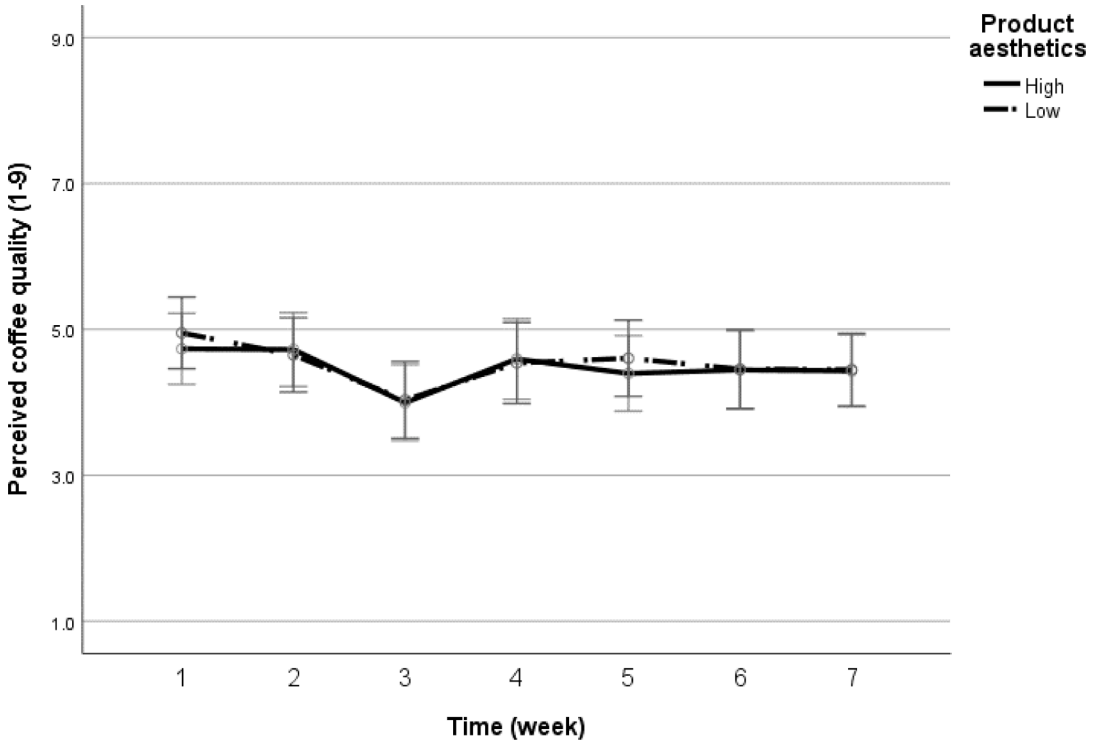


**Fig. 5.** Perceived coffee quality as a function of exposure time and product aesthetics (bars indicate 95% confidence intervals of between-subjects comparisons).

compared to all other weeks (post-hoc test: $p < .001$). No significant interaction was found. The analysis did not reveal any significant effect for any of the three covariates, *all* $F < 1$.

*Affective state (emotional distress).* Ratings of affect revealed no significant effect of aesthetics (see Fig. 7), $F(1, 102) = 0.85$, $p > .05$, $\eta^2_p = 0.008$. The main effects of product usability, $F(1, 102) = 1.19$, $p > .05$, $\eta^2_p = 0.011$, and exposure time, $F(1, 102) = 1.32$, $p > .05$, $\eta^2_p = 0.013$, were not significant. None of the interactions was found to be significant. Finally, the covariates showed no significant influence on emotional distress, *all* $F < 1$.

### 3.3. User performance

*Task completion time.* The data for this measure underwent a z-transformation (see Section 2.9). To facilitate the interpretation of the data, we reported the untransformed data in Fig. 8. The analysis revealed that low inherent usability resulted in significantly longer task completion times than high usability, $F(1, 102) = 7.48$, $p < .01$, $\eta^2_p =$

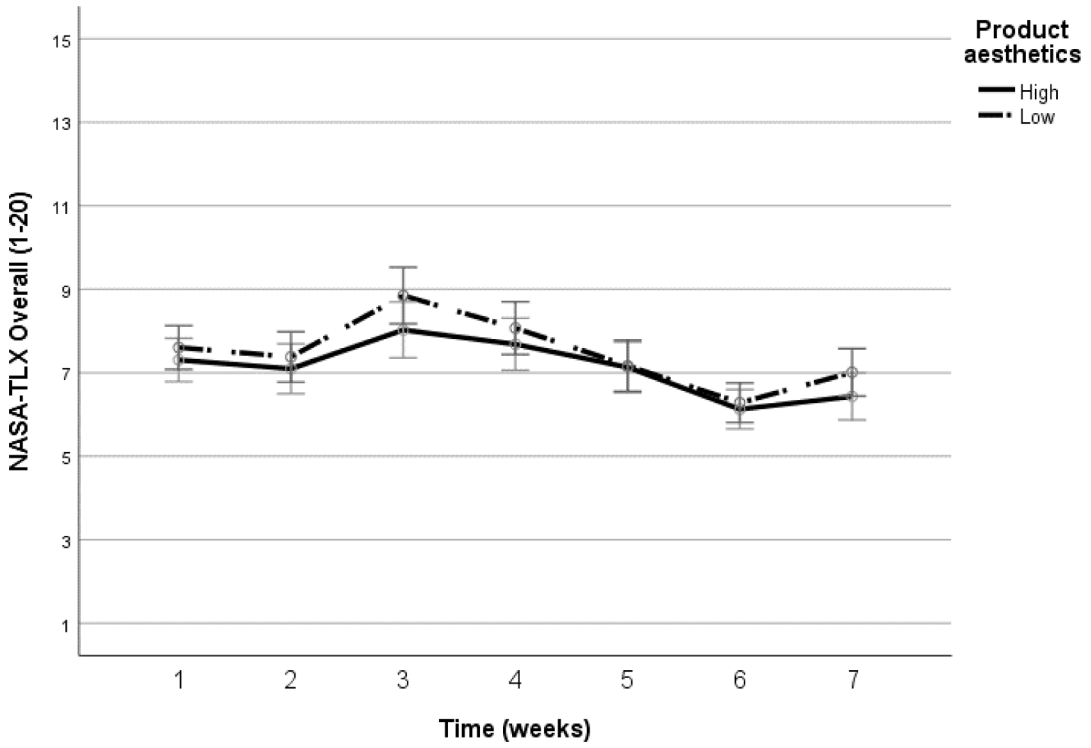


**Fig. 6.** Perceived workload as a function of exposure time and product aesthetics (bars indicate 95% confidence intervals of between-subjects comparisons).

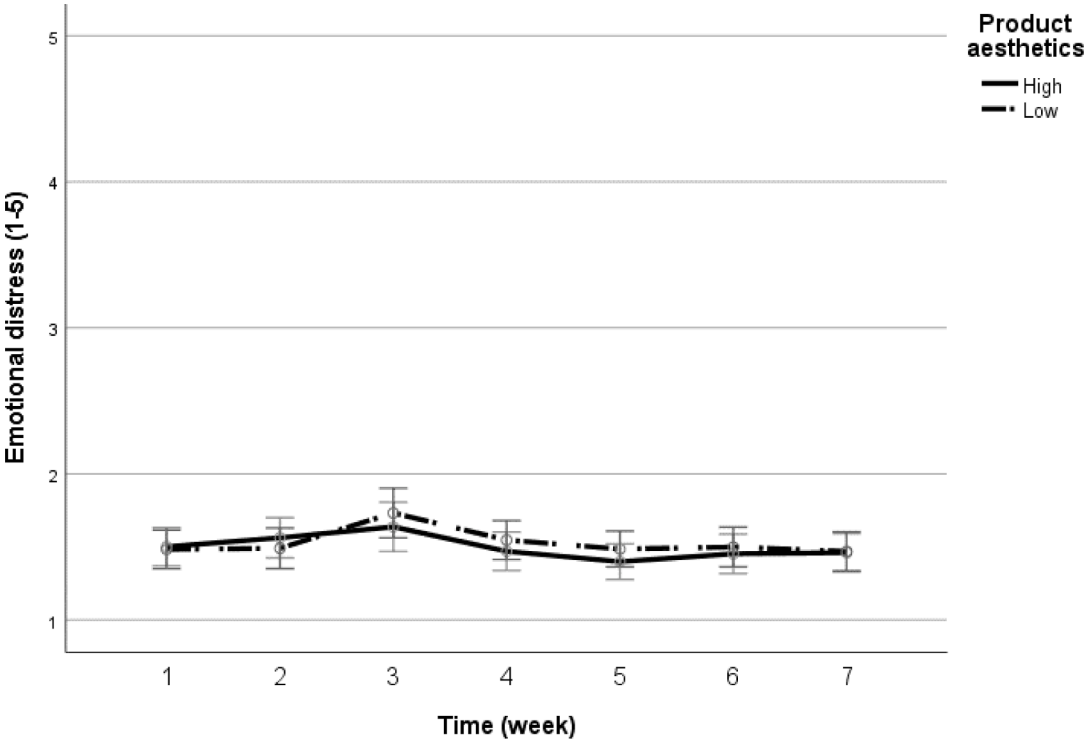


**Fig. 7.** Emotional distress as a function of exposure time and product aesthetics (bars indicate 95% confidence intervals of between-subjects comparisons).

0.068. Furthermore, we observed an interaction between inherent usability and exposure time in that the negative influence of low usability on completion time (especially pronounced in weeks 1 and 2) decreased with increasing experience of the user with the device, $F(6, 612) = 4.46$, $p < .001$, $\eta^2_p = 0.042$. The data show that for some tasks, users in the low-usability were able to compensate the poor usability of the device, showing no difference to the high-usability condition (e.g., weeks 3, 5, and 7). Conversely, the difference between the two conditions was particularly pronounced in week 1 when users were least familiar with the device. There were no effects of product aesthetics and exposure time, both $F < 1$. With regard to the covariates, the analysis revealed that age was negatively related to task completion time, $F(1, 102) =$

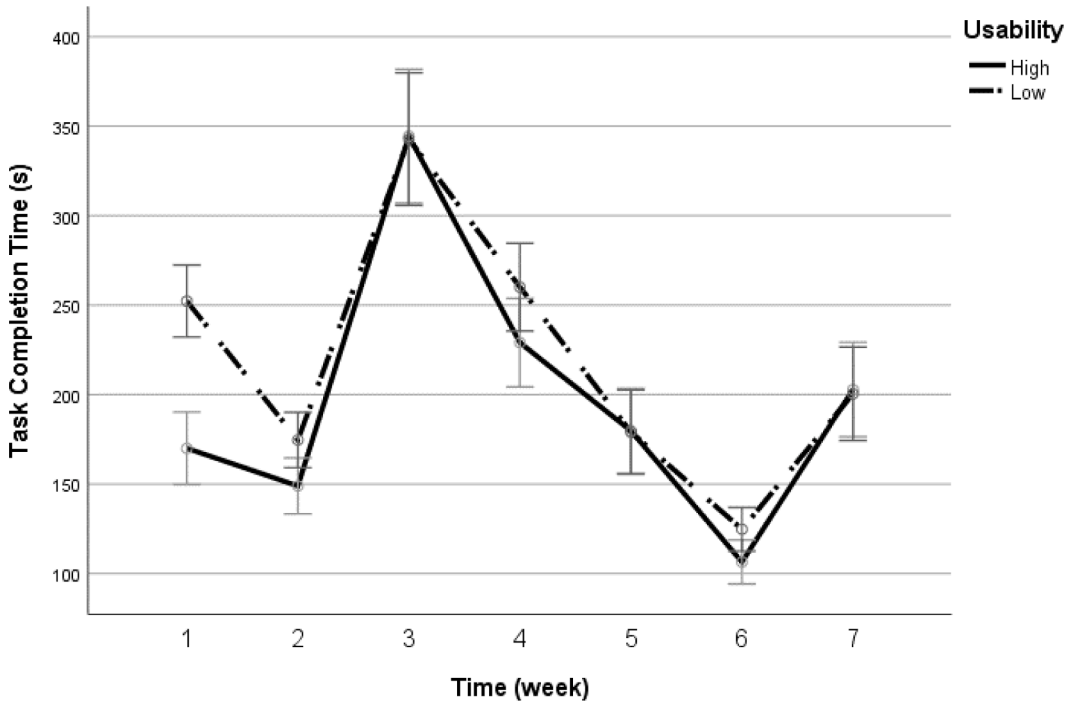


**Fig. 8.** Task completion time as a function of exposure time and inherent usability (bars indicate 95% confidence intervals of between-subjects comparisons).

19.26, $p < .001$, $\eta^2_p = 0.159$ (correlations ranging from $r_{day5} = 0.16$ to $r_{day4} = 0.35$, with five coefficients being significant), while previous technical expertise showed a positive relationship with this measure, $F(1, 102) = 3.76$, $p = .055$, $\eta^2_p = 0.036$ (correlations ranging from $r_{day5} = 0.12$ to $r_{day3} = -0.20$, with one coefficient being significant).

*Interaction efficiency*. The data for interaction efficiency are presented in Fig. 9, indicating the number of unnecessary user inputs that represented deviations from the optimal user-device dialogue. The analysis showed better performance when usability was high than when it was low, $F(1, 102) = 5.95$, $p < .05$, $\eta^2_p = 0.055$. Since this effect was task-dependant, it was observed only in some testing sessions (notably in weeks 1, 2, 4 and 6). This was confirmed by a significant interaction between inherent usability and exposure time, $F(6, 612) = 5.38$, $p < .001$, $\eta^2_p = 0.050$. There were no main effects of product aesthetics and exposure time, both $F < 1$. Furthermore, the analysis revealed a significant three-way interaction between all three independent variables, $F(6, 612) = 2.33$, $p < .05$, $\eta^2_p = 0.022$. We do not report the descriptive data for this effect since it is difficult to interpret. Finally, as for the preceding performance measure, we recorded that age was negatively related to interaction efficiency, $F(1, 102) = 15.57$, $p < .001$, $\eta^2_p =$

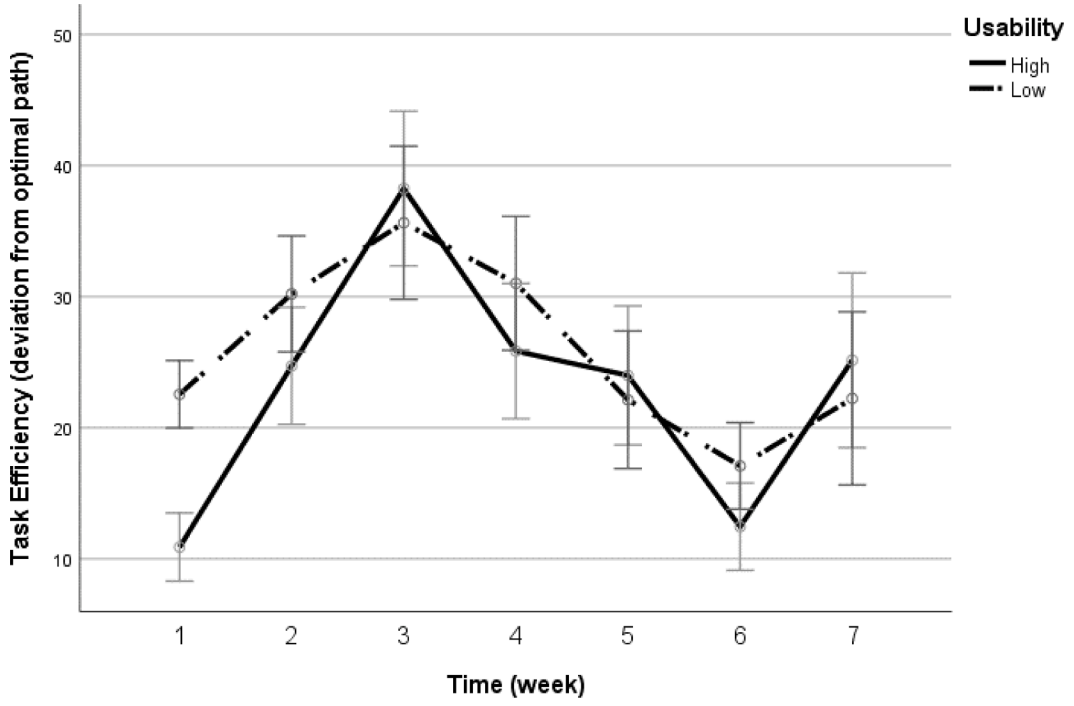


**Fig. 9.** Interaction efficiency as a function of exposure time and inherent usability (bars indicate 95% confidence intervals of between-subjects comparisons).

0.132, while previous technical expertise showed a positive relationship, $F(1, 102) = 6.25$, $p < .05$, $\eta^2_p = 0.058$.

### 3.4. Correlational analysis of outcome measures

Correlation coefficients were computed for all dependant variables, separately for each of the seven points of measurement. Table 1 summarises the range of correlation coefficients computed for each pair of dependant variables across the seven points of measurement. The asterisks indicate which of the seven correlation coefficients were significant at the 5% level. The results showed consistently high correlations between subjective evaluations of usability and visual attractiveness while the link of usability ratings with gustatory attractiveness was less pronounced. Visual and gustatory attractiveness (i.e. coffee quality) both showed generally very little correlations with other measures of the construct of user experience apart from their relation to perceived usability. Emotional distress was related to perceived usability and perceived workload, and to a lesser extent, to perceived attractiveness and task completion time. Finally, the two performance measures correlated with each other.

## 4. Discussion

The present study was one of the very few that examined the impact of visual aesthetics on user experience in a longitudinal experiment (i.e. involving multiple testing sessions rather than a single testing session). The results showed no effects of visual aesthetics on perceived usability as the chief outcome measure. Contrary to expectations, we did not find a reduction of the influence of aesthetics over the course of exposure time, with nil effects of aesthetics being observed already from the beginning of user-device interaction.

A major finding of the present study was that the aesthetically appealing product did not show a positive effect on perceived usability. In contrast to earlier longitudinal studies (Kujala and Miron-Shatz, 2013; Sonderegger et al., 2012), there was no 'what--is-beautiful-is-good'-effect (Tractinsky et al., 2000), which then waned with increasing exposure time. The 'what-is-beautiful-is-good'-effect even failed to emerge in the first place. Put differently, there was no significant interaction between visual aesthetics and exposure time, which has emerged in the two other longitudinal studies examining the relationship between visual aesthetics and perceived usability. While the absence of an effect of visual aesthetics on perceived usability in itself is not surprising after the user has gained some experience with the device, it was unexpected that an effect of visual aesthetics on perceived usability was not recorded at the beginning of the study.

Prior to exploring possible reasons for the unexpected finding, it is important to note that our statistical test confirmed a successful experimental manipulation of visual aesthetics. This manipulation also proved to be effective in the preceding pilot studies. Therefore, we can exclude an ineffective aesthetics manipulation to be at the root of the 'what-is-beautiful-is-good'-effect being absent. There are four possible explanations for the absence of this effect: (a) sensory dominance, (b)

**Table 1**
Range of correlation coefficients for each pair of dependant variables across seven points of measurement.

| Measures | Perceived attractiveness | Emotional distress | Perceived workload | Task completion time | Interaction efficiency | Perceived coffee quality |
|---|---|---|---|---|---|---|
| Perceived usability | | | | | | |
| $r_{min}$ | .319 | −0.184 | −0.217 | −0.140 | −0.088 | .122 |
| $r_{max}$ | .499 | −0.380 | −0.345 | −0.640 | −0.375 | .259 |
| $r_{mean}$ | .437 | −0.267 | −0.293 | −0.321 | −0.208 | .192 |
| SD | 0.06 | 0.08 | 0.05 | 0.16 | 0.12 | 0.05 |
| Week-by-week correlations | ******* | *°*****° | ******* | ******* | *°**°°° | *°°*°** |
| Perceived attractiveness | | | | | | |
| $r_{min}$ | | −0.051 | −0.144 | −0.320 | −0.405 | .041 |
| $r_{max}$ | | −0.237 | .028 | .013 | .149 | .159 |
| $r_{mean}$ | | −0.151 | −0.037 | −0.108 | −0.071 | .096 |
| SD | | 0.07 | 0.07 | 0.11 | 0.18 | 0.04 |
| Week-by-week correlations | | °°*°**° | °°°°°°° | *°°°°°° | *°°°°°° | °°°°°°° |
| Emotional distress | | | | | | |
| $r_{min}$ | | | .207 | .013 | −0.136 | −0.279 |
| $r_{max}$ | | | .392 | .271 | .181 | .014 |
| $r_{mean}$ | | | .302 | .138 | .076 | −0.121 |
| SD | | | 0.06 | 0.10 | 0.12 | 0.10 |
| Week-by-week correlations | | | ******* | *°*°°*° | °°°°°°° | °°°**°° |
| Perceived workload | | | | | | |
| $r_{min}$ | | | | .157 | .116 | −0.163 |
| $r_{max}$ | | | | .284 | .356 | .047 |
| $r_{mean}$ | | | | .223 | .218 | −0.056 |
| SD | | | | 0.05 | 0.09 | 0.07 |
| Week-by-week correlations | | | | °**°*** | °°**°** | °°°°°°° |
| Performance (task completion time) | | | | | | |
| $r_{min}$ | | | | | .527 | −0.130 |
| $r_{max}$ | | | | | .840 | .038 |
| $r_{mean}$ | | | | | .723 | −0.054 |
| SD | | | | | 0.11 | 0.06 |
| Week-by-week correlations | | | | | ******* | °°°°°°° |
| Performance (user interactions) | | | | | | |
| $r_{min}$ | | | | | | −0.106 |
| $r_{max}$ | | | | | | .058 |
| $r_{mean}$ | | | | | | −0.033 |
| SD | | | | | | 0.06 |
| Week-by-week correlations | | | | | | °°°°°°° |

Note: Week-by-week correlations: in each week, an asterisk indicates a significant correlation coefficients (i.e. $p < 05$) whereas a nought indicates a non-significant correlation coefficients during that week (e.g., °***°*** indicates that correlations were non-significant in weeks one and four); $r_{mean}$ is based on correlations that underwent a z-transformation being weighted by the number of cases.

product identification processes, (c) product complexity, and (d) duration of interaction. All four explanations refer to differences in the set-up between the present work and earlier studies.

First, the present study differs from previous work with regard to the sensory experience during user-product interaction. Coffee machines differ from telephones with regard to the different relative weight of sensory experiences (Schifferstein, 2006), with the difference being expected to be even larger when smartphones instead of traditional telephones are being used (leading to a stronger dominance of the visual modality). Work by Fenko et al. (2010) suggests that sensory dominance is not static but is subject to change as the user gains increasing experience with the device. This work shows that users attach a very high importance to the visual modality at the time of purchasing a coffee maker (whereas the other senses only play a subsidiary role). However, four weeks later the same users considered this modality to be of rather low importance while they then reported that audition, olfaction and taste had been of higher importance than the visual modality. While Fenko et al. (2010) did not provide specific data for mobile phones, the pattern for high-tech products as a more generic term was different from coffee makers in that the visual modality did not lose its importance to the same extent (even though its importance decreased to a small extent as a function of increasing usage time). The work of Fenko suggests that non-visual modalities become increasingly important for coffee machines. This may explain why a similar results pattern was found in the present study as in previous work of the authors, which showed that non-visual aesthetics had little effect on perceived usability (Sonderegger and Sauer, 2015).

Second, the degree to which users identified with a product may have been lower in the present study than in previous work. This is because a coffee machine was used rather than a mobile phone. The ideas behind the concept of product identification have been expressed by Karapanos et al. (2009) in a model distinguishing three distinct phases of product use: orientation, incorporation and identification. When beginning to use a product, the user's first experience with the device is characterised by feelings of excitement as well as frustration (orientation phase), followed by reflections on how the device can be sensibly used in the user's daily lives (incorporation phase). In the third phase (identification), personal and social aspects of the user's experience with the product gain in importance. For example, the user may feel good about themselves when using the device, resulting in increased self-esteem but also social esteem (e.g., because he or she is in possession of a more attractive device than his or her friends are). We would argue that for a large part of the population owning an appealing mobile phone is more important for social status than owning an appealing coffee machine, which may explain the different pattern of findings in the present study. Many previous studies demonstrating the effect of visual aesthetics on perceived usability used portable high-status products (i.e. they can be easily presented to other people) such as mobile phones (e.g., Hamborg et al., 2014) or MP3-players (Mahlke and Thüring, 2007), which enjoyed a high status at the time, too. Finally, in a longitudinal experiment the identification of the user with the product may also be higher when it is carried out in the field rather than in the lab. This is because in a field experiment, participants usually take home the device that they are testing (e.g., Sonderegger et al., 2012), which is expected to increase identification since the device has temporarily become part of their domestic environment and can even be used outside the scheduled testing sessions. Conversely, in the present lab-based experiment users only had access to the device once a week during the scheduled testing session.

Third, the devices may differ in the ease with which product usability can be assessed by users. The mechanism of visual aesthetics influencing perceived usability works best if no good cues allow the user to assess the usability of a device (e.g., mobile phones). Therefore, this mechanism may work less well for coffee machines because it is easier to assess the usability of this device than of a mobile phone. This is because the number and complexity of functions is generally smaller for coffee machines than for mobile phones. Based on the data of the present study, it is difficult to say which of the three explanations put forward is most valid. To answer this question, further research would need to test the validity of each of the three explanations.

Fourth, the studies in the research literature differ with regard to the duration of the interaction users had with the device. In the present study, the interaction lasted several minutes while in some of the previous studies there were no or very short interactions (e.g., Iten et al., 2018; Minge and Thüring, 2018). Considering that expected usability ratings were found to be highly correlated with aesthetics ratings (Thielsch et al., 2015), one may speculate that with increasing duration of the interaction the influence of aesthetics loses its impact. However, it is yet uncertain after what time period such a loss of impact may occur.

As a second independent variable, product usability was manipulated. This experimental manipulation was effective because product usability showed a significant effect on perceived usability. This effect was not limited to the subjective component of usability but was also observed in the two performance measures (i.e. task completion time and interaction efficiency). However, more important is the question of whether the 'what-is-good-is-beautiful'-effect (Tuch et al., 2012) would also emerge as a result of such an experimental manipulation. In line with previous work that found such an effect (e.g., Hamborg et al., 2014; Minge and Thüring, 2018; Tuch et al., 2012), it is not clear to what extent the 'what-is-good-is-beautiful'-effect has occurred in the present study. While the data seem to suggest the presence of this effect, this needs to be interpreted with some caution because it cannot be excluded that the manipulation of inherent usability was to some extent confounded with aesthetics. For the manipulation of inherent usability we changed the display contrast (being low in the low-usability condition), which may have also been perceived as aesthetically less pleasing by participants. Since we do not have any data to verify the size of this side effect, we need to exercise some caution with regard to confirming the presence of the 'what-is-good-is-beautiful'-effect.

The correlational analysis revealed several clusters of variables showing associations. Emotional distress was linked to perceived usability and perceived workload, and partly associated with perceived attractiveness and task completion time as an indicator of objective performance. When perceived usability decreased and perceived workload increased, the ratings of emotional distress were found to increase (though the overall rating of this variable was at the lower end of the scale). Perceived usability, perceived workload and objective performance all belong to the cluster of instrumental qualities according to the Components of User Experience Model (CUE model; Thüring and Mahlke, 2007). This model makes important distinction between instrumental product qualities (e.g., effectiveness) and non-instrumental qualities (e.g., visual aesthetics, haptic qualities). The present finding suggests that emotional qualities may be related to this cluster of instrumental qualities. The correlational analysis also revealed that visual attractiveness and perceived coffee quality (as a measure of gustatory attractiveness) showed no correlation. The results showed that the relationship between perceived usability and visual attractiveness was stronger than the link of usability ratings to gustatory attractiveness (i.e. coffee quality). These observations suggest that participants made a clear distinction between visual and gustatory aspects of user-product interaction, reiterating the differences found in previous research between visual and non-visual aesthetics (e.g., Schifferstein, 2006).

The analyses also revealed main effects of exposure time on a number of outcome variables such as perceived usability and perceived workload. Furthermore, there were interactions between exposure time and the two other independent factors, which also pointed to the influence of exposure time. However, the patterns observed were not of a systematic nature (e.g., a linear trend in the form of increasing or decreasing scores, or a non-linear trend in the form of a U-shaped function). The decrease in both performance variables in week 3 was particularly striking. A post-hoc analysis of the difficulty of the tasks assigned revealed that the tasks given in week 3 were particularly challenging. The consequences

of the fluctuation in task difficulty are a result of a fixed task order being used in the present experiment (see limitations below). Therefore, we consider these non-systematic patterns to be difficult to interpret as they may have influenced by differences in task difficulty across testing sessions.

Using a fixed task order represents the first limitation of the study. In a large experiment of this kind using a large set of task, it is difficult to control for the influence of task difficulty. We have presented the tasks to all participants in the same order due to the considerable logistic challenges of a multiple-session experiment. The alternative approach of balancing out the order of tasks was not adopted because such an experimental protocol would have increased the risk of errors being made in the sequencing of tasks. We acknowledge that this decision may have influenced the findings. A possible confounding effect of usability and aesthetics represent the second limitation of the study. It is notoriously difficult to separate these two factors, even in well-designed experiments. In the present experiment, the manipulation of usability (i.e. in the form of changing the contrast of the display) might have had an impact on product aesthetics at the same time (i.e. involving a moderate manipulation of aesthetics). This suggests that the reduced display contrast may have also changed the level of product aesthetics. For future research, there is a need to conceive a manipulation of product usability that does not influence product aesthetics at the same time. A third limitation concerns the only moderate effect size found for the aesthetics manipulation (which is in contrast to the very large effect size in the pilot study). Since there are substantial differences between individuals with regard to aesthetics evaluations (i.e. 'it is a matter of taste!'), the use of aesthetics manipulations in the form of a between-subjects variable (as in main experiment) rather than a within-subjects variable (as in pilot study) has reduced statistical power and hence increased the probability of a type II error. A fourth limitation refers to three scales being used in the present study. The NASA-TLX assessing workload was found to have a lower internal consistency than in its validation studies. The distress dimension of the SSSQ scale showed very low scores, paralleled by low levels of variance. The two items employed to assess perceived attractiveness may have been insufficient to capture fully the attractiveness of a coffee machine since it failed to assess other important facets (e.g., sounds, haptic features). All four limitations may have influenced the findings in the form of reducing the probability of finding significant effects.

In addition to these limitations outlined, which have some implications for future research, there are further issues that should be considered in future work. First, it may be useful to include indicators of positive affective states (e.g., pleasure, sense of accomplishment, pride) in addition to the indicators of negative affect, upon which the current experiment focused (cf. Reppa et al., 2021). Second, future work needs to determine whether the findings of the present experiment may extend to other technical artefacts.

In conclusion, the present study makes an important contribution to the research efforts in the field of product aesthetics for several reasons. First, it represents one of the very few lab-based studies that were carried out over an extended testing period. Second, it employed a substantial sample size (even for a lab-based experiment with a one-off testing session, the sample size would be considered substantial). Third, a different device was used compared to most previous studies (coffee machine rather than mobile phone), which allows a first examination of the generalisability of the findings. Overall, the findings of the present study provide further evidence that it may have been premature to assume a stable and consistent effect of product aesthetics onto user experience and, in particular, its subsidiary component 'perceived usability'. The present work is expected to contribute to the ongoing debate about the relationship between these variables.

## CRediT authorship contribution statement

**Juergen Sauer:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition. **Andreas Sonderegger:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We are very grateful to the Swiss National Science Foundation for their financial support (grant no. 100014/116012). We would also like to express our gratitude to Heidi Fischer, Gianfranco Bastianelli and Sebastian Zumbühl for their support. We would also like to thank the company 'Jura Elektroapparate AG' for their help in completing this study.

## References

Bangor, A., Kortum, P.T., Miller, J.T., 2008. An empirical evaluation of the system usability scale. Int. J. Hum. Comput. Interact. 24 (6), 574–594.
Ben-Bassat, T., Meyer, J., Tractinsky, N., 2006. Economic and subjective measures of the perceived value of aesthetics and usability. ACM Trans. Comput. Hum. Interact. (TOCHI) 13 (2), 210–234.
Bölte, J., Hösker, T.M., Hirschfeld, G., Thielsch, M.T., 2017. Electrophysiological correlates of aesthetic processing of webpages: a comparison of experts and laypersons. PeerJ 5, e3440.
Brady, L., Phillips, C., 2003. Aesthetics and usability: a look at color and balance. Usability News 5 (1), 2–5.
Brooke, J., 1996. SUS: a quick and dirty usability scale. Usability Eval. Ind. 189 (194), 4–7.
Brooke, J., 2013. SUS: a retrospective. J. Usability Stud. 8 (2), 29–40.
Eisinga, R., Te Grotenhuis, M., Pelzer, B., 2013. The reliability of a two-item scale: pearson, cronbach, or spearman-brown? Int. J. Public Health 58 (4), 637–642.
Fenko, A., Schifferstein, H.N., Hekkert, P., 2010. Shifts in sensory dominance between various stages of user–product interactions. Appl. Ergon. 41 (1), 34–40.
Fu, L., Salvendy, G., Turley, L., 2002. Effectiveness of user testing and heuristic evaluation as a function of performance classification. Behav. Inf. Technol. 21 (2), 137–143.
Glass, G.V., Peckham, P.D., Sanders, J.R., 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev Educ Res 42 (3), 237–288.
Hamborg, K.C., Hülsmann, J., Kaspar, K., 2014. The interplay between usability and aesthetics: more evidence for the 'what is usable is beautiful' notion. Adv. Hum. Comput. Interact. 2014, 15.
Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. Elsevier, Amsterdam, pp. 139–183.
Hartmann, J., Sutcliffe, A., De Angeli, A., 2007. Investigating attractiveness in web user interfaces. In: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, pp. 387–396.
Hekkert, P., Leder, H, 2008. Product aesthetics. In: Schifferstein, H.N.J., Hekkert, P. (Eds.), Product Experience. Elsevier Science, Amsterdam, pp. 259–286.
Helton, W.S., 2004. Validation of a short stress state questionnaire. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 48 (11), 1238–1242.
Helton, W.S., Naswall, K., 2015. Short stress state questionnaire. Eur. J. Psychol. Assess. 31, 20–30.
Iten, G.H., Troendle, A., Opwis, K., 2018. Aesthetics in context - the role of aesthetics and usage mode for a website's success. Interact. Comput. 30 (2), 133–149.
Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.B., 2009. User experience over time: an initial framework. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 729–738.
Kujala, S., Miron-Shatz, T., 2013. Emotions, experiences and usability in real-life mobile phone use. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 1061–1070.
Kurosu, M., Kashimura, K., 1995. Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability. In: Proceedings of the Conference Companion on Human factors in Computing Systems. ACM, pp. 292–293.
Lee, S., Ha, T., 2019. Changes in perceived usability and aesthetics with repetitive use in the first use session. Hum. Factors Ergon. Manuf. Serv. Ind. 29 (6), 517–528.
Lindgaard, G., Fernandes, G., Dudek, C., Brown, J., 2006. Attention web designers: you have 50 milliseconds to make a good first impression! Behav. Inf. Technol. 25 (2), 115–126.

Mahlke, S., Thüring, M., 2007. Studying antecedents of emotional experiences in interactive contexts. In: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, pp. 915–918.
Minge, M., Thüring, M., 2018. Hedonic and pragmatic halo effects at early stages of user experience. Int. J. Hum. Comput. Stud. 109, 13–25.
Moshagen, M., Musch, J., Göritz, A.S., 2009. A blessing, not a curse: experimental evidence for beneficial effects of visual aesthetics on performance. Ergonomics 52 (10), 1311–1320.
Moshagen, M., Thielsch, M.T., 2010. Facets of visual aesthetics. Int. J. Hum. Comput. Stud. 68 (10), 689–709.
Nakarada-Kordic, I., Lobb, B., 2005. Effect of perceived attractiveness of web interface design on visual search of web sites. In: Proceedings of the 6th ACM SIGCHI New Zealand Chapter's International Conference on Computer-Human Interaction: Making CHI Natural. ACM, pp. 25–27.
Peres, S.C., Pham, T., Phillips, R., 2013. Validation of the system usability scale (SUS) SUS in the wild. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 57 (1), 192–196.
Postrel, V., 2003. The Substance of Style: How the Rise of Aesthetic Value is Remarking Commerce, Culture, and Consciousness. Harper Collins, New York.
Reber, R., Schwarz, N., Winkielman, P., 2004. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? Pers. Soc. Psychol. Rev. 8 (4), 364–382.
Reppa, I., McDougall, S., Sonderegger, A., Schmidt, W.C., 2021. Mood moderates the effect of aesthetic appeal on performance. Cognit. Emot. 35 (1), 15–29.
Sauer, J., Sonderegger, A., 2009. The influence of prototype fidelity and aesthetics of design in usability tests: effects on user behaviour, subjective evaluation and emotion. Appl. Ergon. 40 (4), 670–677.
Sauer, J., Sonderegger, A., Schmutz, S., 2020. Usability, user experience and accessibility: towards an integrative model. Ergonomics 63 (10), 1207–1220.
Schenkman, B.N., Jönsson, F.U., 2000. Aesthetics and preferences of web pages. Behav. Inf. Technol. 19 (5), 367–377.
Schifferstein, H.N., 2006. The perceived importance of sensory modalities in product usage: a study of self-reports. Acta Psychol. (Amst.) 121 (1), 41–64.
Sonderegger, A., Sauer, J., 2010. The influence of design aesthetics in usability testing: effects on user performance and perceived usability. Appl. Ergon. 41 (3), 403–410.
Sonderegger, A., Sauer, J., 2015. The role of non-visual aesthetics in consumer product evaluation. Int. J. Hum. Comput. Stud. 84, 19–32.
Sonderegger, A., Zbinden, G., Uebelbacher, A., Sauer, J., 2012. The influence of product aesthetics and usability over the course of time: a longitudinal field experiment. Ergonomics 55 (7), 713–730.
Sutcliffe, A., Namoune, A., 2008. Getting the message across: visual attention, aesthetic design and what users remember. In: Proceedings of the 7th ACM Conference on Designing Interactive Systems. ACM, pp. 11–20.
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.
Thielsch, M.T., Engel, R., Hirschfeld, G., 2015. Expected usability is not a valid indicator of experienced usability. PeerJ Comput. Sci. 1, e19.
Thielsch, M.T., Scharfen, J., Masoudi, E., Reuter, M., 2019. Visual aesthetics and performance: a first meta-analysis. In: Proceedings of Mensch und Computer 2019, pp. 199–210.
Thüring, M., Mahlke, S., 2007. Usability, aesthetics and emotions in human–technology interaction. Int. J. Psychol. 42 (4), 253–264.
Tractinsky, N., Katz, A.S., Ikar, D., 2000. What is beautiful is usable. Interact. Comput. 13 (2), 127–145.
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. Comput. Hum. Behav. 28 (5), 1596–1607.
Van der Heijden, H., 2003. Factors influencing the usage of websites: the case of a generic portal in The Netherlands. Inf. Manag. 40 (6), 541–549.
Van Schaik, P., Ling, J., 2009. The role of context in perceptions of the aesthetics of web pages over time. Int. J. Hum. Comput. Stud. 67 (1), 79–89.
Wright, P., McCarthy, J., Meekison, L., 2003. Making sense of experience. Funology. Springer, Dordrecht, pp. 43–53.
Xiao, Y.M., Wang, Z.M., Wang, M.Z., Lan, Y.J., 2005. The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index. Chin. J. Ind. Hyg. Occup. Dis. 23 (3), 178–181.