# LEMtool - Measuring Emotions in Visual Interfaces

**Gijs Huisman**
University of Twente, Human
Media Interaction Group
gijs.huisman@utwente.nl

**Marco van Hout**
SusaGroup
marco.vanhout@susagroup.com

**Betsy van Dijk**
University of Twente, Human
Media Interaction Group
e.m.a.g.vandijk@utwente.nl

**Thea van der Geest**
University of Twente, Media,
Communication &
Organisation
t.m.vandergeest@utwente.nl

**Dirk Heylen**
University of Twente, Human
Media Interaction Group
d.k.j.heylen@utwente.nl

## ABSTRACT

In this paper the development process and validation of the LEMtool (Layered Emotion Measurement tool) are described. The LEMtool consists of eight images that display a cartoon figure expressing four positive and four negative emotions using facial expressions and body postures. The instrument can be used during interaction with a visual interface, such as a website, and allows participants to select elements of the interface that elicit a certain emotion. The images of the cartoon figure were submitted to a validation study, in which participants rated the recognizability of the images as specific emotions. All images were found to be recognizable above chance level. In another study, the LEMtool was used to assess visual appeal judgements of a number of web pages. The LEMtool ratings were supported by visual appeal ratings of web pages both for very brief (50 milliseconds) and for long (free-viewing) stimulus exposures. Furthermore, the instrument provided insight into the elements of the web pages that elicited the emotional responses.

### Author Keywords

LEMtool; emotion; user experience; visual appeal; web pages.

### ACM Classification Keywords

H.5.2. User Interfaces: Evaluation/methodology.

## INTRODUCTION

Emotions play a vital role in our everyday lives. Everything from our most basic perceptions [50] to our deepest, most heartfelt love [37] is influenced by affective processes. This is not only true for encounters in the 'real world', but also extends to our experiences with digital technology [32]. In the past three decades human-computer interaction (HCI) research has started to adopt a more holistic view of the experience of computer interaction, recognizing non-instrumental elements such as fun (see for instance [5]). This field of investigation has been referred to as User eXperience (UX), and is principally concerned with studying emotional responses to HCI [18, 19, 25, 29].

In HCI, emotions can have a broad range of effects on, for instance, the shaping of the interaction, the communication about the interaction, as well as the evaluation of the object of interaction [14, 18]. Consequently, measuring emotions in interaction with a broad range of interactive products [17] and interfaces [24] has been a primary concern of UX researchers. The methods used for such investigations are often validated emotion measurement instruments from the field of experimental psychology (see for instance [18, 28, 40]). A downside of these methods is that they are not always well suited for the highly interactive nature of digital media. Most methods are applied post-hoc, providing a measurement of the overall experience. Consider for instance websites, where users can quickly navigate between different pages through hyperlinks. In such a case a post-hoc measurement would only provide insight into the cumulative experience of all the pages. The range of emotional responses that people experience in relation to individual pages, or elements of individual pages, would be lost. Therefore, the investigation of emotions in UX research would benefit from methods specifically geared towards highly interactive contexts, such as interaction with websites. The development process and initial validation of such an instrument is reported in this paper.

## RELATED WORK

Visual interfaces such as websites, mobile, and desktop operating systems, can elicit emotions in a number of ways. Hassenzahl [16] proposes that an individual's experience with technology depends on the perceived pragmatic and hedonic qualities of the product. Pragmatic quality resembles the notion of usability (e.g. ease of use), while hedonic quality refers to pleasure of use. Pragmatic quality is in essence a

*CHI 2013*, April 27–May 2, 2013, Paris, France.

"hygiene factor" [18] that will not produce positive emotions in itself, but can result in strong negative emotions when, for example, usability breaks down. Unresponsive controls for instance, can cause negative affective responses in the user [36]. Similarly, an overly complex and unclear ordering of visual elements of an interface can lead to heightened arousal and less positive evaluations of that interface [28, 40, 47]. Hedonic quality on the other hand can be a source of positive affect. An oft studied element of hedonic quality, visual appeal, influences the user's experience early on in perception [26, 27, 41], in the form of a rapid affective judgement ([30, 50] see also [27]). Furthermore, a good visual design has the potential to negate existing usability problems, resulting in more favorable evaluations of the interface [39, 42]. The initial affective reaction to the visual appeal of an interface can have a priming effect [51], and influence later evaluations of that interface [7]. The impact of visual appeal is thus not limited to the initial perception.

### Measuring emotions

The measurement of emotions typically involves methods that measure a single component of emotion. These components include activation of the autonomous nervous system, motor expressions, behavioral tendencies, and subjective feelings [38].

#### *Psychophysiological measurements*

Psychophysiological measurements are often used to obtain continuous measurements of emotions during interaction with an interface. Scheier et al. [36] used galvanic skin response (GSR), and blood volume pulse (BVP) to measure activation of the autonomous nervous system as an indication of emotional arousal resulting from a frustrating game. Ward and Marsden [47] used similar measurements of emotional arousal (i.e. GSR, BVP, and heart rate) to detect usability problems occurring during interaction with a website. Physiological responses to usability issues were also studied by Thüring and Mahlke [40] who, next to GSR as a measure of arousal, employed facial electromyography (EMG) to measure minute changes in electrical activity of two facial muscles as a measure of emotional valence (i.e. positive/negative judgements). Similarly, Hazlett and Benedek [20] measured the desirability of certain software functions, using facial EMG as a measure of emotional valence. An issue with using psychophysiological measurements to measure emotions that occur during interaction with a visual interface is that it is often difficult to find which element of the interface is responsible for the emotional reaction. The autonomous nervous system may be activated by non-emotional events, such as increases in concentration, which could confound measurements of emotion [48]. Furthermore, measurements of changes in physiology are an indirect measurement of emotions, thus a delay between the measurement and the actual emotion is to be expected [43]. This is especially problematic when considering the highly interactive nature of most visual interfaces, for instance while rapidly going back and forth between two web pages, or opening and closing programs in an operating system.

#### *Self-report*

The most common method used to ascertain someones affective state is self-report measurements [2, 33]. This is partly due to the complexities of psychophysiological measurements. Malhke and Minge [28], for instance, found ratings of valence and arousal using the Self Assessment Manikin (SAM) [1] to corroborate physiological measurements during interaction with a mobile phone interface. Deng and Poole [7] successfully used self-report measurements of valence and arousal to measure emotional responses to different web page designs. A strength of emotional self-report is that it is easy to apply and interpret. Downsides are that it is a subjective measure, sensitive to bias, as well as subject to priming effects of forced-choice questionnaires [15, 37]. Unlike physiological measurements, which are continuous, emotional self-report is mostly used as a post-hoc measurement. Feldman-Barrett [13] argues that the time that passes between the stimulus and the reported emotion may negatively influence the accuracy of the measurement. The longer the time between the stimulus presentation and the self-report of emotion, the more the respondent will rely on memory to report his or her subjective feelings [13, 33]. In addition, as most self-report methods are verbal measurements, they are difficult to apply cross-culturally, as well as difficult to use with less literate populations (e.g. children) [1, 8]. Furthermore, emotions are difficult to verbalize, thus responding to one's emotional state with verbal labels requires considerable cognitive involvement which may influence the response [8, 49].

#### *Non-verbal self-report*

In an attempt to improve on verbal self-report methods, researchers have developed non-verbal self-report methods to measure the subjective feeling component of emotion. One of the most well-known is SAM [1], which is based on an abstract cartoon figure that conveys emotional valence, arousal and dominance. Each of these dimensions is represented by five different visualized states on a horizontal 9-point scale. While the dominance dimension is often unused because it lacks discriminative power, SAM has proven successful in measuring the valence and arousal dimensions [1], which are considered the underlying dimensions of all emotions [35]. Desmet [8] took a different approach with the development of PrEmo (Product Emotion), an instrument to measure emotional responses to products. PrEmo consists of fourteen animations of a cartoon character that expresses specific emotions through facial, bodily, and vocal expressions. After being exposed to a product, people indicate, for each animation, how strongly the depicted emotion was felt. PrEmo is based on the notion that people can accurately identify discrete emotions from bodily signals such as facial expressions [10] and body language [46]. Recognition of the PrEmo animations in western cultures (i.e. Finland, the Netherlands, and the United States) ranges from 63 to 100 percent [8]. PrEmo has been applied in studies into automobile designs [8] and mobile phone designs [9]. Similarly, the Pictorial Mood Reporting Instrument (PMRI) [45] uses three sets of cartoon figures (i.e. male, female, and genderless) that express one neutral mood and eight specific moods taken from the circumplex model of affect [34]. The initial validation study showed that the PMRI images were recognized with an accuracy between 38 and 94 percent. A considerably lower recognition accuracy was obtained for positive moods compared to negative ones. PMRI is envisioned as a communication tool to share moods between users, as well as a general tool to measure moods. In comparison to emotions, moods are more diffuse affective states that are relatively long in duration, are less intense, and often do not have specific elicitors [37]. This distinction is important in that the PMRI would seem particularly well suited to measure someone's general affective state (i.e. mood) at a point in time, but would be less suitable to measure emotional responses to specific elicitors such as products or interfaces.

Evaluation methods that are strongly related to non-verbal emotional self-report scales, are found in pain assessment scales for children [6, 21, 22]. For example the Faces Pain Scale-Revised [21] uses six facial expressions in a horizontal orientation with the endpoints representing "no pain" and "very much pain". The instrument has been proven to be a highly reliable method of pain assessment for children [22]. Comparable in design to the Faces Pain Scale-Revised, but different in application area, is the Smileyometer. This instrument also consists of a scale of five abstract facial expressions, and is used to measure children's experience with technology [31]. Though not specifically aimed at measuring emotions, these types of scales do provide valuable insights into how pictorial representations might be used instead of words in self-report measurements, especially measurements taken from children.

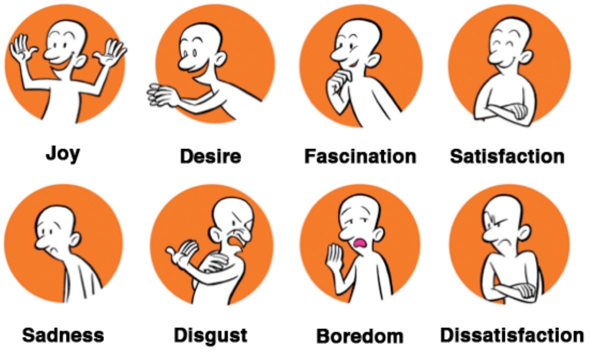


**Figure 1. The eight images of the LEMtool.**

## THE DESIGN OF THE LEMTOOL

As was outlined in the previous section there are a number of ways to measure emotions. However, each of the methods described has its downsides. For in-process measurements, psychophysiological signals can be used, but these measures can be difficult to apply and interpret [48]. Verbal self-report methods might be easier to interpret, but add the difficulty of having to translate words for different languages, as well as adding cognitive load [8, 49]. Therefore, non-verbal self-report measures of emotion provide a viable alternative to verbal self-report measures. However, considering the measurement of emotions during interaction with a visual interface, a number of things stand out in regard to existing non-verbal self-report methods. First, methods either use continuous scales of the underlying dimensions of emotion (SAM [1]), scales of emotion related concepts (pain assessment [21]) or scales for reporting experience with technology [31]. Second, the only method that measures specific emotions uses an animated cartoon figure (PrEmo [8]), while the only method that uses still images of discrete affective states is geared towards measuring moods instead of emotions (PMRI [45]). Finally, no method is specifically aimed at measuring emotions during interaction with visual interfaces.

The LEMtool (Layered Emotion Measurement tool) [23] was developed with the requirements that it should be easily deployable during interaction with a visual interface (i.e. measurements in-process), that it should be comprehensible (i.e. not too demanding for the participant), and that it should have the possibility to be used cross-culturally. Similarly to PrEmo [8], the idea behind the LEMtool was to measure a finer granularity of discrete emotions rather than general emotional states. However, PrEmo uses animations, which take time to play in their entirety. This would severely disrupt the interaction with a visual interface, making the use of animations unsuitable for deployment during interaction. For these reasons it was decided that the development of a new set of visualizations was necessary.

### *Selecting and visualizing emotions*

The LEMtool consists of a cartoon figure that expresses eight discrete emotions using facial expressions and body postures. The instrument consists of four positive and four negative emotions (see Figure 1). The emotions were selected from a study into the emotional impact of web pages, in which emotion terms from the circumplex model of affect [34] were divided into eight octants along the valence and arousal dimensions [4]. Emotion words that represented states with a neutral valence were not considered, as these words might denote non-emotional states (see also [37]). In the selection of the eight emotion words from the remaining six octants, the possibilities for visualizing each emotion term was considered. Findings from studies into facial expressions [10], as well as the emotions that were visualized in PrEmo [8] were taken into account here. Furthermore, the concept of emotion families was considered [11]. This concept states that although there are numerous emotion terms, each term may belong to a group of related affective states. For instance, dissatisfaction would belong to the same emotion family as anger and rage, but represent a less intense emotional state. It is suggested here that dissatisfaction might be a more relevant emotion than anger or rage in the context in which the LEMtool will be employed (i.e. evaluations of visual interfaces)[2]. That is not to say these more intense emotions cannot be elicited during such interactions, just that respondents using the LEMtool are more likely to experience dissatisfaction with an interface than to be enraged by it.

The design of the LEMtool is based on the notion that people can identify specific facial and bodily expressions of emotion [10, 12, 46], especially when such expressions are caricatured (i.e. undone of any elements distracting from the expression) [3]. The LEMtool images were created in collaboration with a professional cartoonist, who was provided with general guidelines about the composition of the facial expressions and body postures. The LEMtool was designed as an interactive instrument deployed during interaction with a visual interface, allowing participants to indicate responses in-process. The way the LEMtool is used during interaction with a web page, is depicted in Figure 2.

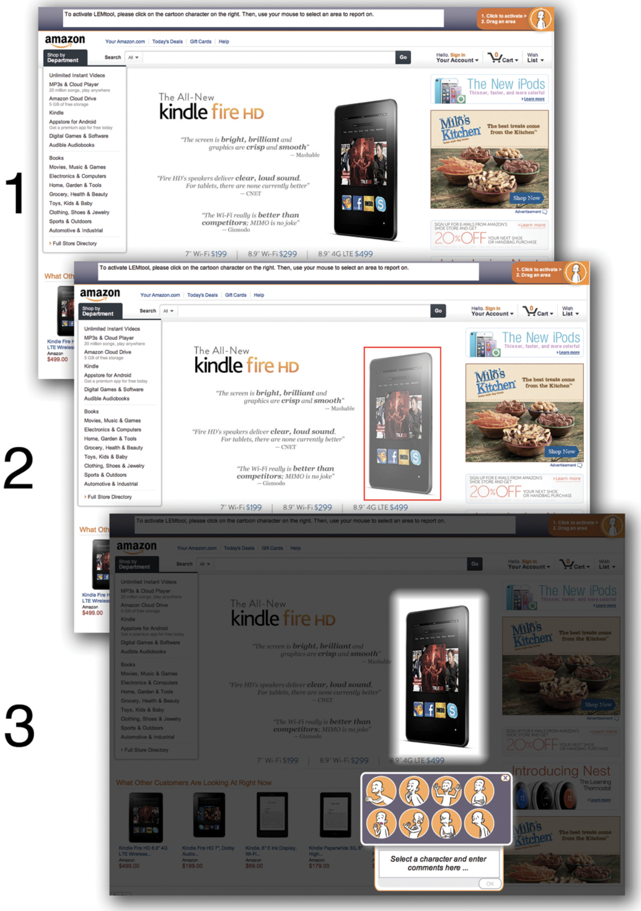

1
2
3
To activate LEMtool, please click on the cartoon character on the right. Then, use your mouse to select an area to report on.
1. Click to activate > 2. Drag an area
Select a character and enter comments here ...

**Figure 2. The steps required to indicate an emotion for a certain area of a visual interface, in this case a web page. Step 1: activate the LEMtool. Step 2: select an area of the website. Step 3: indicate an emotion.**

## VALIDATION OF THE LEMTOOL IMAGES

A validation study was conducted in order to assess the recognizability of the LEMtool images. Using a number of different response formats, participants were asked to indicate which emotions they thought the LEMtool images were displaying.

### Pilot study

Prior to the validation study, a pilot study was conducted. The goal of the pilot study was to obtain a baseline for the recognizability of the LEMtool images.

| Image | Target label percentage |
|---|---|
| Joy | 83.9 |
| Desire | 83.9 |
| Fascination | 81.7 |
| Satisfaction | 80.6 |
| Sadness | 80.6 |
| Disgust | 96.8 |
| Boredom | 100 |
| Dissatisfaction | 87.1 |

**Table 1. Recognition accuracy of the LEMtool images in the pilot study. All images were recognized at above chance level (i.e. 50%, $p$ <.001).**

#### *Participants*

Participants were Master's students at the Technical University of Delft enrolled in a course on product experience. In total 38 male and 55 female students participated (N = 93). Age ranged from 21 to 31 (M = 23.4, SD = 1.9).

#### *Apparatus*

Participants were presented with two sheets of A4 paper stapled together. The LEMtool images were printed in black-and-white with a size of 5 by 5 centimeters. The eight emotion terms corresponding to the images were presented in English next to each image.

#### *Procedure*

The procedure was explained to the entire group by the lecturer. Participants were asked to select either one or more of the eight given emotion terms, or add a word of their own, that according to them, would best describe the emotion expressed by the LEMtool image. This last option was added to reduce response bias as a result of the forced-choice format [15]. Finally the participants were instructed to indicate their gender, age and first language. Participants were specifically told that they were not allowed to talk to each other. The entire procedure took no more than five minutes.

#### *Results*

Table 1 lists the percentages of participants who selected the target label for each LEMtool image. Only selection of the target label was considered a correct response. Responses containing selection of more than one label, or responses consisting of an added label were considered an incorrectly selected label. Binomial tests were computed for the proportion of participants who selected each emotion label for a given target label. Chance was set to 50% for each emotion. This chance level was based on Ekman's [10] considerations on how some facial expressions might be most likely confused with similar expressions. Here the chance level reflects a choice between two emotions that may be expected to be confused based on their morphology (i.e. joy-satisfaction, desire-fascination, sadness-boredom, and disgust-dissatisfaction). Note that the chosen chance level is more stringent than that typically suggested for forced-choice facial expression recognition tasks (see [15]). Table 1 shows that all of the LEMtool images were recognized as the emotions they were intended to display at above chance level ($p$ <.001). These results are comparable to, and in some cases

| Image | Rating for target label (5-point scale) | Std. Deviation |
|---|---|---|
| Joy | 4.9 | 0.3 |
| Desire | 4.8 | 0.6 |
| Fascination | 4.4 | 0.9 |
| Satisfaction | 4.9 | 0.6 |
| Sadness | 4.7 | 0.6 |
| Disgust | 4.6 | 0.8 |
| Boredom | 4.9 | 0.4 |
| Dissatisfaction | 4.7 | 0.7 |

**Table 2. Average ratings for target emotion label in the first task of the validation study. All target emotion labels were rated significantly higher than all other labels ($p$ <.001).**

| Image | Target label percentage |
|---|---|
| Joy | 95.1 |
| Desire | 89.0 |
| Fascination | 89.0 |
| Satisfaction | 91.5 |
| Sadness | 89.0 |
| Disgust | 86.6 |
| Boredom | 100 |
| Dissatisfaction | 93.9 |

**Table 3. Recognition accuracy of the LEMtool images in the second task of the validation study. All images were recognized at above chance level (i.e. 50%, $p$ <.001).**

exceed, recognition accuracies obtained in studies into basic facial expressions [10], as well as results from studies using similar visualizations of emotions [8, 45]. This indicates that the LEMtool images were relatively accurately recognized as the emotions they were intended to display. Based on these findings, it was decided not to make any changes to the images at this point.

## Validation study

The validation study was carried out to assess possible confusions between images, using a broader sample of participants, and different response formats as compared to the pilot study.

### *Participants*

A notice of the study was posted on the student website of the University of Twente, and on a design and emotion related website. In total 46 male and 36 female participants took part in the study (N = 82). Participants' ages ranged from 18 to 59 (M = 30.0, SD = 10.2).

### *Apparatus*

The study was conducted using a purpose-built website. The website was compliant with the most popular browser standards (i.e. Microsoft Internet Explorer, Mozilla Firefox, Opera) presenting the contents of the study consistently across browsers. The LEMtool images were 200 by 200 pixels in size and displayed in an orange circle (see Figure 1).

### *Procedure*

Participants were first presented with an introductory text detailing the goal of the study and outlining the general procedure. In the first task, participants were presented with a list of the eight emotion terms and for each of the eight subsequently presented LEMtool images were asked to rate, on a five-point scale, the terms that they thought were most prominently present in the presented image. The scale ranged from 1 ("This emotion is not present") to 5 ("This emotion is strongly present"). Intermittent response options were not labeled. The order of the eight emotion terms, as well as the order in which the images were presented, was randomized for each participant. In the second task, participants were asked for each of the eight subsequently presented images, to select one of eight emotion terms that they felt best described the emotion expressed in the image. The response option "none of these terms" was added to reduce response bias. Again, the order of the eight emotion terms, as well as the order in which the images were presented, was randomized for each participant. All texts in both tasks were presented in Dutch.

### *Results*

Table 2 lists the average rating given to the target label for each LEMtool image. For all images, a one-sample $t$-test was performed, comparing the average rating for each target label with the average rating of all other labels. For all images the target label was rated as significantly higher ($p$ <.001) than all other labels. This indicates little confusion between the emotions depicted in the LEMtool images. Additionally, the low standard deviations in Table 2 indicate consensus among participants about the most appropriate label. Table 3 lists the percentages of participants who selected the target label for each LEMtool image. Only selection of the target label was considered a correct response. Identical to the pilot study, chance level was set to 50%. Binomial tests were computed for the proportion of participants who selected each emotion label for a given target label. Table 3 shows that all of the LEMtool images were recognized as the emotions they were intended to display at above chance level ($p$ <.001). Again, these results are highly comparable to previous research [8, 10, 45].

### *Conclusions*

Findings from the validation study show that participants were relatively successful in decoding the emotions intended by the LEMtool images. Similarly high recognition accuracy was obtained using a number of different response formats, limiting the possibility that findings are an artifact of the response format (see also [15]). The fact that the recognition accuracy of the images is comparable to, and in some cases exceeds, that of existing instruments such as PrEmo [8] and PMRI [45] is encouraging, and shows the potential of using the LEMtool images to measure emotional responses. As a first application of the LEMtool a study was designed to see if the instrument could be used successfully to indicate emotional responses to visual stimuli.

## CASE STUDY: VISUAL APPEAL

Visual appeal is an important element of the hedonic quality of an interface [18]. Research suggests that judgements of visual appeal are in essence emotional judgements that occur rapidly, and can bias later judgements of visual appeal [27].

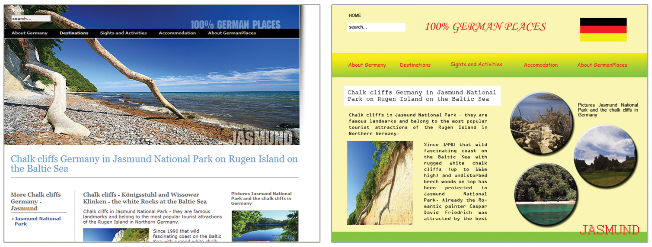
Therefore it is important for the LEMtool to be able to measure responses to the visual appeal of interfaces. A study was conducted to assess whether the LEMtool could be used to measure emotional responses to the visual appeal of web pages. As a first application of the LEMtool to measure visual appeal of web pages, the focus was on visual appeal as a general concept. The accent of the case study was therefore on measuring the difference between high and low visual appeal web pages. Both brief (50 ms) and long (free-viewing) exposure times were used. Additionally, the case study served as a first evaluation of the way the LEMtool was envisioned to be used, namely by having participants indicate positive and negative emotions on different areas of a number of web pages.

### Manipulation check

Two independent web designers created a high and a low visual appeal version of the same web page. The designers were only provided with guidelines aimed at keeping the type, organization, and presentation of information, as well as the content and perceived functionality of the web page, consistent. No specific instructions were given to the designers regarding manipulations of visual appeal. The manipulations of visual appeal relied on the expertise of the designers. The 24 web pages covered three topics, namely: Einstein, a holiday island, and medical information about headaches [44].

In an online survey, promoted through social media (i.e. Twitter, Facebook), a total of 31 participants rated the visual appeal of the 24 web pages. Similar to [41] a ten-point rating scale (1 = very unappealing, 10 = very appealing) was used. Each web page screenshot was displayed for 500 ms, in a resolution of 1024x768 pixels, without visible browser elements. All texts were presented in Dutch.

A paired-samples *t*-test was performed, comparing the visual appeal ratings for high visual appeal and low visual appeal web pages. Overall, high visual appeal web pages received a mean rating of 5.39 (SD = 1.34) while low visual appeal web pages received a mean rating of 3.08 (SD = 1.15). This difference was significant ($t(30) = 14.4$, $p < .001$). From this result it was concluded that the web pages did indeed differ on visual appeal, and were therefore suitable for use in the main study.



Chalk cliffs Germany in Jasmund National Park on Rugen Island on the Baltic Sea

More Chalk cliffs Germany - Jasmund

› Jasmund National Park

Chalk cliffs - Königsstuhl and Wissower Klinken - the white Rocks at the Baltic Sea

Chalk cliffs in Jasmund National Park - they are famous landmarks and belong to the most popular tourist attractions of the Rugen Island in Northern Germany.

Since 1990 that wild fascinating coast on the Baltic

Pictures Jasmund National Park and the chalk cliffs in Germany

Chalk cliffs Germany in Jasmund National Park on Rugen Island on the Baltic Sea

Chalk cliffs in Jasmund National Park - they are famous landmarks and belong to the most popular tourist attractions of the Rugen Island in Northern Germany.

Since 1990 that wild fascinating coast on the Baltic Sea with rugged white chalk cliffs (up to 161m high) and undisturbed beech woods on top has been protected in Jasmund National Park. Already the Romantic painter Caspar David Friedrich was attracted by the best

Pictures Jasmund National Park and the chalk cliffs in Germany

**Figure 3. Example of a holiday island high visual appeal web page (left) and a holiday island low visual appeal web page (right) used in the study.**

## Visual appeal study

### *Participants*

In total 43 (13 male, 30 female) individuals participated in the study. Twenty-four of these participants were first and second year psychology and communication science students who received course credits for participation. The remainder of the participants were approached by the researcher and asked to volunteer in the study. Participants age ranged from 18 to 31 (M = 22.4, SD = 3.19). Participants with color blindness were excluded.

### *Apparatus*

The study consisted of two separate phases. In the 50 ms phase, stimuli were presented using E-Prime 2.0 displayed on a 17 inch Samsung SyncMaster 750s CRT monitor (Samsung Electronics, Seoul, South Korea). The monitor was set to a screen resolution of 1024x768 pixels at 60 Hz. Brightness was set to 85 and contrast to 100 with color temperature set to 9300 °K. Participants used a standard computer keyboard to indicate their responses. In the free-viewing phase, stimuli were presented using a purpose built online environment running in Firefox 3.6.3, displayed on an HP Compaq LE1711 17 inch LCD monitor (Hewlett-Packard, Palo Alto, California). The monitors native resolution of 1280x1024 pixels at 60 Hz was used. The monitor was set to a brightness level of 90 and contrast to 80 with color temperature set to 6500 °K.

### *Procedure*

Participants were given a written explanation of the procedures. After informed consent was obtained, participants took place behind the CRT monitor and followed the instructions on the screen. First, five test web page screenshots (e.g. amazon.com, cnn.com) were shown for 50 ms each, to allow participants to get used to the short exposure time. Participants pressed the space bar to present the next web page, and used the keys 1-0 on the keyboard to indicate a response from 1 (visually unappealing) to 10 (visually appealing). After participants gave a response the eight LEMtool images appeared. Each image corresponded to a numbered key (1 to 8) on the keyboard. By pressing a single key corresponding to a single LEMtool image, participants indicated the emotion that the web page elicited in them. Participants could use the 0-key to indicate that the web page did not elicit any emotion. Once the participants had rated all five test web pages, a selection of twelve web pages (6 high and 6 low visual appeal) was subsequently presented in random order to the participants. The procedure was identical to that of the test pages. After rating all web pages in the 50 ms phase, participants moved on to the free-viewing phase. Instructions for the use of the LEMtool (see Figure 2) were given in the online environment used to present the stimuli. Participants were again presented with the same five test web pages displayed in random order. Each web page would stay on the screen until participants pressed a key 1 to 0 representing a rating of 1 (visually unappealing) to 10 (visually appealing). After giving a visual appeal rating, the LEMtool would appear in the top-right corner of the screen and participants had to select areas of the web page using the computer mouse, and attach a LEMtool image to that area. Participants could give as many LEMtool indications as they liked, but were instructed to only rate elements that were related to the visual appeal of the web page. They were told that reading texts on the web pages was not required.

|  |  | Visual Appeal ratings grouping |  |  |  |
|---|---|---|---|---|---|
| **LEMtool image** | N | 1 | 2 | 3 | 4 |
| Joy | 14 | 7.50 | - | - | - |
| Desire | 36 | 7.67 | - | - | - |
| Fascination | 108 | - | 6.19 | - | - |
| Satisfaction | 63 | - | 6.08 | - | - |
| Sadness | 70 | - | - | 4.39 | - |
| Boredom | 106 | - | - | 3.70 | - |
| No emotion | 18 | - | - | 4.00 | - |
| Disgust | 78 | - | - | - | 2.35 |
| Dissatisfaction | 23 | - | - | - | 2.30 |

**Table 4. Tukey's HSD for average visual appeal ratings per LEMtool indication in the 50 ms phase. All groups differ from each other at *p* <.001**

Once participants had rated the five test web pages, the twelve stimulus web pages (6 high and 6 low visual appeal, for each participant different from those in the 50 ms phase) were displayed in random order. The exact same procedure as with the test web pages was followed. The second phase was concluded after participants had rated all 12 web pages using the 10-point rating scale and LEMtool and had indicated their age, gender, and native language. All texts in both phases were presented in Dutch.

### *Results*

First a paired-samples *t*-test was calculated for the average visual appeal ratings given to high visual appeal and low visual appeal web pages with a 50 ms exposure time. Overall, high visual appeal web pages were rated with a 5.61 (SD = .82) while low visual appeal web pages were rated with a 3.85 (SD = 1.12). This difference was significant ($t(42) = 11.47$, $p < .001$). An identical analysis was conducted for the free-viewing phase, during which high visual appeal web pages were rated with a 5.93 (SD = 0.99) and low visual appeal web pages with a 3.47 (SD = 1.04). Again, this difference was significant ($t(41) = 16.34$, $p < .001$). In addition, a significant correlation was found between ratings in the 50 ms and free-viewing conditions (r = .88, $p < .001$). These results match findings by Lindgaard et al. [27] showing that the visual appeal of web pages can be judged in 50 ms, and that this judgement does not change substantially after longer exposure.

To determine whether the LEMtool would show a similar difference between high and low visual appeal web pages, for both the 50 ms and free-viewing phases, a cross-tabulation for visual appeal (two rows, high and low visual appeal) and LEMtool image (nine columns for the 50 ms phase, and eight columns for the free-viewing phase) was constructed. A Chi-square test of independence was performed to assess significance. For both the 50 ms phase ($\chi^2 = 89.56$, $df = 8$, $p < .001$) and the free-viewing phase ($\chi^2 = 251.01$, $df = 7$, $p < .001$) the difference between high and low visual appeal web pages was significant. Thus, for both brief and extended exposure times, the LEMtool differentiated between high and low visual appeal web pages.

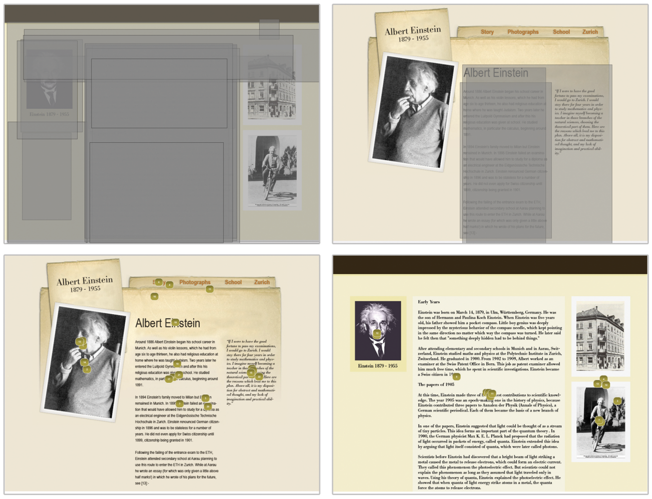


**Figure 4. Two high visual appeal web pages about Einstein (Einstein - early years on the left, and Einstein - Albert Einstein on the right) used in the case study. The top two pages show a visualization for the LEMtool boredom emotion. The grey areas were selected by participants indicating boredom. The bottom two pages show a visualization for the LEMtool fascination emotion. The dots indicate the centre of a selection area indicated by participants.**

To assess the relation between the visual appeal ratings and the LEMtool indications more in-depth, a one-way ANOVA with LEMtool indications (9 levels) as the independent variable and visual appeal ratings as the dependent variable, was computed ($F(8) = 113.68$, $p < .001$). A post-hoc Tukey's HSD revealed that certain LEMtool images were related to different visual appeal judgements (see Table 4). These findings indicate that the LEMtool images covered a range of emotional responses related to visual appeal. Moreover, the LEMtool images that display positive emotions were related to higher visual appeal ratings, while LEMtool images displaying negative emotions were related to lower visual appeal ratings.

To illustrate how the LEMtool can provide more detailed information about individual web pages, Figure 4 depicts two data visualization methods for two similar looking high visual appeal web pages from the same topic (i.e. Einstein). Note that this is only an example of a method for treating LEMtool data. A more comprehensive analysis of LEMtool indications for elements of all the web pages would be beyond the scope of this paper. Table 5 shows the cross-tabulation for both pages. What can be observed from both the visual data, as well as the table, is that, while both web pages are high visual appeal web pages, the composition of LEMtool emotions attached to each page differs. This was most apparent for the most selected emotions for each page (i.e. fascination, satisfaction, and boredom). The visualizations at the top of Figure 4 show that the "early years" page on the left elicited boredom in participants, more than the "Albert Einstein" page on the right did. For both pages, boredom was mainly indicated for the central text area, but considerably more so for the early years page. In addition, for the early years page the image of Einstein was also indicated as boring, which was not the case for the image on the Albert

| Web page | Joy | Desire | Fascination | Satisfaction | Sadness | Boredom | Disgust | Dissatisfaction | Total |
|---|---|---|---|---|---|---|---|---|---|
| Einstein - Early Years | 3 | 0 | 13 | 16 | 4 | 17 | 2 | 2 | 57 |
| Einstein - Albert Einstein | 1 | 4 | 22 | 23 | 3 | 3 | 3 | 0 | 59 |
| Total | 4 | 4 | 35 | 39 | 7 | 20 | 5 | 2 | 116 |

**Table 5. Cross-tabulation for the LEMtool emotions indicated for two high visual appeal Einstein web pages. A Chi-square test of independence shows the web pages differ significantly on the LEMtool emotions ($\chi^2$ = 20.69, *df* = 7, *p* <.01)**

Einstein page. Furthermore, the visualizations at the bottom of Figure 4 show that the Albert Einstein page elicited more fascination in the participants than the early years page. For both web pages, the images, as well as the main text elicited fascination, but this was more so for the Albert Einstein page. Additionally, the header and quoted text in the right web page elicited fascination. The differing LEMtool indications for the web pages in Figure 4 are supported by the visual appeal ratings for each page. The early years page was rated at 5.14 (SD = 1.68) overall, while the Albert Einstein page was rated at 6.67 (SD = 1.32) overall. This difference was significant ($t(40) = -3.27$, $p < .01$). As indicated by the cross-tabulation (Table 5) the LEMtool revealed a similar difference between both pages. Moreover, as shown by the visualizations in Figure 4, the LEMtool provides additional insight into why these web pages differ on visual appeal. This difference mainly stems from differing LEMtool indications for boredom, fascination and satisfaction (Table 5). The Albert Einstein page was rated as more satisfying, fascinating and less boring than the early years page.

### *Conclusions*

The goal of the case study was to demonstrate that the LEMtool can be used to differentiate between high and low visual appeal web pages. Furthermore, the case study served as a first evaluation of the way the LEMtool was envisioned to be used (see Figure 2). First of all, the case study demonstrated that participants were able to use the LEMtool to select specific areas of a web page and indicate their emotional response. While a test session was required for participants to familiarize themselves with the way the LEMtool is used, all participants were capable of indicating their responses without issues. Second, the case study supports findings by Lindgaard et al. [27] by showing that the visual appeal of a web page can be judged accurately after participants have seen the web page for only 50 ms. Moreover, this judgement remained consistent for visual appeal ratings after longer stimulus exposure. The results from the case study showed that the LEMtool revealed a similar differentiation between high and low visual appeal web pages for both the 50 ms phase and the free-viewing phase. Moreover, findings showed that the LEMtool images relate to a range of visual appeal judgements. The LEMtool emotions Joy and Desire were related to high visual appeal judgements, Fascination and Satisfaction to moderately high judgements, Sadness and Boredom to moderately low judgements, and Disgust and Dissatisfaction to low visual appeal judgements. The alternative explanation that these findings represent confusion between the images is unlikely, because little confusion was found between the LEMtool images in the validation study (see Table 2). Thus, these findings indicate that the positive LEMtool emotions are related to positive visual appeal judgements and that the negative LEMtool emotions are related to negative visual appeal judgements. Third, results from the free-viewing phase of the case study, in which participants used the LEMtool to select areas of web pages that elicited a certain emotion, illustrated how the LEMtool can provide additional insights. Analysis of two high visual appeal web pages showed that the LEMtool could aid in revealing which elements of a web page are mainly responsible for the outcome of a certain visual appeal judgement.

## CONCLUSIONS AND DISCUSSION

In this paper the development process and initial validation of the LEMtool were outlined. A validation study revealed that the recognition accuracy of the images was comparable to, and in some cases exceeded, recognition ratings found in other research into non-verbal self-report of emotions [8, 45]. In a case study on visual appeal judgements, results obtained with the LEMtool were supported by findings from visual appeal ratings. Using an interactive version of the LEMtool, participants were able to select areas of web pages to indicate their emotional responses. Results revealed that the LEMtool could provide additional insights into which elements of the web pages were most prominent in forming the visual appeal judgements.

A number of limitations of the current research deserve mentioning. First, the validation study was carried out for one culture only. While the current validation study offers a good starting point for validating the LEMtool images, additional studies are required to further assess the validity of the images across different cultures. Second, while the use of screenshots in the case study allowed for better experimental control in studying the relation between visual appeal judgements and the LEMtool, emotions resulting from usability issues during interaction with a visual interface were not studied. However, considering the vital role visual appeal plays in the perception of and interaction with visual interfaces [26, 27, 39, 42], it can be argued that studying the LEMtool's capabilities to measure visual appeal judgements is crucial for the validation of the instrument. Finally, one could argue that in the free-viewing phase of the case study participants not only rated the web pages on visual appeal, but were also influenced by texts and images. However, the use of complete web pages instead of, for instance, abstract mock-ups without texts and images, makes the study more ecologically valid. Moreover, visual appeal ratings between the 50 ms phase and the free-viewing phase of the study, were highly comparable. Considering that it is unlikely that participants in the 50 ms phase reported on anything other than a first visual impression [27], it would seem that participants were actually able to focus on rating visual appeal in the free-viewing phase.

Taking these limitations into account, the investigation presented in this paper provides a good starting point for the further development of the LEMtool. Our aim is to provide a useful tool for designers in different stages of the design process of a visual interface. Early on in the design process, the LEMtool could be used to compare different prototype designs, in a similar fashion to the example in the case study. This could provide valuable insights into different design decisions. Furthermore, the LEMtool could provide insights during interactions with a visual interface. Here, the changing emotional responses of users can be studied over time, and, based on the reported emotions, interventions can be made during different stages of the interaction. For example, if a user indicates dissatisfaction at a certain point during a search task, such indications could be used to actively prompt users with information they might be looking for. Finally, we hope to further develop the LEMtool as a general research tool for measuring emotional responses during interaction with a visual interface.

## ACKNOWLEDGEMENTS

This publication was supported by the Dutch national program COMMIT, and SusaGroup.

## REFERENCES

1. Bradley, M. M., and Lang, P. J. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry 25*, 1 (1994), 49 – 59.
2. Brave, S., and Nass, C. Emotion in human-computer interaction. In *Handbook of Human-Computer Interaction*, J. Jacko and A. Sears, Eds., Lawrence Erlbaum Associates. (2008), 77–92.
3. Calder, A. J., Rowland, D., Young, A. W., Nimmo-Smith, I., Keane, J., and Perrett, D. I. Caricaturing facial expressions. *Cognition 76*, 2 (2000), 105 – 146.
4. Capota, K., van Hout, M., and van der Geest, T. Measuring the emotional impact of websites: a study on combining a dimensional and discrete emotion approach in measuring visual appeal of university websites. In *Proceedings of DPPI '07*, ACM (2007), 135–147.
5. Carroll, J. M., and Thomas, J. M. Fun. *SIGCHI Bulletin 19*, 3 (1988), 21–24.
6. Chambers, C., Giesbrechta, K., Craiga, K., Bennett, S., and Huntsman, E. A comparison of faces scales for the measurement of pediatric pain: children's and parents ratings. *Pain 83*, 1 (1999), 25–35.
7. Deng, L., and Poole, M. S. Affect in web interfaces: a study of the impacts of web page visual complexity and order. *MIS Quarterly 34*, 4 (2010), 711–730.
8. Desmet, P. *Designing Emotions*. Unpublished doctoral thesis. Delft University of Technology, Delft, The Netherlands, 2002.
9. Desmet, P., Porcelijn, R., and Van Dijk, M. HOW to design WOW Introducing a layered-emotional approach. In *Proceedings of DPPI'05*, J. Jacko and A. Sears, Eds. (2005), 71–89.
10. Ekman, P. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin 115*, 2 (1994), 268 – 287.
11. Ekman, P. Basic emotions. In *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power, Eds. John Wiley & Sons Ltd., 1999, 45–60.
12. Ekman, P., and Friesen, W. V. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology 17*, 2 (1971), 124 – 129.
13. Feldman-Barrett, L. Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology 87*, 2 (2004), 266 – 281.
14. Forlizzi, J., and Battarbee, K. Understanding experience in interactive systems. In *Proceedings of DIS '04*, ACM Press (2004), 261–268.
15. Frank, M. G., and Stennett, J. The forced-choice paradigm and the perception of facial expressions of emotion. *Journal of Personality and Social Psychology 80*, 1 (2001), 75 – 85.
16. Hassenzahl, M. The Thing and I: Understanding the Relationship Between User and Product. In *Funology*, M. Blythe, K. Overbeeke, A. Monk, and P. Wright, Eds., vol. 3 of *Human-Computer Interaction Series*, Springer Netherlands (2005), 31–42.
17. Hassenzahl, M. User experience (ux): towards an experiential perspective on product quality. In *Proceedings of IHM '08*, ACM Press (2008), 11–15.
18. Hassenzahl, M., Diefenbach, S., and Gritz, A. Needs, affect, and interactive products facets of user experience. *Interacting with Computers 22*, 5 (2010), 353 – 362.
19. Hassenzahl, M., and Tractinsky, N. User experience-a research agenda. *Behaviour & Information Technology 25*, 2 (2006), 91–97.
20. Hazlett, R. L., and Benedek, J. Measuring emotional valence to understand the user's experience of software. *International Journal of Human-Computer Studies 65*, 4 (2007), 306 – 314.
21. Hicks, C. L., von Baeyer, C. L., Spafford, P. A., van Korlaar, I., and Goodenough, B. The faces pain scale revised: toward a common metric in pediatric pain measurement. *Pain 93*, 2 (2001), 173 – 183.
22. Huguet, A., Stinson, J. N., and McGrath, P. J. Measurement of self-reported pain intensity in children and adolescents. *Journal of Psychosomatic Research 68*, 4 (2010), 329 – 336.

23. Huisman, G., and Van Hout, M. The development of a graphical emotion measurement instrument using caricatured expressions: the LEMtool. In *Emotion in HCI-designing for people: proceedings of the 2008 international workshop*, C. Peter, E. Crane, M. Fabri, H. Agius, and L. Axelrod, Eds., Fraunhofer (2010), 5–7.

24. Kim, J., Lee, J., and Choi, D. Designing emotionally evocative homepages: an empirical study of the quantitative relations between design factors and emotional dimensions. *International Journal of Human-Computer Studies 59*, 6 (2003), 899 – 940.

25. Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., and Kort, J. Understanding, scoping and defining user experience: a survey approach. In *Proceedings of CHI '09*, ACM Press (2009), 719–728.

26. Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., and Noonan, P. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction 18*, 1 (2011), 1–30.

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

28. Mahlke, S., and Minge, M. Consideration of multiple components of emotions in human-technology interaction. In *Affect and Emotion in Human-Computer Interaction*, C. Peter and R. Beale, Eds., vol. 4868 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg (2008), 51–62.

29. McCarthy, J., and Wright, P. *Technology as Experience*. MIT Press, 2004.

30. Norman, D. *Emotional design: Why we love (or hate) everyday things.* Basic Books, 2004.

31. Read, J. Validating the fun toolkit: an instrument for measuring childrens opinions of technology. *Cognition, Technology & Work 10* (2008), 119–128.

32. Reeves, B., and Nass, C. *The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places*. CSLI Publications, 2002.

33. Robinson, M. D., and Clore, G. L. Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin 128*, 6 (2002), 934 – 960.

34. Russell, J. A. A circumplex model of affect. *Journal of Personality and Social Psychology 39*, 6 (1980), 1161 – 1178.

35. Russell, J. A. Emotion, core affect, and psychological construction. *Cognition & Emotion 23*, 7 (2009), 1259–1283.

36. Scheirer, J., Fernandez, R., Klein, J., and Picard, R. W. Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers 14*, 2 (2002), 93–118.

37. Scherer, K. R. What are emotions? And how can they be measured? *Social Science Information 44*, 4 (2005), 695–729.

38. Scherer, K. R. The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion 23*, 7 (2009), 1307–1351.

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

40. Thüring, M., and Mahlke, S. Usability, aesthetics and emotions in human-technology interaction. *International Journal of Psychology 42*, 4 (2007), 253–264.

41. Tractinsky, N., Cokhavi, A., Kirschenbaum, M., and Sharfi, T. Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human-Computer Studies 64*, 11 (2006), 1071–1083.

42. Tractinsky, N., Katz, A., and Ikar, D. What is beautiful is usable. *Interacting with Computers 13*, 2 (2000), 127–145.

43. Van den Broek, E., Janssen, J., Westerink, J., and Healey, J. Prerequisites for affective signal processing (ASP). In *Proceedings of Biosignals '11*, Springer (2009), 426–433.

44. Van Dongelen, R. *What is beautiful is good and usable*. Unpublished master thesis. University of Twente, Enschede, The Netherlands, 2008.

45. Vastenburg, M., Romero Herrera, N., Van Bel, D., and Desmet, P. PMRI: development of a pictorial mood reporting instrument. In *Proceedings of CHI '11*, ACM (2011), 2155–2160.

46. Wallbott, H. G. Bodily expression of emotion. *European Journal of Social Psychology 28*, 6 (1998), 879–896.

47. Ward, R., and Marsden, P. Physiological responses to different web page designs. *International Journal of Human-Computer Studies 59*, 12 (2003), 199 – 212.

48. Ward, R., and Marsden, P. Affective computing: problems, reactions and intentions. *Interacting with Computers 16*, 4 (2004), 707 – 713.

49. Wiles, J. A., and Cornwell, T. B. A review of methods utilized in measuring affect, feelings, and emotion in advertising. *Current Issues and Research in Advertising 13*, 1-2 (1991), 241–275.

50. Zajonc, R. B. Feeling and thinking: Preferences need no inferences. *American Psychologist 35*, 2 (1980), 151 – 175.

51. Zhou, H., and Fu, X. Understanding, measuring, and designing user experience: The causal relationship between the aesthetic quality of products and user affect. In *Human-Computer Interaction. Interaction Design and Usability*, J. Jacko, Ed., vol. 4550 of *Lecture Notes in Computer Science*, Springer (2007), 340–349.