doi:10.1093/iwcomp/xxxxxx

Relating Experience Goals with Visual User Interface Design

Jussi P.P. Jokinen[1] Johanna Silvennoinen[2] Tuomo Kujala[2]

[1] *Department of Communications and Networking, Aalto University*
[2] *Cognitive science, Faculty of Information Technology, University of Jyväskylä*

This paper examines the cognitive process of visually experiencing user interfaces. It contributes to a theory- and methodology-grounded understanding of how UIs are experienced with regard to various aesthetic criteria. This aids in considering the targeted experience goals in relation to visual design choices – a problem that designers usually have to tackle intuitively. The issue in explicitly relating designs to experiences stems from the complexity of the process in which visual stimuli are processed and turned into experiences. The authors present a cognitive top-down approach to this process, rooted in the appraisal theory and the theory of the predictive brain. Several predictions are derived via this approach, and an eye-tracking experiment with Web sites is presented that provides evidence of them. The experience goals and repeated exposure to stimuli are shown to affect appraisal times and visual scanpaths in Web pages' evaluation; this supports the top-down approach described. Researchers can use the findings to inform their theoretical and empirical pursuits as they strive to understand what makes design artefacts emotionally evocative, and the methodology outlined can assist designers in locating the visual regions and elements relevant for experiential design goals.

Categories and subject descriptors: xxx

Keywords: Human–computer interaction (HCI); laboratory experiments; HCI theory, concepts, and models

Responsible Editorial Board Member: Name

RESEARCH HIGHLIGHTS

- Visual experience can be tied in with interface design that applies cognitive appraisal theories of emotion.
- Different visual experience goals are shown to result in different visual search and appraisal behaviour.
- Familiarity with the visual stimulus predicts changes in visual search and appraisal behaviour.
- The methodology can be used as a tool for analysing how designs correspond with the targeted experience goals.

1. INTRODUCTION

User interface (UI) designers are constantly faced with a difficult problem: which visual elements are relevant, and in what way, when there are particular target user experiences? As one of the main problems of user experience research and design, this question has understandably received a large amount of attention (Schenkman and Jönsson, 2000; Hall and Hanna, 2004; Tuch et al., 2012; Seckler et al., 2015; Seo et al., 2016; Karvonen, 2000). The conclusion from the majority of the research is that the connection between user experience and visual UI design is far from simple. In fact, while there is agreement that visual UI design *does* influence user experience, little consensus exists on the nature of this connection and the ways in which designers can ascertain that their design choices influence experiences in a predictable way.

The number of distinct methodologies in user experience research is large (Bargas-Avila and Hornbæk, 2011;

Rogers, 2012; Thanh Vi et al., 2017), with specific data-collection methods ranging from questionnaires, interviews, focus-group discussions, and probes through various observation-based approaches to psychophysiological measurements. While methodological variety is not necessarily symptomatic of a theory crisis, it has been argued that user experience research does need a stricter theoretical approach with a methodologically explicit and verifiable basis (Bargas-Avila and Hornbæk, 2011; Jokinen, 2015; Jokinen et al., 2015; Silvennoinen and Jokinen, 2016a). This does not mean that data-collection methods would have to be ranked in terms of some universal validity; rather, for a given problem, the choice of methodology should be justifiable and its grounding verifiable. The worst-case scenario entails many hours of research and design going to waste on account of ineffective research methodologies and poor understanding of how the elements and work-flow of a UI can be connected to the targeted experiences.

This paper contributes to addressing the theoretical and practical problem of relating UI elements to experience goals. We demonstrate how appraisal theories of emotion (Scherer, 2009; Frijda, 1988; Ellsworth, 2013) and the theory of the predictive brain (Clark, 2013) can predict mental responses, such as experiences of users who are exposed to visual UIs. In the appraisal theory, emotion is considered to be an evaluative cognitive process, wherein information from several sources contributes to affective judgements. The predictive brain theory proceeds from the concept of the human brain as a prediction machine, with perception and experience resulting from integration of top-down and bottom-up sources of information.

Here, we contextualise the appraisal-theory-based understanding of visual user experience (Jokinen et al., 2015; Silvennoinen and Jokinen, 2016a,b) with the predictive brain theory. This is done by showing that they are similar enough in their perspective on information processing that common hypotheses about affective experiences can be derived. That leads to empirically testable predictions, related to such factors as processing fluency, familiarity, and associations between UI elements and visual user experience. The results support the notion that user experience is a complex phenomenon that cannot be described with any one simple theory or captured with a single methodological approach. Firstly, there are often no clear one-to-one mappings between visual experiences and design choices. The experiment we report upon here demonstrates that such connections are heavily confounded, as is predicted by the appraisal theory and predictive brain theory. Nevertheless, using the experimental procedure outlined in this paper to control for such confounding factors provides a way to improve evaluation of the visual experience of UIs.

In the experiment presented here, the participants appraised Web pages in terms of cued experiential adjectives, such as 'beautiful', 'modern', and 'professional', chosen for their representation of different experience-related aspects of visual Web site design. The primary finding from the study is that, although the relationship between visual UI elements and experience is complex, dynamic, and difficult to predict, practitioners *can* utilise the methodology described herein to understand how their designs tie in with the experience goals and how changes in those designs might affect the relevant experiences. For instance, the methodology can be used to uncover whether there are salient visual elements associated with given experience goals (at least in comparative studies). With regard to generalisation, however, the results reported here should not be taken as rules or suggestions pertaining to Web site design choices. Instead, our hope is that designers will benefit from our work, using it to improve their understanding of the connection between visual design and user experience as it occurs via the cognitive appraisal process, while also finding that we have offered a fruitful approach for investigating said process in the context of their own designs.

2. RELATING EXPERIENCE AND VISUAL UI DESIGN

Recent attempts at solving the problem of connecting visual UI elements with user experience feature two main approaches, along with an additional one combining the two. In the first, the *objective approach*, screen-based design is utilised for detection of specific bottom-up design factors that influence aesthetic experience (Bauerly and Liu, 2006; Lin et al., 2013). This approach is an attempt to identify visual features, such as symmetry and balance, that consistently seem to contribute to perceived beauty. Such enterprises have a long history (Arnheim, 1974; Gombrich, 1995) and are undertaken also in the context of contemporary HCI research (Kim et al., 2003; Tuch et al., 2012; Ngo and Byrne, 2001).

The second approach to investigating the relationship between design elements and aesthetic appeal focuses on users' perceptions of aesthetics from a top-down perspective (Lavie and Tractinsky, 2004; Moshagen and Thielsch, 2010). This *subjective approach* can be characterised with the saying 'beauty is in the eye of the beholder'. In contrast to the screen-based design approach, the top-down approach often relies on self-reporting methods, such as questionnaires (Seckler et al., 2015). The third approach, combining the bottom-up and top-down approaches, is an *interactionist approach*. It has not been utilised to a great extent in examining the interplay of visual UI elements and experience outcomes

(Seckler et al., 2015). The interactionist perspective on aesthetic experience is based on the view that 'beauty is grounded in the processing experiences of the perceiver that emerge from the interaction of stimulus properties and perceivers' cognitive and affective processes' (Reber et al., 2004, p. 365).

In recent HCI research, the interactionist approach has been utilised for a combination of the objectivist and subjectivist perspectives. For instance, *classical and expressive aesthetics* (Lavie and Tractinsky, 2004) and *VisAWI* (Moshagen and Thielsch, 2010), which can be characterised as widely accepted measures applied in the subjective approach (Seckler et al., 2015), have been examined in terms of correlations with design elements detected in objectivist studies (Michailidou et al., 2008; Altaboli and Lin, 2011; Silvennoinen and Jokinen, 2016b). Another approach to connecting design elements and experience outcomes is *computational aesthetics* (Miniukovich and De Angeli, 2015; Ivory et al., 2001; Reinecke et al., 2013), which can be considered to follow the bottom-up approach in how it addresses detecting visual UI design elements and compositional structures important in affective design.

While these approaches do seem promising in prediction of UI experience, they rarely entail forming hypotheses about the mental processes that produce the experience in combination with the visual stimulus (Silvennoinen and Jokinen, 2016b). Scholars of computational aesthetics, for example, list various measurable visual UI aspects, such as symmetry and visual clutter, and correlate these with subjectively perceived beauty. However, the approach offers no explanation for these correlations. For example, why does visual clutter correlate negatively with perceived beauty? Without a theoretical foundation, correlation-based results more readily fall prey to confounding factors, which a grounding theory would identify and allow the researcher to control methodologically (Seckler et al., 2015). Further, a methodology of assessing experience that does not rely on hypotheses surrounding the mental processes that produce the experience has difficulties in dealing with *contrafactuals*, statements that are not true but could be. For example, pixel-based computational aesthetic models, automatically predicting certain aesthetic features such as *beauty* from screenshots of UIs (e.g., Miniukovich and De Angeli, 2015), are theoretically unsuitable for answering certain contrafactual questions, such as 'what would the visual appeal of this site be if the user were to have this specific goal?'; 'if the user were an expert with the layout, how would her aesthetic perception of it change?'; and 'how would a highly motivated user perceive the layout as compared to one with no motivation?'. If we are to answer such questions empirically, we need a methodology that connects empirical observations with cognitive hypotheses about the relationships between perception, goals, motivations, knowledge, and emotions.

One of the theories proffered for explaining the correlations observed between visual UI features and user experience is evolutionary aesthetics, which suggests that the adaptive function of the senses can explain why certain visual aspects are found to be pleasing (Hekkert, 2006). This approach can be used to derive design principles, such as 'maximum effect for minimum means', which suggests that visually pleasing patterns are simple but at the same time still reveal information (Hekkert, 2006); 'most advanced yet acceptable' (Hekkert et al., 2003), which points to an optimal combination of novelty and prototypicality; or 'unity in variety' (Post et al., 2016), which encompasses a similar idea of optimal combination, now between simplicity and complexity, thereby suggesting that visual clutter correlates negatively with perceived aesthetics. Evolutionary aesthetics seems to give a partial answer to the call for theory-based explanation of correlations between experience and design. That said, evolutionary explanations are often challenged by the problem of begging the question (Ketelaar and Ellis, 2000): the assumption that humans have adapted to prefer minimal visual stimuli stems from an observation that we do not prefer cluttered visual aesthetics, but this is the very observation for which evolution was appealed to for explanation in the first place.

In addition to evolutionary explanations of what is aesthetically pleasing, one needs psychological explanations that are testable in laboratory environments. The early 'Gestalt laws' are an example of this approach. Working from their assumptions as to how humans perceive patterns, the Gestalt psychologists suggested a set of heuristics, such as the 'law of continuation', which states that the eye has a natural tendency to follow lines or curves in a direction derived from the visual field (Chang et al., 2002). Although such explanations have a psychological basis, they arguably share the problem of evolutionary explanations to at least some extent. An experiment may be able to test and extend Gestalt laws, but their generality renders experimental manipulations and subsequent hypothesis-testing difficult (Luchins, 1951). Furthermore, both Gestalt and evolutionary explanations assume, sometimes explicitly but often implicitly, that our perception of what is beautiful is biologically hard-wired, yet this assumption is difficult to justify on any other than a very general level (e.g., Thanh Vi et al., 2017). Again, the grounding offered may not be very useful in tackling practical design problems.

A recent brain-imaging study (Thanh Vi et al., 2017) seemed to identify distinct and shared neural correlates of perceived usability and perceived aesthetics. The researchers were able to link these evaluations with brain areas associated with separate cognitive functions.

Hence, the method may improve our understanding of the commonalities of neural processes related to these two constructs and aid in connecting particular visual designs with neural activity associated with distinct cognitive functions. On account of the level of analysis, however, the neural correlates of general cognitive functions might not provide us with a sufficiently detailed theoretical or methodological basis for explaining how and why visual experiences of a visual UI design emerge. These considerations again lead one to ask about the role of the user's previous knowledge, goals, motivations, etc. in the emergence of experience as a result of exposure to visual stimuli.

3. VISUAL EXPERIENCE AS A COGNITIVE PROCESS

3.1. Operationalising Visual Experience

Before the designer can even start investigating how to design for a given experience, she needs a list of possible experience goals. The number of visual experience goals that a UI designer might have is large, and it may be difficult or even impossible to find a common denominator for all of these goals. In addition, it is not enough to provide a list of experiences: an operationalisation of how to measure or otherwise assess them is necessary too Bargas-Avila and Hornbæk (2011). Often in the HCI field, the operationalisation of aesthetics is handled on a very abstract level without breaking of aesthetics into subcomponents. In various studies, the participants have been presented with stimuli designed, for instance, to have low or high aesthetic value (Tractinsky et al., 2000), be pleasant or unpleasant (Sondenegger and Sauer, 2010), or be non-appealing or appealing (Thielsch and Hirschfeld, 2012). In such work, the decision on whether a particular stimulus has a high or low level of the aesthetic properties is left to an intuitive understanding, and no connection can be made between the elements of the design and experience responses. Intuitive designs – either of UIs or of evaluation studies – are not necessarily flawed by nature, but it should be possible to at least try to offer more operable definitions that address what visual appeal is and how it is connected to visual user interfaces' design and evaluation.

Various, overlapping concepts have been used to conceptualise and operationalise appealing visual experience. For example, one can extract meaningful dimensions of visual experience with an Osgoodian method, wherein participants report their impressions of stimuli, using Likert or semantic differential scales that feature various adjectives or quality-related statements. The responses are analysed through factor analysis or similar statistical methods, which are aimed at revealing the latent dimensions of affective experience (Osgood et al., 1975). Osgood's original model had three dimensions (evaluation, potency, and activity), but these perhaps are too general to suffice for a comprehensive assessment of visual experience of UIs. Studies more specific to the HCI context that have focused in particular on user experience have found such sets of dimensions of visual appeal as overall impression, beauty, and meaningfulness (Schenkman and Jönsson, 2000); classical aesthetics (represented by the aesthetic, pleasant, clean, clear, and symmetrical) and expressive aesthetics (involving creativity, use of special effects, originality, sophistication, and an element of fascination) (Lavie and Tractinsky, 2004); simplicity, diversity, colour, and craftsmanship (Moshagen and Thielsch, 2010); and axes of beautiful and practical (Saariluoma et al., 2013).

Dimension-based operationalisation of visual aesthetics is an important approach for creating questionnaires that evaluate UIs. Also, experimentation with such questionnaires can lead to improved understanding of what makes UIs visually appealing. However, the methodology itself does not offer any hypotheses as to the factors behind why certain UIs are found appealing. Although a generative causal model connecting UI design to user experience might not be possible in practice, it is important to try to theorise about the causal connections between visual stimuli and aesthetic experience. If one is able to derive testable hypotheses from such theorising, it is reasonable to assume that further work could embark on causally connecting UI design to concrete experiences. Below, we review a possible causal mechanism leading from visual stimuli to visual experiences, and we derive testable hypotheses from this review. With this we intend not to present a causal model tying any visual input to any experience, but, rather, to demonstrate that such a model is theoretically and methodologically plausible.

3.2. Experience As Appraisal

If we are to causally connect subjective experience with objectively perceivable visual properties, a theory of mental mechanisms is needed. The reason for this lies in the vagueness of the concept of subjective experience. Furthermore, the same theory should posit how the mind interacts with the apparent reality. We concur with previous assertions (Brave and Nass, 2003; Demir et al., 2009; Jokinen, 2015; Mahlke and Minge, 2008) that one tenable psychological theory responding to this need is the *appraisal theory*, which describes emotion as a process wherein the subjective significance and coping potential represented by an event are appraised (Frijda, 1988; Folkman and Lazarus, 1985; Scherer, 2009; Moors et al., 2013; Ellsworth, 2013). In effect, this is a theoretical formulation of the idea that emotional response depends not only on the perceived stimulus but also on the

subject who encounters it and on the circumstances of the encounter. In other words, we usually encounter the world with our existing goals, motivations, and knowledge, and we use existing appraisal patterns to evaluate the events in our environment. Moreover, as emotion is a process, the emotional responses to an encounter may change as the situation progresses, either within our cognitive system or beyond it.

The appraisal theory has been successfully used in exploration of emotional responses to technology use generally (Demir et al., 2009; Jokinen, 2015) and also specifically in the context of aesthetic experience (Jokinen, 2015; Jokinen et al., 2015; Silvennoinen and Jokinen, 2016a,b). The key benefit of using the appraisal theory in this context is that it provides explanations. For example, in a study examining the effect of design era and the processing fluency connected with icons on how they would be perceived aesthetically, the appraisal theory was used to predict and explain how long it takes to appraise such stimuli and how these appraisal times correlate with subjective preference (Silvennoinen and Jokinen, 2016a). In another study, the appraisal theory was applied to predict and explain the influence of individual-specific coping skills on emotional responses in computer use (Jokinen, 2015).

The appraisal theory's model of emotion as a process has practical relevance for HCI research and design. Appraisal progresses in two main stages. In *primary appraisal*, the subject (e.g., a user) evaluates the relevance of a situation or stimulus with regard to her personal goals and values. In *secondary appraisal*, the subject evaluates her ability to cope with the situation or her emotional responses to it (Folkman and Lazarus, 1985; Jokinen, 2015). These two stages of the emotion process are related to different aspects of visual designs. For example, the user's goal may be to add a product to the online 'shopping cart' of a Web shop. If an error occurs, the user evaluates this event as goal-incongruent. Furthermore, the user may evaluate her ability to cope with the situation (that is, to recover from the error) as low (perhaps because she is inexperienced in this respect), which results in distress or frustration (Jokinen, 2015).

Information that is evaluated in appraisal can be analysed as coming from three, separate sources: perceptual stimuli, association, and reasoning (Smith and Kirby, 2001; Jokinen et al., 2015). These might tie in with different experience goals in design. For example, a *light* design may be processed differently than a *civilised* design, because the information requirements associated with these appraisals clearly differ. Appraising something as *light* is probably related more closely to perceptual processing, whereas an appraisal of it as *civilised* should entail association and reasoning. Examining the hypothesis that these appraisals differ in nature was one of the goals behind this paper.

The main contribution of the appraisal theory for addressing the problem of the relationship between experience and design is that it predicts how the subjective experience arises from the process in which the design is evaluated. This allows researchers to design experiments wherein the details of this process can be manipulated for purposes of revealing the relationship between design and experience. With this paper, we demonstrate how the appraisal theory can connect visual experience with interface design, along with how this connection can be exploited experimentally. The principal insight gained from the appraisal theory is that different experience goals involve different mental processes and that, therefore, to target certain experiences, the designer needs to understand the related appraisal processes.

3.3. Predictive Processing and Visual Attention

The three appraisal sources are not distinct and separate. They are interwoven in a complex process. During appraisal, information from bottom-up perceptual stimuli is integrated with top-down associative information and reasoning. This interplay involving two types of information sources, bottom-up and top-down, can be clarified via the *predictive brain theory*. Recent developments in cognitive neuroscience and cognitive science have led to significant empirical evidence and corresponding unified theories of cognition suggesting that brains are a highly developed hierarchical prediction engine (Clark, 2013).

The function of this machine (i.e., the brain) is to support perception and action by continuously matching incoming sensory inputs with top-down expectations or predictions of the same. From the perspective of predictive brains, the target of visual attention is prediction error and the target of attention is selected by modulating the gain (precision weighting) of the error signal (Clark, 2013). Only the sensory signals that represent error in the top-down predictions that produce expectations of what will be seen and, secondly, the associated weight propagate upward in the hierarchy of prediction models for an event and thereby get processed at the higher levels in the hierarchy. This kind of predictive information-processing approach saves bandwidth and enables, for instance, efficient multitasking, ability to function in suboptimal conditions, and high plasticity for learning prediction models that are more accurate for dealing with the causal relationships in our body and the environment.

The theory applies also to an experimental situation with visual stimuli; a participant learns to expect certain tasks during the experiment. With task repetition, the participant also learns (at least) the task-relevant features

of the visual stimuli and can rely more and more on internal representations (i.e., top-down models) of the stimuli instead of active visual search or scene-perception processes. This should become visible during an experiment in a decreasing number of fixations on the visual stimuli and decreased scanpath lengths after repeated exposure. It should be noted that the participant does not have to be aware of this progress (Clark, 2013). In an additional factor, fixation on a visual element does not necessarily mean that the observer's attention is assigned to that visual element; there can occur endogenous mental processing of another target at the time of the fixation (Irwin, 2004). With predictive brains, this means that the visual error signal available is not given weight (i.e., attended to). That complicates the analysis of fixations as indicative of the targets of visual attention.

Nevertheless, the theory of the predictive brain provides clear predictions for an experiment with visual stimuli. While for the first tasks the participant's gaze patterns may be dominated by bottom-up effects, such as the salient features of the (yet unfamiliar) stimuli (Itti et al., 1998) and involve lengthy exploratory gaze paths, increasing repetition can be expected to be accompanied by a decrease in the number of fixations and in the lengths of the gaze paths. Furthermore, the fixations should become more focused on task-relevant elements of the visual stimuli (given that such elements are recognised) (Goldberg and Kotval, 1999).

3.4. The Cued Comparisons Method

For studying how visual experience can be understood as an appraisal process wherein bottom-up and top-down sources of information dynamically influence the subjective experience, we needed an experimental setup that requires participants to appraise how well various visual stimuli served visual experience goals. We chose a method of cued comparison for this purpose (Jokinen et al., 2015). In it, participants are required to quickly choose between two stimulus images once a cue, such as an experience goal, was supplied. The cue can be given via any of several modalities. We opted for words. The data we collected are participants' preferences between the stimuli for the individual cues and the reaction time (RT) required for the participants to arrive at preferential judgements.

The benefit of the cued comparison method over a non-comparison method wherein the participants judge whether, or how much, a stimulus corresponds to the cue supplied is that the comparison task makes the appraisal relative. The participant has to choose between two stimuli, which means not being able to simply agree or disagree with statements. Subjects are forced to make a judgement. This helps to 'calibrate' the cues to the stimuli, supporting appraisal that is sensitive to the cue irrespective of the stimuli used (of course, if the results indicate that the preferences were random, the interpretation following from this is that the cues were not able to distinguish between the stimuli). Additionally, the method connects the concrete process of preferring one stimulus over another with the theoretical appraisal process connected with emotion, as it requires the participants to make their choice as quickly as possible. This is evident in, for example, the ability of the method to capture the effect of differences in appraisal levels on the time it takes to arrive at the preferential judgement. The method thus allows investigation of the cognitive process that ultimately results in a conscious decision that the participant prefers one stimulus over the other.

A clear limitation of the comparison method is that all results are relative, so any absolute statements about the experiential qualities of the stimuli are hypothetical at best. For example, if a certain stimulus is always preferred over all other stimuli in the experiment when they are considered in relation to the cue 'beautiful', one cannot conclude that the stimulus is, in fact, beautiful; one can state only that it is judged more beautiful than the others. However, it should be noted that previous studies using the comparative method have found high correlations (Pearson rs between .80 and .82) between appraisals from comparisons and appraisals made in isolation by means of a pen-and-paper semantic differential questionnaire Jokinen et al. (2015). The advantage conferred by the comparison method and the reason we have used it instead of a pen-and-paper questionnaires, is twofold. Firstly, the automated procedure allows collection of preferential judgements in large numbers, along with the associated RTs. A pen-and-paper method is slower, and it does not lend itself easily to analysis of processing times (even if the participants have been encouraged to respond as quickly as possible), which in our case is important for the analysis of the mental processes behind the appraisals. Secondly, while a questionnaire that involves appraising a single stimulus at a time could be implemented electronically, it can be argued that the comparison causes the participants to appraise the stimuli more robustly. This is because the stimuli are now appraised relative to each other, and a preferential choice is forced. In contrast, presenting one stimulus at a time permits the existence of little variance in the responses. That is, the comparison method guarantees elicitation of preferences (if we assume that the participant does not 'cheat' and make random choices), whereas single-stimulus ratings might result in very similar (e.g., consistently very low or very high) scores across all stimuli for all cues. Finally, questionnaires are not immune to the relativity issues found in the cued comparison method, wherein the participants 'calibrate' their responses to the

overall quality of the stimuli (Sudman et al., 1996). An absolute score on a given rating scale might, accordingly, depend on the set of stimuli appraised. We can conclude, then, that the method presented here is useful specifically for studying the process of appraisal, because of the comparative nature of the task, though simultaneously the interpretation of the results must be sensitive to the fact that the appraisals are relative.

3.5. The Hypotheses for the Research

To enable investigating the application of appraisal theory and thereby improving our ability to connect design and experience, the following hypotheses were derived from the appraisal theory and the predictive brain theory in the methodological context of primed product comparisons. We use the term 'research hypothesis' below to maintain a distinction from a test hypothesis. Each research hypothesis involves multiple statistical hypothesis tests. These more detailed tests will be elaborated upon after the description of the experiment has been provided.

RH1. The speed of the appraisal correlates with the amount by which the object of the appraisal is preferred.

This research hypothesis is centred on the claim, articulated several times above, that more easily processed stimuli are also more preferred. The cued comparison method has already yielded evidence to support this hypothesis (Silvennoinen and Jokinen, 2016a), but the observations in question have not been framed in relation to experience goals. The relevance of this examination for a practitioner is in revealing whether the rule by which preference is correlated with fast information processing is universal from one experience goal to the next.

RH2. Previous encounters with a stimulus change the dynamics of how that stimulus is appraised.

This research hypothesis ties in with the foregoing discussion about the top-down and bottom-up influences in the appraisal process. The clearest test hypothesis implied here is that as the experiment progresses and the participants become familiar with the stimuli and the cues, their RTs get shorter. Similar tests have been conducted in respect of eye-tracking measurements, for inspecting how experiment progress affects fixation count and duration.

RH3. Different areas of interest in the stimuli are associated with different cues, and the associations change, depending on previous encounters.

This two-part research hypothesis suggests that different cues, such as experience targets, are associated with different visual elements of stimuli. Furthermore, this association, in line with **RH2**, should depend on the change in the top-down and bottom-up dynamics.

4. METHOD

4.1. The Participants and Stimuli

The participants ($N = 40$) were recruited via a mailing list for people who are interested in participating in scientific experiments. Everyone with normal or corrected-to-normal vision who responded and expressed interest in taking part was included. The call for participants was up for several weeks, until the required pool of 40 subjects was achieved. The participants' mean age was 30 years ($sd = 9.5$, age range 18–65), and half of them were women. Analysis of the background demographic variables in connection with the research hypotheses was not performed, for reason of the small number of participants in any given age band or gender group. The study was run in a quiet laboratory environment with controlled lighting.

The stimulus context was chosen to be Web pages, as this is a familiar context for almost everyone and therefore allows the expectation that all participants could make judgements as to their perceived aesthetics. The stimuli (see Figure 1) were selected from the CSS Zen Garden Web site (`http://www.csszengarden.com/`), which lists multiple templates that differ in visual appearance but present the same textual content. The idea was that this would let the participants focus on visual experience and not the content of the Web sites. The pages chosen were selected to present the same theme (flowers), for less unpredictability in the variables. All participants were to make judgements of Web pages that shared a common theme. Selection of the Web pages involved finding all flower-themed CSS Web pages firstly. Through an iterative selection process, seven Web pages were selected as stimuli, with the same theme and content but with the visual appearance varying. The seven pages chosen differ in their colour schemes; the concreteness of the flowers displayed; layout structure; typography; and overall use of space – including white space, alignment, etc. It should be noted, however, that, although the sites themselves contained no explicit context, there is always unavoidable contextual information that informs people's appraisal of stimuli. Here, one source of context was the pictures of flowers involved in all stimuli. Hence, the comparison of any two stimuli was always conducted in the context established by this commonality, alongside (probably with a stronger effect) the cue for which the comparison was made.

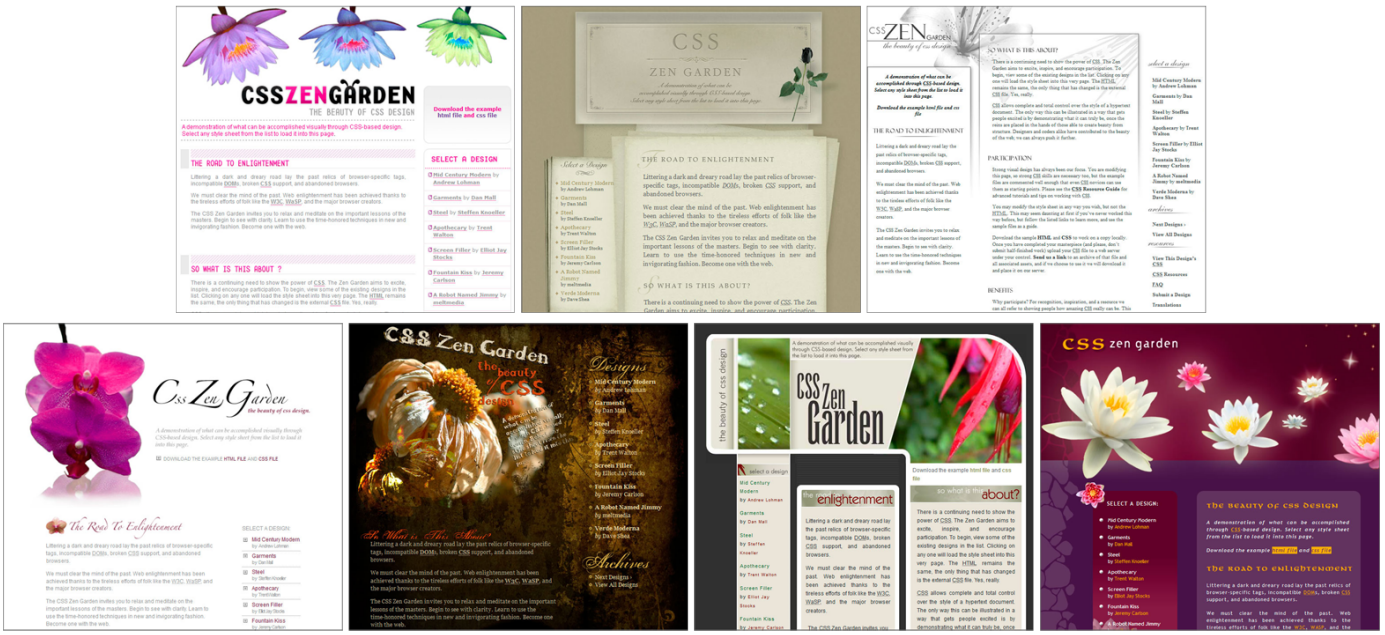

Figure 1. The stimuli in the experiment were screenshots from the Web site CSS Zen Garden. The participant's task was to make pairwise comparisons between stimuli, given a specified experience target.

The participants evaluated the stimuli with regard to the 25 adjectives listed in Table 1. The adjectives were hand-picked from prior research on affective evaluations of various stimuli, with a focus on Web sites (e.g., Jokinen et al., 2015; Lavie and Tractinsky, 2004; Saariluoma et al., 2013). The adjectives were presented to the participants in their native language (*anonymised*). We assumed that all participants had familiarity with the adjectives, and we did not explain their meaning. This does mean that participants may have differed in their interpretations for the adjectives within the context of the Web sites. For example, the word 'light' (the sense of 'not heavy', conveyed in the original language of the study <*anonymised*>, makes this definition explicit) could be related to how quickly the site is assumed to load in a browser or, alternatively, to how complex it is in appearance (these two interpretations are, of course, related to each other). We do not go into detailed analysis of the meaning of the cues here, as the goal is not to provide an evaluation of the Web sites used in the study but to explore the mental processes that can be revealed through their evaluation.

Table 1. The adjectives used in the experiment.

aesthetic	beautiful
cheap	civilised
clean	clear
consistent	courageous
creative	deep
dynamic	exciting
familiar	fascinating
heavy	intelligent
inventive	manageable
modern	old-fashioned
pleasant	professional
sophisticated	symmetrical
technical	

4.2. Procedure

At the beginning of the experiment, an informed consent form was presented, discussed, and signed. The participants were informed that they could withdraw from the experiment at any time without completing it, with no need to disclose the reason for doing so; however, no-one left before the experiment was finished. In all, the experiment took each participant between 50 and 60 minutes, including the experiment proper, breaks, and preparations. Before the collection of data began, the participants were shown screenshots of two CSS Zen Garden designs that were not used in the actual experiment and did not contain pictures of flowers. The participants were told that the experiment would consist of them looking at similar Web sites. They were encouraged to read the text in both of the images and

confirm that they had the same content. They were told in addition that the textual content of the experimental stimuli would always be the same as in these samples and that, therefore, they should not focus on reading the text in the screenshots during the experiment.

In the experiment, the participants were shown combinations of adjectives with stimulus pairs. A trial consisted of *a cue*, which was one of the 25 adjectives listed in Table 1, and *a pair of stimulus images*, a combination consisting of two of the seven screenshots from the CSS Zen Garden site (see Figure 1). The total number of pairwise combinations from 7 is 21, so the total number of trials was $25 \times 21 = 525$. Pilot studies with various numbers of stimuli and cues were run to find a combination that would not take too long yet would still provide enough material for the analyses. The participants were shown all of the 525 adjective–stimulus-pair combinations sequentially in a fully randomised order of all possible adjective–pair combinations. Their task was to select whichever of the two stimuli on the screen they preferred, given the cue adjective. Firstly, the cue was shown on a computer screen for 2 s, after which it was replaced with the stimulus pair. The order of the two stimuli (that is, which was on the left and which on the right) was randomised. The participants were asked to indicate their preference as quickly as possible, using a two-button RT switch in front of them (pressing the left button for the left-hand stimulus or the right button for the right-hand one). For example, if the cue was 'beautiful', the participant pressed the button on the side of the stimulus that he or she appraised as more beautiful than the one it was paired with. A new cue appeared as soon as the RT button was pressed, except for the two rest periods balanced within the experiment. These breaks effectively split the experiment into three blocks, but here we analyse all data together (having ensured that there were no noteworthy differences between blocks). All within-subject counter-balancing of stimulus pairs and cues in the experiment was done purely randomly. Because of the high quantity of tasks and the relatively large number of participants, we expected that there would be no clear bias toward certain stimulus pairs or cues in terms of ordering. After collecting the data, we made sure that the expected randomness had been achieved, by checking that all cues and stimulus pairs had roughly even distributions between individual phases of the experiment. Figure 2 illustrates the flow of a trial. It presents a photograph of the setup with two stimuli shown (the cue is not visible here, because it is displayed only before the stimulus pair).

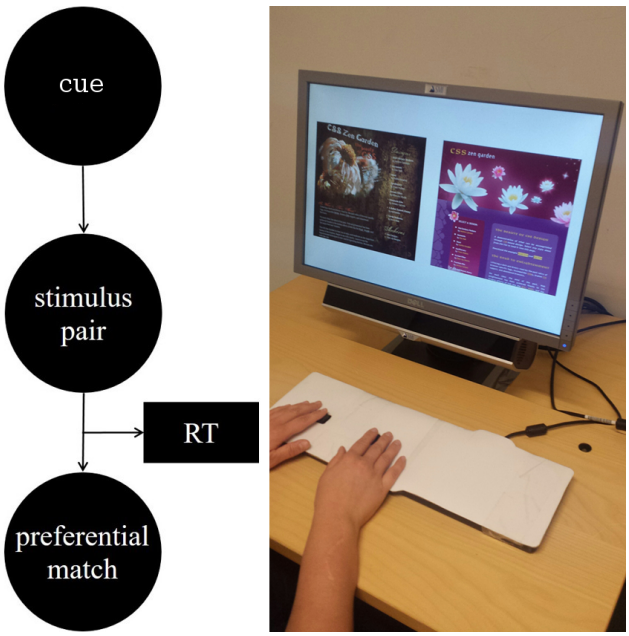

Figure 2. The experiment setup. The participant, who is seated in front of a computer, is shown cues and pairs of Web pages. The task is to indicate his or her preference as quickly as possible, using a two-button RT switch.

4.3. Collection and Analysis of the Data

The two main sets of data[1] generated in the cued comparison task are RTs and preference scores (PSs) (Jokinen et al., 2015). The *reaction time* values simply indicate the time it took for the participant to indicate the preference by using the RT switch, from the time the stimulus pair appeared (that is, not counting the time for which the cue was visible). By always having to choose between two stimuli with regard to a given cue, the participants provided information for quantitatively ranking the stimuli on the basis of the cues. This quantitative ranking, called the *preference score*, indicates the probability of a stimulus being preferred over all other stimuli in the experiment with the given cue. Therefore, there is a PS for each stimulus–cue pair (e.g., there are 25 PSs for stimulus 1, etc.). From the PS of a stimulus for a given cue, one obtains the average probability of it being chosen over any other stimulus with this cue. A PS of 1.0 for the stimulus with the given cue means that this stimulus was always preferred over the other stimulus with regard to that cue. A PS of 0.5 means that half of the time this stimulus was preferred over the other one for the relevant cue.

Eye tracking was conducted with an SMI RED remote binocular eye-tracking system at 500 Hz sampling rate.

[1] All data and the R code used in the data analysis are available at https://userinterfaces.aalto.fi/emotion

The eye-tracker was attached below a Dell 22" stimulus display that displayed the stimuli at 1680 × 1050 pixel screen resolution. The size of all stimuli was 730 × 660 pixels. For all participants, the distance from the display was kept between 65 and 70 centimetres. The calibration deviation for both left and right eye in the x and the y direction was less than .67 degrees for all participants. A fixation was defined as very low-velocity eye movement corresponding to the participant staring at a particular point within a stimulus, in line with the definition provided by the eye-tracking system manufacturer. A fixation is recorded by the SMI BeGaze 3.6 software through application of a dispersion-threshold identification algorithm (Salvucci and Goldberg, 2000) with two parameters, a maximum movement dispersion threshold and a minimum duration threshold. The software's default thresholds of 100 pixels and 80 ms were used for recording the fixations.

The three research hypotheses for this study were analysed via various statistical techniques. The first hypothesis, pertaining to the connection between the speed of appraisal and the amount of preference, was addressed by correlating RTs and PSs individually for each of the 25 cues. Each cue had seven PSs, one for each stimulus. Corresponding RTs were calculated as an average across the decisions wherein the cue in question was considered with a given stimulus present. Here, negative correlations would mean that the more preferred a stimulus was for a cue, the less time the appraisal took. Accordingly, high negative correlations are taken to support **RH1**. However, we also assume the magnitude of these correlations to vary, and we expect their comparison to reveal details about the hypothesised mechanism.

The second research hypothesis, to do with the effect of exposure on task speed (i.e., RT) and eye movements, was considered by means of three, quite different multilevel models. The dependent variables for the three models were RT, number of fixations, and average fixation duration. Support for **RH2** would be indicated by task progress (trial number) having an impact on these variables. Further control variables were included in the model, such as RT for predicting the number of fixations. Inclusion of RT as a control variable in the model enabled analysis of the influence of the other independent variables while the effects of changes in RT were controlled for. This is because the number of fixations probably correlates with how long the comparison took, which again might depend on the number of trials completed. Therefore, adding the number of fixations to the model as a control lets us inspect the correlation between trial number and the number of fixations, without these different correlations confounding each other in the model. Finally, participant was entered as a random effect in the model; thereby, the estimates were adjusted for the fact that there were multiple observations for each participant (Hox, 2010).

Also, we examined the possible confounding effect of fatigue on RTs. It is possible that the participants grew tired of the experiment as it progressed and started to choose stimuli more randomly, thereby producing shorter RTs. We checked for this in two ways. Firstly, we tested for dominance of either hand, on the assumption that participants who were cheating would have used just one hand and, in essence, chosen at random. The data showed no evidence of such cheating, either on aggregate level or within individual phases of the experiment. Secondly, we binned the experiment into 10 segments, calculated PSs separately for each of those segments, calculated the absolute deviation of PS from the random expected value of 0.5, and correlated that value with the segment. If the participants had started choosing at random, the segment number should correlate negatively with PS values. The data showed no evidence of such a mechanism either, with $r = -.04$ between segment number and PS, and the boxplots of PSs were practically identical across the 10 segments by visual inspection (in addition, no correlation could be observed between segment number and mean PS; $r = -.02$). The average standard deviations of both deviation and grand mean PSs within the 10 experiment segments did not change noticeably and did not follow any of the correlative patterns tested for. This suggests that, although the experiment was demanding, requiring the participants to make more than 500 comparisons, the participants did not suffer from fatigue so greatly that this would have affected their decision-making.

The third and final research hypothesis, pertaining to changes in the visual search patterns due to differences between adjectives (user experience targets) was studied through a metric capturing scanpath similarity. Whereas simple fixation heatmaps would provide a possible qualitative means for comparing the areas of Web sites that are investigated with regard to a given experience goal, this route of analysis would not produce a satisfactory answer for the research hypothesis. Firstly, visual inspections of heatmaps can yield wildly different interpretations, which depend on the interpreter. Creating fixed areas of interest (AOIs) and comparing fixation statistics between these would be less susceptible to subjectivity of interpretations, but identifying the AOIs in the first place would require subjective judgements by the researchers. Secondly, as posited in **RH2**, the scanpaths change throughout the experiment as the participants gain exposure to the stimuli.

To enable objective comparison addressing the effect of cue on how the stimuli are inspected, a scanpath similarity metric was calculated for each stimulus in each trial. Each trial was split into two segments (because there were two stimuli per trial) and considered separately. The

scanpath similarity metric chosen, referred to as *scasim* (Von der Malsburg and Vasishth, 2011), was used to create a large similarity matrix covering all scanpaths for a given stimulus; accordingly, seven distinct scanpath similarity matrices were generated. The reason for splitting between stimuli was that, since the position of the key visual elements differed between the various Web pages, it makes sense to analyse the effect of a particular experience goal (a given cue) on scanpaths only with regard to a given stimulus. After that, a two-dimensional map was created from the similarity measurements, through multidimensional scaling (MDS). Individual scanpaths on the 2D map were then classified via mixture modelling, which does not require the researcher to dictate the number of classes. The *scasim* calculation is a technical procedure, and the reader is referred to the original source for details (Von der Malsburg and Vasishth, 2011). Non-technically, it can be interpreted as giving a number that represents how similar two scanpaths seem: if the ordered fixations appear to travel along the same route, they are more similar than are two scanpaths that look vastly different from each other. This measure is not unlike *Levenshtein distance*, which represents how much two text strings differ from each other in terms of the minimum number of edit actions needed for rendering them identical (Von der Malsburg and Vasishth, 2011).

The reason for using multinomial regressions for scanpath classes instead of simply plotting gaze heatmaps across the cues for a particular stimulus is the theoretically complex nature of appraisal. The analysis conducted goes beyond visual inspection of heatmaps: although some heatmaps will be provided in the results section, their interpretation requires the results from the regressions. Further, the results of the multinomial regression runs for the clusters reveal qualitative differences in how the stimuli are appraised with regard to different cues. Support for **RH3** is provided by the observation that stimuli, cues, and experiment progression affect the overall scanpath patterns.

5. RESULTS

The grand mean RT of all appraisals, by all participants, was 2.0 s (sd = 0.9). This means that, at least for the most part, the participants complied with the request to refrain from thoroughly inspecting the Web sites (e.g., they were asked not to read the textual content of the pictures), and they indeed made their appraisals on the basis of first impressions. Also, inspection of the results for the first few trials revealed that the participants were able to make immediate appraisals from the very earliest trial onward. Mean RTs and their sd values were no larger for the few first trials than for the rest of them. The number of individual PSs is large ($25 \times 7 = 175$), so they are not all presented here; moreover, as the stimuli were mock-up Web pages, the results for individual sites were not of particular interest. Nevertheless, for illustration purposes, three stimuli, one making a favourable overall impression, one receiving a clearly theme-based appraisal, and one gaining a generally unfavourable appraisal, are presented here in more detail. The stimulus image in the lower left in Figure 1 was appraised, relative to the various alternatives, as more *professional*, *aesthetic*, *manageable*, *sophisticated*, *beautiful*, *fascinating*, *pleasant*, *modern*, *clean*, *civilised*, and *familiar*. All of these cues had $PS > .70$ for this stimulus (this cutoff point was suggested by Jokinen et al. (2015)). The stimulus in the middle in the upper row in Figure 1 was appraised as *sophisticated*, *civilised*, *old-fashioned*, and *consistent*. Conversely, the second stimulus from the left in the bottom row received generally negative relative appraisals. When compared to its alternatives, it was considered less *professional*, *manageable*, *sophisticated*, *beautiful*, *intelligent*, *pleasant*, *clean*, *clear*, *symmetrical*, and *familiar*, also being deemed more *heavy*.

The correlations between RTs and PSs are shown in Table 2, broken down by cue. Let us consider cases with a high negative correlation, such as that seen with *creative*. When the comparison of two stimuli for their creativeness involved a stimulus that was regarded as creative overall, the judgement took less time than if the stimulus was not regarded as creative in general. In simpler terms, participants quickly 'knew (relative) creativity when they saw it'. Analogous statements can be made for other cue words with high negative $RT \times PS$ correlations ($r < -.50$): 'technical', 'old-fashioned', 'cheap', 'fascinating', 'familiar', and 'beautiful'. Conversely, such cues as 'civilised', 'symmetrical', 'exciting', and 'clear' exhibited positive correlations in general, although these were not as striking as the negative correlations.

The results for the three multilevel models are presented in Table 3. With the first model, predicting changes in RT with trial number, cue, and the interaction between the two, the number of trials completed was statistically significant. Each trial (where the counter started at 1 and ended with 525) saw RT decrease by 3.1 ms; hence, on average, the final trials were 1.6 s faster than the very first trials. This means that the speed-up due to familiarity with the tasks (the cues and stimuli) was considerable. The effect, without the controls included in the multilevel model, is shown in Figure 3 with boxplots. Since the horizontal centre lines in the boxplots represent median values, the result can be considered robust and not susceptible to effects of outliers. Looking at other variables of Model 1 shows that the cue variable had a statistically significant effect: some cues took longer to appraise than others. On account of the large number of

Table 2. Pearson correlations between RTs and PSs for each of the cues.

Cue	Cor.	Cue	Cor.
beautiful	−0.77	aesthetic	−0.21
familiar	−0.71	deep	−0.18
fascinating	−0.70	clean	−0.05
cheap	−0.65	inventive	0.02
creative	−0.65	professional	0.08
old-fashioned	−0.62	consistent	0.09
technical	−0.51	intelligent	0.10
sophisticated	−0.50	manageable	0.14
heavy	−0.44	clear	0.25
courageous	−0.38	exciting	0.28
dynamic	−0.30	symmetrical	0.36
modern	−0.28	civilised	0.44
pleasant	−0.27		

Table 3. The three multilevel models employed in the study, with coefficients displayed for continuous variables (the effect of RT in models 2 and 3 is standardised).

Fixed variable	Estimate	Pr(Wald χ^2)
Model 1 (RT)		
Intercept	2907.7	< 0.001
Trial	−3.1	< 0.001
Cue	–	< 0.001
Trial × cue	–	0.26
Model 2 (total fixation count)		
Intercept	6.6	< 0.001
Trial	−0.002	< 0.001
RT (std)	5.8	< 0.001
Cue	–	0.01
Stimulus pair	–	0.005
Trial × RT	$-3 \cdot 10^{-7}$	< 0.001
Trial × cue	–	0.22
Model 3 (average fixation duration)		
Intercept	216	< 0.001
Trial	0.04	< 0.001
RT (std)	102	< 0.001
Fixations (std)	−108	< 0.001
Cue	–	0.55
Stimulus pair	–	0.01
Trial × cue	–	0.03

cues, there were no planned *post hoc* tests to examine in detail which cues in particular were connected with fast and which with slow appraisals. Nevertheless, some descriptives can be provided. For example, appraisal times for 'aesthetic' were 480 ms shorter than the grand mean RT. Similar effects were seen for appraisals as *beautiful* (550 ms shorter), *clean* (770 ms shorter), and *consistent* (440 ms longer).

The first model demonstrates the decrease in RTs that occurred as the experiment progressed. The phenomenon is visible also in the total number of fixations within a trial falling: in the first 100 trials in the full set of 525, the average number of fixations on a stimulus pair was 8.4, whereas for the final 100 trials it was 4.8. The grand mean for fixations was 6 ($sd = 3$). The second model presented in Table 3 demonstrates this effect too but in more detail: the total number of fixations on a stimulus pair decreased by 5.7 for each 1 *sd* change in RTs. The remaining effect of trial on total number of fixations is small but still statistically significant, showing that when one controls for the confounding effect of RT, there were 0.8 fixations fewer toward the end of the experiment. Finally, even with controlling for the confounding effect of cue on RT (Model 1), different cues prompted a different number of fixations from the participants. These effects were small, though: even at maximum, as with 'exciting', the average decrease in total number of fixations was 0.2, and the increase for 'consistent' was 0.3.

The third model described in Table 3 predicts changes in average fixation duration. Grand mean fixation duration was 221 ms ($sd = 53$). The average fixation duration early in the experiment, and in a trial with average RT, was 216 ms. Each RT increase of 1 *sd* raised the fixation duration by 102 ms, when trial number is controlled for. Furthermore, fixations during the final trial were 21 ms longer on average than fixations made in the beginning. In addition, the number of fixations was related to fixation durations: when the number of fixations rose by 1 *sd*, average fixation duration decreased by 108 ms.

The scanpath similarity, or *scasim*, analysis, conducted separately for each of the individual stimuli from the fixation data, yielded 6–8 classes, with the number depending on the stimulus. Using multinomial regression – again separately for the various stimuli – we used trial and cue to predict scanpath classes. A separate model run was performed for each of the seven stimuli, with the *scasim* class as the dependent variable and trial number and cue as independent variables. Trial number was a statistically significant predictor for all seven stimuli, while cue was for three of them (those in the top row in Figure 1).

Description of individual classes for all stimuli is omitted for reasons of space and also because the prototypical nature of the stimuli would render it

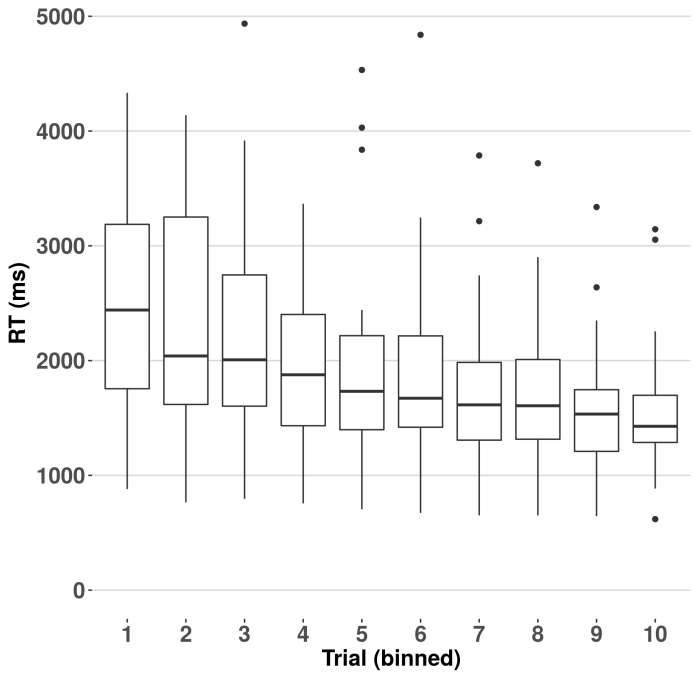

Figure 3. The effect of experiment event on RTs: as the experiment progressed, RTs became lower. The 525 events have been grouped into 10 bins, and the RTs are aggregated within the bins separately for each participant. The y-axis has been limited to 5000 ms, leaving a small number of outliers (the isolated black dots) at the periphery in bins 1–4. These results are for illustration purposes – for more detailed statistics, please refer to the estimates for the models provided in Table 3.

not so informative. Some example descriptions can be given nonetheless, to demonstrate how the *scasim*-based analysis utilised here works. Firstly, the effects analysed with the multilevel models can be investigated further. Figure 4 shows two *scasim* classes, the first of them (see the centre image) more likely to be associated with early trials and the second (see the image on the right) with later trials. This result can be used to develop a richer description of the results produced by the multilevel models, predicting that fewer fixations accompany later trials. The finding is clearly that the fixations decrease in number and converge under a smaller area rather than merely decreasing in number without converging. This gives evidence of the hypothesised top-down appraisal process that directs fixations to areas that offer the greatest efficiency for making the appraisal.

Further analysis can be conducted with respect to how the cues are distinguished between *scasim* classes. Figure 5 shows scanpath heatmaps from three of the seven *scasim* classes for one of the three stimuli with which the cue was predictive of the *scasim* class. If we compare the odds ratios between the first two of these classes for different adjectives, the large differences can be identified. The scanpath class depicted on the left in Figure 5 was more likely to be encountered when the cue was 'manageable' (1.8 times as probable for this class than for the one in the centre), 'inventive' (1.9), or 'pleasant' (2.0). Conversely, the class depicted in the centre was more probable with 'creative' (2.4), 'clean' (1.9), and 'civilised' (1.7). The scanpath results in Figure 5 do not easily lend themselves to interpretation. However, the goal here is just to show that such distinctions can be made. A designer who wishes to compare design ideas, for example, might use the method demonstrated here for better understanding which areas of the designs correspond to which experience goals.

Another way to use the classes is to consider a single cue and investigate the classes that were probable and improbable with it. For example, appraisals as *beautiful* had the greatest likelihood with the class that is depicted in the third (rightmost) pane of Figure 5. When this pane is compared with the scanpaths next to it, for two other classes, the scanpaths of this third class are relatively concentrated. In another example, Figure 6 illustrates the effect of two cues, in relation to which the stimulus in question was preferred, on the *scasim* class and subsequently on fixation heatmaps. The *scasim* class for which 'sophisticated' was more probable (2.6) is shown in the centre. It reflects scattering of fixations that is centred on the middle of the stimulus. Conversely, the final image in the figure depicts the scanpaths likely found in conditions of appraisal for 'civilised' (4.8) as clearly directed in a more focused manner toward the top of the Web page's left panel. Again, such comparison might assist a designer who wants to understand which areas of her visual designs might correspond to which experience goals. Again, we will refrain from interpreting the results for the designs themselves, which are not subject to inspection or development.

6. DISCUSSION

The main finding from the inspection of correlations between RTs and PSs (see Table 2) is that Web pages deemed creative, familiar, old-fashioned, technical, fascinating, cheap, or beautiful in comparison to other Web pages are quite quickly appraised to be so (indicated by a high negative correlations between preference and RT). This could mean that if a designer targets these experiences, establishing the targeted experience swiftly in the mind of the user is more important than with such aesthetic qualities as being civilised and exciting. Conversely, if the designer targets a 'civilised experience', there is more time to impress the user. As the user

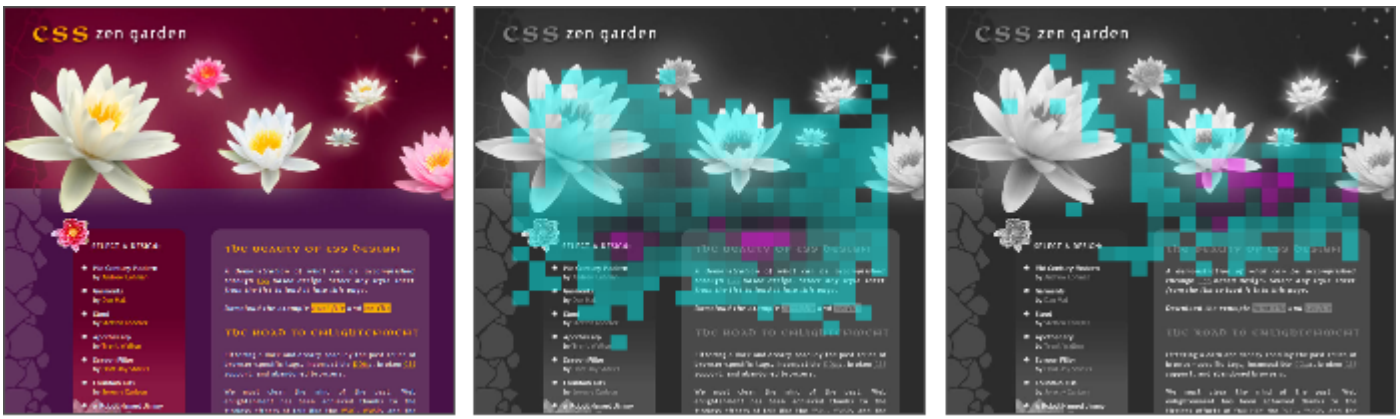

Figure 4. An example of the effect of trial on *scasim* classes. The figure presents heatmap aggregation of scanpaths for two classes for a single stimulus. At the left is the pure screenshot of the stimulus, and the other two images are heatmaps for scanpath aggregates for the respective *scasim* classes. As trial number increases, the probability of the scanpath belonging to the class represented on the right increases. The base image has been desaturated so that the scanpaths are more clearly visible.

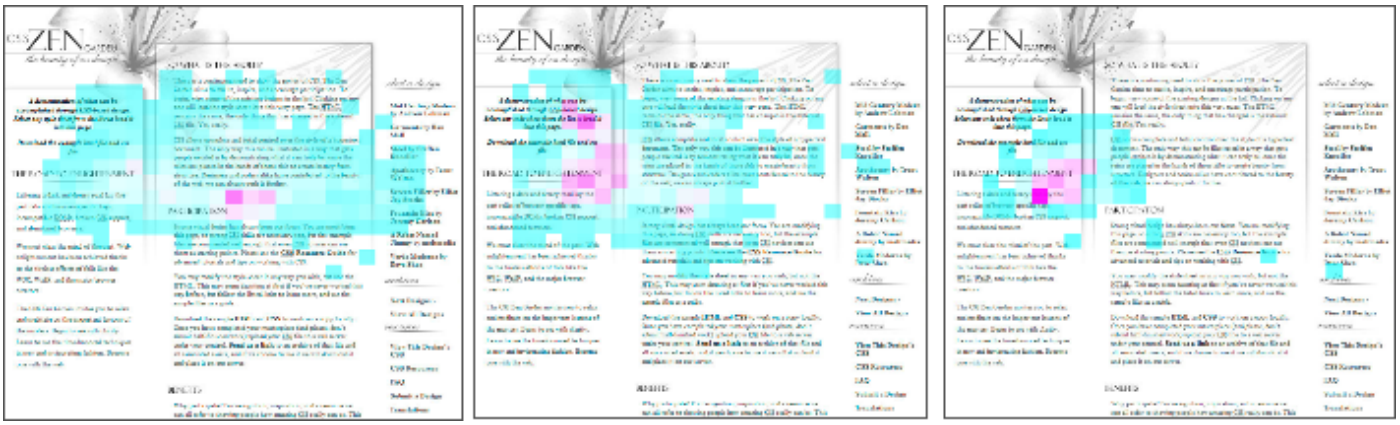

Figure 5. An example of the effect of cue on scanpaths. The image shows heatmaps aggregating scanpaths for three *scasim* classes for the given stimulus. For the leftmost class, probable cues were 'manageable', 'inventive', and 'pleasant'. The probable cues for the class in the second image were 'creative', 'clean', and 'civilised'. The rightmost class was likely to be encountered when 'beautiful' was the cue.

devotes time to these appraisals, more information input to the appraisal is collected via the senses and mental processes, such as associative memory and reasoning. More information does not mean more valid appraisals, but the results do suggest that there is a threshold of 'enough information' for which the participants wait before feeling confident enough to make the preferential decision.

It should be noted that, because all of the appraisal tasks were comparative, the methodology applied does not lend itself to absolute statements: it is possible that comparison involves an appraisal process that is vastly different from that entailed by a single stimulus. Also, previous studies using the same comparison method showed very high correlations between preference scores calculated from the comparisons and absolute scores arrived at in isolation via a questionnaire (Jokinen et al., 2015). In the interest of time, we did not add a pen-and-paper questionnaire (e.g., a sematic differential) to the apparatus, to check whether it produced similar appraisals of the stimuli. Any practitioners utilising the method are encouraged to perform an initial check that their design domain produces such agreement.

The finding related to fast appraisal of beauty when the stimulus is regarded as relatively beautiful is in line with an earlier finding that people can appraise beauty very quickly (Lindgaard et al., 2006). The reasons for this are not necessarily clear: maybe we are just accustomed to making appraisals of beauty, whereas, for instance, we are less used to appraising such qualities as being civilised. Also, the correlations probably depend somewhat on the stimuli: the results presented here are

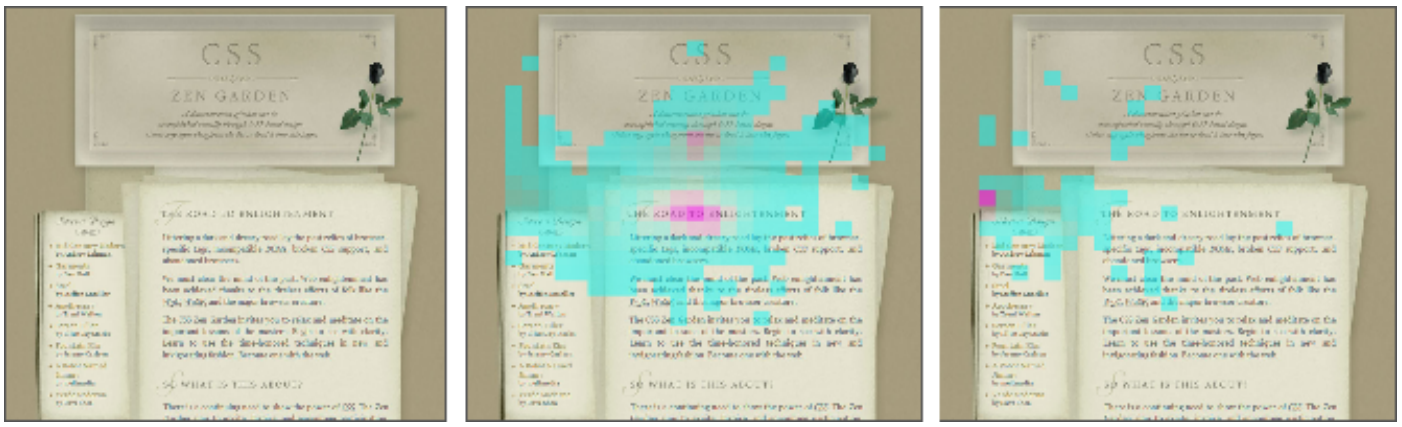

Figure 6. An example of the effect of two preferred cues on scanpaths. The heatmaps aggregate scanpaths for two *scasim* classes for a given stimulus. At the left is the pure stimulus screenshot, and the following panes in the image are heatmaps for scanpath aggregates. The centre image represents a *scasim* class for 'sophisticated' and the rightmost for 'civilised'.

situated in the context of Web sites, and the results might be very different for icons or drinking glasses, for example. Therefore, our general conclusion about the negative correlation between appraisal time and preference in terms of the listed experience goals should be generalisable to comparisons of other Web page stimuli but not necessarily to other visual design contexts. Nevertheless, this result supports **RH1**, the hypothesis that processing fluency and preference are correlated.

The results lend plausibility to the aesthetic-usability effect, a posited phenomenon wherein more usable solutions are perceived as also more aesthetically pleasing (Hassenzahl and Monk, 2010), on the assumption that faster processing is connected in addition to more aesthetically pleasing products, as the relevant information is quickly assimilated and processed (Reber et al., 2004). The negative correlations between task-completion times and preferences (see Table 2) for many cues would support the claims of the aesthetic-usability effect. However, low RTs do not necessarily mean that the stimulus was more aesthetically pleasing. For example, although an *old-fashioned* stimulus is appraised quickly (at least when being compared to other, similar products), this may not be the experience targeted by the Web page's designer (Silvennoinen and Jokinen, 2016a). Cheapness, for instance, is probably not an impression that a Web designer wishes to target. Accordingly, more swiftly and smoothly processed stimuli are not necessarily more usable or aesthetically pleasing – these qualities depend on what is being processed. It would be interesting for a future study to establish a set of very desirable and very undesirable qualities as cues, then use these to investigate correlations between RTs and PSs, preferably in various contexts, which would have a plausible hypothetical effect in terms of the (un)desirability of the qualities. For example, being *delicate* might be more desirable in the context of wine glasses than that of cars. One could then hypothesise that the RT–PS correlations for 'delicate' would differ between these two contexts.

Finally, it is important to note that the confounding effect of fatigue is difficult to manage, which is why the analyses took it into consideration. It was possible that the participants not only became more familiar with the stimuli as the experiment progressed but also grew more fatigued with the procedure, with this leading to more random preference choices and hence faster responses. However, as shown above, no such fatigue effects were observed in the participants' preferences. In addition, the tasks themselves were not difficult to learn. As there was no 'learning overhead' in the first few trials, the participants could immediately start performing the tasks quickly – as they had been asked to do.

Comparing the RTs obtained in this study to others obtained with the same method, one can conclude that relative Web page evaluation (2.0 s) requires slightly more processing time in the mind than does comparative evaluation of icons (1.3 s) (Silvennoinen and Jokinen, 2016b) or drinking glasses (1.8 s) (Jokinen et al., 2015). However, the differences, while clear, are not huge, and in any case the RTs of 1–2 s indicate that people can make relative aesthetic judgements quickly in various design contexts. Both the study reported upon here and earlier work cited above produced clear and face-valid results, and the participants in all three studies can be assumed to have made actual, informed judgements instead of random (i.e., uninformed and senseless) choices. Our study was consistent with the earlier ones in showing that different cues have different processing times, which hints at the information requirements depending on the aesthetic judgement one is making. This finding is difficult

to explain without a cognitive theory of emotion (the appraisal theory), which explicitly states that emotion is a process, in which the mind processes information from multiple sources – and at different speeds – to form a subjective feeling about a situation or an object (Scherer, 2009; Smith and Kirby, 2001).

Our study also supports the predictive processing theory of the human mind (Clark, 2013). Models 2 and 3 (see Table 3) indicate that the total fixation count grew smaller but fixation durations larger as the experiment progressed (when we controlled for changes in RT). The effect was small, but it was statistically significant and can be meaningfully explained. Toward the end of the experiment, less visual information was needed, as the participants relied increasingly on their memory. This can be taken as a top-down predictive process (Clark, 2013) – that is, as the participants learned processing of the stimuli, they needed less external information. The reason for this pattern is that the appraisal patterns in the mind became tuned to process the tasks, and the role of external information lay in making slight corrections to the predictions of this appraisal process. The latter process relies mostly on internal information. The effect manifests itself also in longer fixations.

When we controlled for the effects of trial, an increase in RT was associated with increased average fixation duration. This suggests that the cognitive demands were higher for the high-RT tasks. Increased fixation duration has been linked with increased cognitive processing of a stimulus at the fovea (e.g., a difficult word in reading; see Just and Carpenter, 1980). However, the link between fixation duration and the cognitive processing demands of the given stimulus at the point of fixation is not obvious, since, again, the cognitive processing may be targeted at things that are not in the foveal area and may take place between fixations too (Irwin, 2004). Here, the slightly greater fixation durations toward the end of the experiment can be interpreted to suggest higher levels of top-down cognitive processing instead of increased visual processing of the stimulus at the fovea. An alternative explanation is that with experience the fixations are more efficiently targeted at task-relevant targets and the fixation durations rise simply on account of decreased requirements for shifting the gaze (Goldberg and Kotval, 1999).

However, this increased efficiency in locating the task-relevant targets still presumes a well-developed internal model of the cues and stimuli (i.e., the task). In addition, as shown with Model 2, while larger RT values correlate with larger numbers of fixations, this connection is reduced as the experiment progresses. Accordingly, RT explains the number of fixations less as time elapses in the experiment. The reason for this is that, overall, fixations are required less, as more of the processing time is devoted to endogenous cognitive processes. Hence, the decrease in fixation count results from other factors than merely the RT decreasing. It is not self-evident whether the RT is less because there is no need for as many fixations or, rather, time for that many fixations is not available within the time used. We can nonetheless explain both findings as being caused by the same phenomenon: with accumulation of experience, the participant develops more accurate prediction models (i.e., appraisal patterns) for the cues and stimuli and also for their relations (what to attend in the given cue, and where); therefore, there is less need to find information by consulting the screen.

The scanpath lengths decreased with repetition, in line with the fixation data and our predictions (see Figure 4). Reliance on top-down models of the visual stimuli, which are created through repeated exposure, is evident here for all the stimuli, decreasing the need to perform extensive visual search actions for comparing a pair of Web pages in terms of the given cue. People tend to decrease their scanpath lengths and are able to gather more visual information via peripheral vision with increasing experience of a task (Summala et al., 1996), which is consistent with our findings and theoretical framework.

More support for the suggested explanation of transitioning from bottom-up processing to top-down processing can be found in the Model 1 results. The model's predicted 1.3 s RTs near the end of the experiment already represent very fast responses: in this time, the participant is able to make only three fixations of 200 ms per stimulus. This is difficult to explain otherwise than by concluding that the participants started to rely on a quick top-down process with minimal need for external information. The essential function of this external information is in correction of the existing top-down prediction model, but, as the experiment proved, the participants found little reason for doubting themselves. Throughout its course, the participants' reliance on internal top-down information sources was extensive. Because all trials (tasks) were unique, this is not a *mere exposure effect*, wherein repeated exposure to a stimulus increases the likelihood of positive evaluation (Bornstein, 1989; Zajonc, 1968). The explanation is, hence, more complex: the participants formed a template of how to appraise certain stimuli or certain experiences. These appraisal patterns were transferred between trials. This conclusion is supported by the results for the third research hypothesis, on the scanpath similarities. The appraisal patterns must be associated with information for certain task-relevant features of the stimuli in relation to a cue, as shown in figures 5 and 6.

The results pertaining to how participants adjusted over time, moving from bottom-up- to top-down-oriented appraisals, do not mean that experiments such as the

one we conducted should involve initially training the participants for a long time before assessment of which stimuli are preferred for which cues begins. This is because the same process of encountering a stimulus, later coming to rely more on bottom-up information sources, then with repeated encounters moving to reliance on top-down sources is a very natural phenomenon. This notion has been explored in a less cognitivistic manner through the concept of *user experience over time* (Karapanos, 2013). Designers of any products can expect their designs to be used by people with any of various levels of familiarity and expertise with relevant contexts. As the methodology presented here allows studying the effect of exposure on appraisal, the designer can experimentally manipulate and investigate required phases of interaction.

Finally, the analyses of the *scasim* (Von der Malsburg and Vasishth, 2011) classes indicate that different cues were associated with different types of scanpaths. Although the methodology utilised here was able to establish correlations between experience targets and visual locations of the stimuli, it is beyond the scope of our research to examine particular stimuli and ask why certain experience goals were associated with certain scanpath patterns. To study such effects, one would need to design the stimuli in such a way that the visual elements are controlled. In contrast, the stimuli in the study reported on here were not carefully designed to contain certain visual elements (rather, they were chosen to contain enough variety for eliciting reliable comparative preferences). Nevertheless, the method we used did reveal relations between visual design and experience goals, and future work could simply adopt it as-is while using more carefully designed stimuli. In addition, although the stimuli in this study were Web pages, the approach is easily adopted to other contexts, as long as a pictorial representation of the design artefact can be produced.

An interesting avenue for future study would be to investigate the effect of small incremental design changes. How many small changes are needed before the method presented here identifies the location as relevant for a given experience goal? Expectations of technological artefacts have been shown to affect interaction experiences (Raita and Oulasvirta, 2011); hence, the examination of visual elements affecting experience goals in terms of expectations could be utilised for understanding the experience outcomes in greater depth. This idea is closely related to the predictive processing theory utilised here (Clark, 2013). Additionally, visual elements experienced as visually appealing differ with the Web page genre (Papachristos and Avouris, 2013). The method presented in this paper can be utilised in examination of genre-specific appealing design factors.

In the cued comparison method employed in our work, the cues were textual words. Alternatively, one could investigate the use of other cue modalities, such as sound or animation, or even smell. If the cues were, for instance, sounds or images, the cognitive appraisal process would be different. We assume, however, that written words have more predictability in their meaning than do images or sounds; accordingly, the results with words should be more robust. Nevertheless, it would be interesting to see the results from an experiment wherein, for example, images that elicit powerful emotions were used as cues for comparing ordinary Web sites that have little semantic relation to the cue images.

In models of aesthetic appraisal and aesthetic judgement (e.g. Leder et al. (2004)), it is assumed that the stimulus needs to be recognised as an object of art for cognitive-affective processes of aesthetic experience to occur. An additional consideration is that, in a broader sense, in psychology of arts and in empirical aesthetics – which encompass almost all the visual aesthetics in HCI research, whether implicitly or explicitly – phenomena such as those explored in this paper would not belong to this line of research, on account of the selection of stimuli. There is healthy discourse nonetheless that involves critically scrutinising other visual artefacts and representations that fall under the same methodological paradigm (Tinio and Smith, 2014). Technological artefacts such as visual UIs can be considered to be visually appealing and involve appraisals similar to those in encounters with art, on account of the nature of the process. That is, the visual experience is not in the physical properties of an object but occurs in the perceiver's mind is informed by the object in question. Clearly, then, a stimulus in itself is not the sole determiner of such cognitive-affective processes.

7. CONCLUSION

Here, we have presented a study utilising the appraisal theory of emotion and the theory of predictive processing to investigate the connection between the visual elements of Web pages and various experiences. A cued-comparison-based method was used to obtain data related to the participants' preferences with respect to the visual aspects of seven sites. Participants' relative preference for the various stimuli when presented with certain experience goals, such as *beautiful* and *civilised*, were assessed. The theoretical frameworks of appraisal and the predictive brain were integrated, and the joint model thereby produced was able to predict and explain measurements such as reaction times and eye movements of the participants. The key conclusion is that the forming of the subjective visual experience is a complex mental process, involving numerous information sources and both top-down and bottom-up information processing. These complexities notwithstanding, visual experience

can be studied via a theoretically and methodologically strong grounding such as the one presented here, with acknowledgement of the various confounding factors that are present in studies of this nature.

The methodology presented in this paper provides a tool by which designers can analyse how well their visual designs correspond to the experience goals targeted. In addition to understanding the users' preference for the various designs and how long it takes for them to identify this preference, the designer can quantitatively compare the users' scanpath patterns as they appraise the designs in question in relation to various experience goals. The results of our work should aid the designer in focusing on the relevant visual areas of the work, while also informing understanding of how particular experience goals are associated with specific areas of interest.

ACKNOWLEDGEMENTS

The authors thank Kirsi Heiskanen for assisting with the experiment. The work was performed with support from the Academy of Finland (310947).

REFERENCES

Altaboli, A., and Lin, Y. (2011). Investigating effects of screen layout elements on interface and screen design aesthetics. Advances in Human–Computer Interaction, 2011, 1–10.

Arnheim, R. (1974). Art and visual perception: A psychology of the creative eye. University of California Press.

Bargas-Avila, J.A., and Hornbæk, K. (2011). Old wine in new bottles or novel challenges: A critical analysis of empirical studies of user experience. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2689–2698). ACM.

Bauerly, M., and Liu, Y. (2006). Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. International Journal of Human–Computer Studies, 64(8), 670–682.

Bornstein, R.F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. Psychological Bulletin, 106(2), 265–289.

Brave, S., and Nass, C. (2003). Emotion in human–computer interaction. In Jacko, J., and Sears, A. (Eds.), The human–computer interaction handbook (pp. 81–93). Lawrence Erlbaum Associates.

Chang, D., Dooley, L., and Tuovinen, J.E. (2002). Gestalt theory in visual screen design: A new look at an old subject. In Proceedings of the Seventh World Conference on Computers in Education: Australian Topics – Volume 8 (pp. 5–12). Australian Computer Society, Inc.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181–204.

Coursaris, C.K., Swierenga, S.J., and Watrall, E. (2008). An empirical investigation of color temperature and gender effects on web aesthetics. Journal of Usability Studies, 3(3), 103–117.

Demir, E., Desmet, P.M., and Hekkert, P. (2009). Appraisal patterns of emotions in human–product interaction. International Journal of Design, 3(2), 41–51.

Ellsworth, P.C. (2013). Appraisal theory: Old and new questions. Emotion Review, 5(2), 125–131.

Fokkinga, S.F., and Desmet, P.M. (2013). Ten ways to design for disgust, sadness, and other enjoyments: A design approach to enrich product experiences with negative emotions. International Journal of Design, 7(1), 19–36.

Folkman, S., and Lazarus, R.S. (1985). If it changes it must be a process: Study of emotion and coping during three stages of a college examination. Journal of Personality and Social Psychology, 48(1), 150–170.

Frijda, N.H. (1988). The laws of emotion. American Psychologist, 43(5), 349–358.

Goldberg, J.H., and Kotval, X.P. (1999). Computer interface evaluation using eye movements: Methods and constructs. International Journal of Industrial Ergonomics, 24(6), 631–645.

Gombrich, E.H. (1995). The story of art. London: Phaidon.

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

Hassenzahl, M., and Monk, A. (2010). The inference of perceived usability from beauty. Human–Computer Interaction, 25(3), 235–260.

Hekkert, P. (2006). Design aesthetics: Principles of pleasure in design. Psychology Science, 48(2), 157–172.

Hekkert, P., Snelders, D., and Wieringen, P.C. (2003). 'Most advanced, yet acceptable': Typicality and novelty as joint predictors of aesthetic preference in industrial design. British Journal of Psychology, 94(1), 111–124.

Hox, J. (2010). Multilevel analysis, 2nd ed. Hove: Routledge.

Irwin, D.E. (2004). Fixation location and fixation duration as indices of cognitive processing. The Interface of Language, Vision, and Action: Eye Movements and the Visual World, 217, 105–133.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254–1259.

Ivory, M.Y., Sinha, R.R., and Hearst, M.A. (2001). Empirically validated web page design metrics. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 53–60). ACM.

Jokinen, J.P.P. (2015). Emotional user experience: Traits,

events, and states. International Journal of Human–Computer Studies 76, 67–77.

Jokinen, J.P.P., Silvennoinen, J., Perälä, P., and Saariluoma, P. (2015). Quick affective judgments: Validation of a method for primed product comparisons. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 2221–2230). ACM.

Just, M.A., and Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87(4), 329–354.

Karapanos, E. (2012). Modeling users' experiences with interactive systems. Springer.

Karvonen, K. (2000). The beauty of simplicity. In Proceedings of the 2000 Conference on Universal Usability (pp. 85–90). ACM.

Ketelaar, T., and Ellis, B.J. (2000). Are evolutionary explanations unfalsifiable? Evolutionary psychology and the Lakatosian philosophy of science. Psychological Inquiry, 11(1), 1–21.

Kim, J., Lee, J., and Choi, D. (2003). 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), 899–940.

Lavie, T., and Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. International Journal of Human–Computer Studies, 60(3), 269–298.

Law, E.L.C., Roto, V., Hassenzahl, M., Vermeeren, A.P., and Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 719–728). ACM.

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

Lin, Y.C., Yeh, C.H., and Wei, C.C. (2013). How will the use of graphics affect visual aesthetics? A user-centered approach for web page design. International Journal of Human–Computer Studies, 71(3), 217–227.

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

Luchins, A.S. (1951). An evaluation of some current criticisms of Gestalt psychological work on perception. Psychological Review, 58(2), 69–95.

Mahlke, S., and Minge, M., 2008. Consideration of multiple components of emotions in human–technology interaction. In Peter, B. (Ed.), Affect and emotion in human–computer interaction (pp. 51–62). Berlin: Springer.

Michailidou, E., Harper, S., and Bechhofer, S. (2008). Visual complexity and aesthetic perception of web pages. In Proceedings of the 26th Annual ACM International Conference on Design of Communication (pp. 215–224). ACM.

Miniukovich, A., and De Angeli, A. (2015). Computation of interface aesthetics. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 1163–1172). ACM.

Moors, A., Ellsworth, P.C., Scherer, K.R., and Frijda, N.H. (2013). Appraisal theories of emotion: State of the art and future development. Emotion Review, 5(2), 119–124.

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

Ngo, D.C.L., and Byrne, J.G. (2001). Application of an aesthetic evaluation model to data entry screens. Computers in Human Behavior, 17(2), 149–185.

Osgood, C., May, W., and Miron, M. (1975). Cross-cultural universals of affective meaning. Chicago: University of Illinois Press.

Papachristos, E., and Avouris, N. (2013). The influence of website category on aesthetic preferences. In IFIP Conference on Human–Computer Interaction (pp. 445–452). Berlin: Springer.

Post, R.A.G., Blijlevens, J., and Hekkert, P. (2016). 'To preserve unity while almost allowing for chaos': Testing the aesthetic principle of unity-in-variety in product design. Acta Psychologica, 163, 142–152.

Raita, E., and Oulasvirta, A. (2011). Too good to be bad: Favorable product expectations boost subjective usability ratings. Interacting with Computers, 23(4), 363–371.

Reber, R., Schwarz, N., and Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? Personality and Social Psychology Review, 8(4), 364–382.

Reinecke, K., and Gajos, K.Z. (2014). Quantifying visual preferences around the world. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 11–20). ACM.

Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., and Gajos, K.Z. (2013). Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2049–2058). ACM.

Rogers, Y. (2012). HCI theory: Classical, modern, and contemporary. Morgan & Claypool.

Saariluoma, P., Jokinen, J., Kuuva, S., and Leikas, J. (2013). User experience as mental contents. In Proceedings of the 10th European Academy of Design Conference. Gothenburg: Chalmers University of Technology.

Salvucci, D.D., and Goldberg, J.H. (2000). Identifying fixations and saccades in eye-tracking. In Proceedings of the Eye Tracking Research & Applications Symposium 2000 (pp. 71–78). ACM.

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

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

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

Seo, K.K., Lee, S., and Chung, B.D. (2016). Effects of perceived usability and aesthetics on emotional responses in different contexts of use. International Journal of Human–Computer Interaction, 32(6), 445–459.

Silvennoinen, J.M., and Jokinen, J.P.P. (2016). Aesthetic appeal and visual usability in four icon design eras. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 4390–4400). ACM.

Silvennoinen, J.M., and Jokinen, J.P.P. (2016). Appraisals of salient visual elements in web page design. Advances in Human–Computer Interaction, 2016.

Silvennoinen, J.M., Rousi, R., Jokinen, J.P.P., and Perälä, P.M. (2015). Apperception as a multisensory process in material experience. In Proceedings of the 19th International Academic Mindtrek Conference (pp. 144–151). ACM.

Silvia, P.J. (2009). Looking past pleasure: Anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. Psychology of Aesthetics, Creativity, and the Arts, 3(1), 48–51.

Smith, C.A., and Kirby, L.D. (2001). Toward delivering on the promise of appraisal theory. In Scherer, K.R., Schorr, A., and Johnstone, T. (Eds.), Appraisal process in emotion (pp. 121–138).

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

Sudman, S., Bradburn, N.M., and Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco: Jossey-Bass.

Summala, H., Nieminen, T., and Punto, M. (1996). Maintaining lane position with peripheral vision during in-vehicle tasks. Human Factors, 38(3), 442–451.

Thanh Vi, C., Hornbæk, K., and Subramanian, S. (2017). Neuroanatomical correlates of perceived usability. In Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (pp. 519–532). ACM.

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

Tinio, P.P.L., and Smith, J.K. (2014). The Cambridge handbook of the psychology of aesthetics and the arts. Cambridge University Press.

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

Tuch, A.N., Bargas-Avila, J.A., and Opwis, K. (2010). Symmetry and aesthetics in website design: It's a man's business. Computers in Human Behavior, 26(6), 1831–1837.

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

Von der Malsburg, T., and Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? Journal of Memory and Language, 65(2), 109–127.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. Journal of Personality and Social Psychology, 9(2), 1–27.