Towards a theory of user judgement of aesthetics and user interface quality

JAN HARTMANN
ALISTAIR SUTCLIFFE
ANTONELLA DE ANGELI
Manchester Business School, University of Manchester (UK)

The paper introduces a framework for users' design quality judgements based on Adaptive Decision Making theory. The framework describes judgement on quality attributes (usability, content / functionality, aesthetics, customisation and engagement) with dependencies on decision making arising from the user's background, task and context. The framework is tested and refined by three experimental studies. The first two assessed judgement of quality attributes of websites with similar content but radically different designs for aesthetics and engagement. Halo effects were demonstrated whereby attribution of good quality on one attribute positively influenced judgement on another, even in the face of objective evidence to the contrary (e.g. usability errors). Users' judgement was also shown to be susceptible to framing effects of the task and their background. These appear to change the importance order of the quality attributes; hence, quality assessment of a design appears to be very context dependent. The third study assessed the influence of customisation by experiments on mobile services applications, and demonstrated that evaluation of customisation depends on the users' needs and motivation. The results are discussed in the context of the literature on aesthetic judgement, user experience and trade-offs between usability and hedonic / ludic design qualities.

Categories and Subject Descriptors: H5.2 [**Information Interfaces and Presentation**]: User Interfaces - *User-centered design*; *Interaction styles*; *Theory and methods*
General Terms: Design, Experimentation, Human Factors
Additional Key Words and Phrases: Aesthetics, Usability, Interaction Styles, Judgement and Decision-Making

1. INTRODUCTION

The new usability agenda [Hallnäs and Redström 2002; Norman 2004] has broadened the determinants of quality in user interfaces by drawing attention to aesthetic design and the influence of emotions in design. Some heuristic evaluation techniques make reference to aesthetics and user engagement [Nielsen 1993, 2000; IBM 2000]. However, few experimental investigations into aesthetic components of usability have been carried out apart from those by Tractinsky [1997], who demonstrated that users' perception of aesthetic qualities was an important, and culturally variable, component of quality in experiments with ATM user interfaces. Further evidence for the importance of aesthetics can be found in the study by Hassenzahl et al. [2000], who asked users to compare six different designs of a process control application, with questionnaire inventories for experience, hedonic and appeal qualities. Hassenzahl et al. concluded that both experience and hedonic qualities contributed approximately equally to the overall judgement of appeal. However, these studies either did not specify which design features they varied in testing aesthetic quality, or varied only simple aspects such as colour and layout consistency, as in Tractinsky's studies.

Manchester Business School, University of Manchester, Booth Street West, Manchester M15 6PB, UK
Contact author: a.g.sutcliffe@manchester.ac.uk

In his influential book, Norman [2004] claimed that aesthetic design can be even more influential in affecting user preferences than traditional operational usability. This claim reflects well-established knowledge in marketing, product design, and even social psychology: namely, that beauty matters. The aesthetic quality of a product influences consumers' attitudes, and is a major determinant of its marketplace success [Bloch 1995]. Beauty can be an important quality of a product; furthermore it seems to influence other judgements about the same person or object. For example, not only do people associate positive personality traits with attractive individuals [Dion et al. 1972], but they also tend to make more positive judgements on other attributes, such as intelligence [Meiners and Sheposh 1977] – often referred to as a halo effect.

Consistent with the halo effect, several studies have suggested a correlation between the aesthetic quality of an interface and its perceived usability [Tractinsky 1997; Tractinsky et al. 2000], as well as the overall user satisfaction with that system [Lindgaard and Dudek 2003]. More recently, these findings have been contradicted by experimental studies which found no, or only a weak, correlation between perceived aesthetic quality of MP3-player skins and judgement of pragmatic (perceived and experienced usability) product attributes, thus suggesting that positive aesthetic appreciation may not be strongly affected by usability evaluation [Hassenzahl 2004]. This inconsistency indicates the need for a better conceptualisation of what constitutes the "user experience" and in particular what is beauty in interaction [Overbeeke et al. 2002] and teasing apart the visually perceived aspect of aesthetics from engagement during interaction. However, with a limited number of exceptions [Sutcliffe and De Angeli 2005; De Angeli et al. 2006], few studies have been undertaken on how different interaction styles might influence aesthetics or other aspects of users' judgement of interactive experience.

In this paper we report investigations into user judgement of UI quality that focus on aesthetics and its interaction with usability, content, and customisation. Our motivation is ultimately to inform trade off judgements in design and guidance about which quality criteria to invest in for particular genres of interactive products. In the following section we review related work. Section 3 introduces our initial motivating theoretical framework, followed by results from the first set of experiments in section 4; the latter section summarises and reworks previously published investigations [Sutcliffe and De Angeli 2005; De Angeli et al. 2006]. Section 5 describes our more recent experimental investigations, which extend the initial framework. Section 6 proposes a revised theory which presents a summary of our understanding to date, and hypotheses that will motivate future work. The paper concludes with a discussion of the contribution and evolving state of knowledge on aesthetics and affective interaction.

2. RELATED WORK

Several authors have drawn attention to aspects of user interfaces that are variously described as user engagement, experience, or emotion in design [Reeves and Nass 1996; De Angeli et al. 2002; Tractinsky and Zmiri 2006]. McCarthy and Wright [2005] view interactive technology as an experience, introducing a framework that describes compositional, sensual, emotional, and spatio-temporal threads of experience as "ways of talking about technology"; while Norman [2004] claims that aesthetic design can outweigh usability in the users' overall experience and argues for the emotional impact of good design. Hallnäs and Redström [2002] see aesthetics as the logic of expressionals and a foundation for designing for p*resence*, defining an expressional as "a thing that is designed to be the bearer of a certain expression". Aesthetics as the logic of expressionals is then "concerned with how material builds expressive things".

Based on Shusterman's concept of Pragmatist Aesthetics [1992], Petersen et al. [2004] proposed a framework to distinguish between the aesthetics of use and appearance. They argued that interaction and playful design is an important influence on aesthetic perception, while Djajadiningrat et al. [2000] discussed the role of affordances in aesthetic design and Norman's propositions for affective design with feedback dimensions, and note that the socio-cultural context of aesthetic perception needs to be taken into account. Lindgaard and Dudek [2003] compared websites with different aesthetic design styles and found that a site with high aesthetic appeal but very low perceived usability yielded very high overall satisfaction. Thus they suggest that what is considered to be "beautiful" need not necessarily also be perceived to be usable; consequently, designers may need to pay attention to both visual appeal and usability.

Lavie and Tractinsky [2004] proposed a model of website aesthetics which differentiated between *classical* aesthetics, referring to traditional aesthetic notions emphasising orderly and clear design, and *expressive* aesthetics, which the authors associate with the design's creativity and originality. Their questionnaire-based measurement instrument for *classical* aesthetics emphasises orderly and clear design and is related to design rules advocated by usability experts, e.g. principles such as consistent and structured layout, symmetry, clean and clear design; whereas *expressive* aesthetics is "manifested by the designer's creativity and originality and by the ability to break design conventions" with questions focused on users' perception of user interface qualities, such as "beautiful", "challenging" and "fascinating". However, these questionnaires only elicited users' high-level perceptions of design quality, essentially the interface look and feel, rather than requesting judgement about components of the design.

Kim et al. [2003] related design features such as colour, texture, and menu layout to aesthetic and emotional responses which users frequently feel when viewing emotionally evocative home pages. They found 13 emotional dimensions which clustered design features and perceived qualities, but no model of aesthetic judgement was produced and the dimensions were closer to aesthetic perceptions than affective responses. Park et al. [2004] analysed critical factors that determine the degree to which users feel the impressions intended by the designers (aesthetic fidelity), and found that variability of user perception was closely related to the designers' intent. In spite of a number of studies, no model or theory has emerged to explain how and why users judge particular design to be more or less aesthetic, apart from Norman's proposal that emotion may play a role via users' reflective and visceral reaction to specific designs; even though Norman does not describe how such reactions influence users' judgement or behaviour.

3. THEORETICAL FRAMEWORK

Based on previous evaluations of websites [Ivory and Hearst 2001; Sutcliffe 2002b; Sutcliffe and De Angeli 2005; De Angeli et al., 2006; de Bruijn et al. 2007], we propose a theoretical framework for judgement of user interface quality based on Adaptive Decision Making theory [Payne et al. 1993], illustrated in Figure 1. ADM theory asserts that people's decision making is adaptive and contingent upon the task, context, and their background-experience.

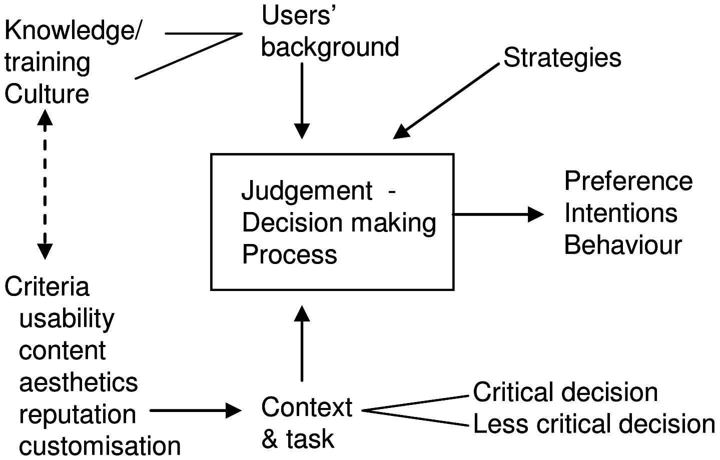


Fig. 1. Initial theoretical framework, based on Adaptive Decision Making theory.

It proposes that people make decisions by adapting their strategies to the task in hand, and that decision making is contingent on their background experience and the criticality of the decision. For example, for an important decision such as the purchase of high-value goods, most people will adopt a slow-path reasoning-intensive process with filtering and multi-attribute comparisons. Our elaboration of ADM hypothesises that users' judgement will also depend on interactions between decision-making criteria (e.g. design qualities such as content, aesthetics, functionality, usability) that are conditioned by the task context. In the context of ADM theory we posit that users will adopt a mixture of filtering on different criteria (Elimination by Aspects), and multi-attribute comparison (Weighted Attributes Decision) strategies to trade off between different quality criteria. For example designs could be selected by filtering based on perceived aesthetics to select the most 'beautiful'; alternatively, the overall assessment might be made by weighing the actual usability experience against the gain in content and services delivered by the application. Different strategies of this nature are implicit in Hassenzahl's theory of user judgement of beauty and goodness of interactive products [Hassenzahl 2004]. We hypothesise that preferences for user interface designs when the scenario of use is critical will be based on more in-depth consideration of quality attributes, whereas for less serious contexts of use, preferences will be governed by selecting designs based on general aesthetic impressions. The outcomes of users' judgement are preferences between designs, intention to use, and actual use (behaviour). The key research questions we wished to investigate are how design quality attributes bear upon users' overall judgement, the importance weightings of the attributes, and their possible interaction. In particular we wish to explore the relationship between usability, aesthetics and content in more depth than previous studies.

Five judgement criteria are proposed in the initial framework for attractiveness ("pleasing or appealing to the senses, arousing interest" Oxford English Dictionary, Sutcliffe and De Angeli 2005): *usability* following the widely accepted ISO definition [ISO 1997], emphasising efficiency, effectiveness, and satisfaction, as well as including ease of learning, efficiency of use, memorability, low error frequency. Appropriate and interesting *content* is widely cited to be a key factor in successful website design [Nielsen and Molich 1990; Mullet and Sano 1995; Lynch and Horton 2001] with *services* to describe the functions. Content and services correspond to the utility of an application. *Aesthetics* reflects the format in which the content and services are presented as well as the design look-and-feel of a system [Hallnäs and Redström 2002; Norman 2004; McCarthy and Wright 2005]. *Reputation / identity* relates to the identity of the website owner and the brand-product identity which can be projected by a consistent visual style, logos, and product presentation [Merrilees and Fry 2002]. Finally, *customisation* describes the extent to which the user can adapt the system to his or her needs; this can encourage users to take ownership over a system, and has been found to influence perceived usability and aesthetics [Blom and Monk 2003].

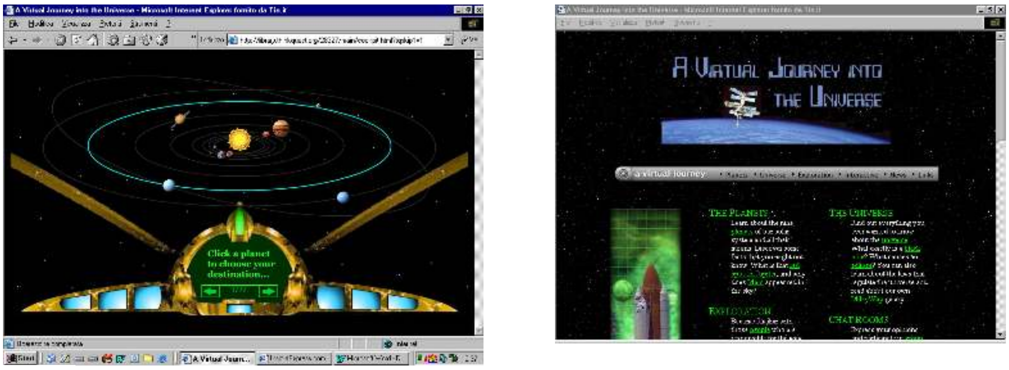
The first two experiments concentrate on aesthetics and usability, while the third investigates the effect of customisation on aesthetics and content. Customisation is different from the other attributes since it influences judgement in a pre-use phase, during which customisation effort has to be expended; and then in the use phase, the potential reward may be gained, so a cost-benefit analysis is added to address this issue.

4. STUDIES 1 AND 2: AESTHETICS AND USABILITY

This section reports a combined analysis of data collected in two separate experiments [Sutcliffe and De Angeli 2005; De Angeli et al. 2006]. These studies addressed the evaluations of educational websites and followed a common theme of comparing the influence of aesthetics and usability on user judgement of the user interfaces (UIs). In both studies we selected a live website with different design versions which should afford different perceptions of aesthetics and usability, and compared users' reaction to them. The two versions of each website presented exactly the same information but with different UI styles. One was a traditional menu-based style; the other exploited animated metaphors and more aesthetic features (metaphor-based).

4.1. Materials

The website [Sutcliffe and De Angeli 2005] was selected from the ThinkQuest Library, a collection of educational sites entered in an international competition sponsored by the Oracle Education Foundation. Two versions of the website presented the same content on astronomy with very different design styles, one metaphor-based, the other a traditional menu-based style. The versions were designed by students who were not HCI experts. In this paper we will refer to them as the Astronomy website. Example screen-shots are shown in Figure 2.

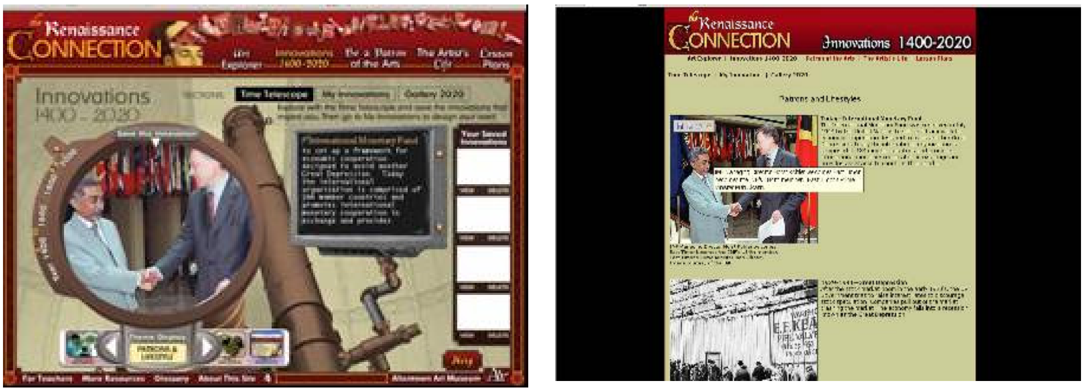

A Virtual Journey into the Universe - Microsoft Internet Explorer fornito da Tin.it
Click a planet to choose your destination...
A Virtual Journey into the Universe
a virtual journey
The Planets
The Universe
Exploration
Chat Rooms

(a) (b)

Fig. 2. (a) Metaphor-style interface for the Astronomy site, showing the planets and cockpit metaphors; and (b) menu-based user interface

The Astronomy site (http://library.thinkquest.org/28327/) had 210 content pages organised in a number of sections: planets, universe, exploration, news, resources and glossary, with further pages linking interactive features such as chat rooms and feedback forms, most of which had been disabled. In the interactive metaphor version, an animated representation of the solar system was displayed in the first page with a "cockpit" metaphor of controls at the bottom of the screen (see Figure 2a). Moving the mouse over the planets caused an orbit to highlight, then clicking on the planet or its orbit caused the planet's name to appear in the cockpit display. Information about the planet could then be accessed by "Go" in the cockpit metaphor. Some fast moving planets such as Mercury were difficult to access, while Uranus and Neptune were not easily visible. The side panels in the cockpit contained other navigation links to chat rooms, universe, exploration, etc., although the user had to move the mouse over the panel to display the option name. The menu interface version, in contrast (Figure 2b), had a standard link menu to access information without any intervening displays.

The second website [De Angeli et al. 2006] was designed by Eduweb for the Allentown Art Museum in Pennsylvania, for history classes by middle-school students (10-14 years) and their teachers. (Figure 3: http://www.renaissanceconnection.org). In this paper we will refer to it as the History website. The metaphor-based version (Figure 3a) adopted a playful and engaging interaction style, with animated characters providing information by speech bubbles, and generating other pictures and information from inside their heads. The menu-based version (Figure 3b) adopted a more serious interaction style, displaying a static picture instead of the animated head, and with no humorous effects. The website was structured in five sections, three of which were the object of the evaluation study. The *Innovation 1400-2020* section (Figure 3a) was the most different as regards interaction style and was used to perform four of the eight information retrieval tasks in the study. A telescope metaphor was used to select information themes by moving a sequence of pictures at the bottom of the telescope (leading to a change of the displayed picture in the lens) and the year was selected by moving a slide on the lens. This website had already been subjected to a comparative usability evaluation [Schaller et al. 2004], which addressed unstructured behavioural observations and the comments of middle-school and college students. In our study, we removed most of the fun features from the menu-based interface, and replaced them with pictures of a more serious content to maximise look and feel differences between the two interaction styles.



(a) (b)

Fig. 3. Metaphor-based version (a) of the History website, showing the telescope and slide controls for timeline and subject matter filters; and (b) menu-based version.

In the Astronomy website, the metaphor was used for “higher-level” navigation purposes only, e.g. in the selection of a particular planet and a sub-category. In the History website, the metaphors were used for lower-level navigation, to filter items by time period. The content was identical within each pair of sites. The Astronomy website was live, but was not altered during the course of data-collection, so all participants viewed exactly the same site. For the History website, we created a copy of the live website to be hosted on our servers to ensure no change in content for the course of the experiment.

4.2. Method

Participants

The Astronomy website was evaluated by 25 students (21 male; 4 female), and the History website by 29 students (23 male; 5 female). All were students of the University of Manchester and had basic knowledge of HCI and usability evaluation techniques from HCI courses that they had recently attended. All the participants were expert web-users; none had any prior knowledge of the websites.

Procedure

The experimental procedure was very similar in both studies. Data was collected in a group setting, with each participant working individually for almost 3 hours, followed by a group discussion. On arriving for the experimental session, the participants received verbal and written instructions, followed by a brief pre-test questionnaire recording personal data, internet experience, and level of interest and knowledge of the website’s subject area. Then each participant was asked to perform a series of information retrieval tasks with both the metaphor and the menu-based versions of the websites. There were two sets of equivalent tasks for each UI version, so none of the tasks were repeated. The orders were counterbalanced for UI-styles and tasks. For each condition the experimental sequence was:

1. Perform information retrieval tasks (e.g. what is the orbital period of Jupiter; find Renaissance artists' names and painting dates).
2. While performing the tasks, describe any usability errors encountered and rate their severity.
3. After task-completion, participants completed a free recall memory test. This required listing the first ten facts / items / issues they could remember about the website, and rating the quality of these memories on a five-point scale as favourable, neutral or adverse.
4. Then, they briefly revisited the site and evaluated the version of the website regarding usability, and visual aesthetics.
5. After this had been completed for both UI-style versions, a group discussion was run to investigate overall preferences and reasons behind them. For the History websites, participants also individually completed a post-test questionnaire, which captured their overall preferences and the reasons behind them. They also selected the ideal UI-style for different target populations and environments (e.g. children aged 7-10 at home; children aged 7-10 at school).

Evaluation Instruments

In both studies, websites were evaluated for usability, memorability of content and interface features, aesthetics, information quality, engagement and overall preferences. Several techniques and instruments were used to gather evidence on these dimensions. They are summarised in Table I.

Table I: Summary of evaluation instruments used in the two studies.

Usability	• *Performance analysis* • *Self-report and severity rating of usability problems (1=minor problem; 5=major problem)* • *5-item usability scale on a 7-point Likert scale [Lavie and Tractinsky 2004]*
Memorabilty	• Free recall memory test and memory rating (1=very negative, 5=very positive)
Aesthetics	• Heuristics for attractiveness [Sutcliffe 2002a, 2002b] • 10-item perceived website aesthetic scale on a 7-point Likert scale [Lavie and Tractinsky 2004]
Engagement	• 3 items on a 7-point Likert scale (engaging, entertaining, pleasant)
Information/service quality	• Astronomy websites: 3 items measuring service-quality [Lavie and Tractinsky 2004] • History websites: *Scale A* - six items on a 7-point Likert scale expanding the service quality measure and adapting it to the evaluation of information quality in educational software *Scale B*: nine items on a 7-point Likert scale extracted and adapted from the Bernier Instructional Design Scale BIDS [Bernier 1996]
Overall preference	• Astronomy websites: group discussion • History websites: Dichotomous choice on the post-test questionnaire, including different scenarios

In both studies, usability was assessed by objective measures (performance analysis and report of usability problems) and subjective measures (questionnaire). Aesthetics was assessed by two distinct yet complementary approaches. The perceived aesthetics scale [Lavie and Tractinsky 2004] was used to collect holistic impressions on two apparently separate dimensions, namely classical aesthetics and expressive aesthetics. The first dimension includes items such as pleasant, clear, clean, symmetrical and aesthetic design. The dimension of expressive aesthetics is characterised by qualities that capture the user's perception of creativity and originality of the site's design. Relevant items in this dimension are creative, fascinating, original, sophisticated design, use of special effects. In addition, participants evaluated the site applying the heuristics for attractiveness, proposed by Sutcliffe [2002a]. These address the quality of individual design features linked to the perception of aesthetics. Engagement was measured by the following 3 items: engaging, entertaining and enjoyable.

Information quality addressed the educational impact of the website and was directly assessed in the History websites. A first measure employed in this study builds upon Lavie and Tractinksy's [2004] service-quality construct (measured by the items "makes no mistakes", "provides reliable information", "reliable" which were used also in the Astronomy web-site evaluation), and localised to the analysis of educational software by adding the following items: "provides enough details", "informative", "educational". The second measure builds on the Bernier Instructional Design Scale (BIDS), a psychometric instrument developed to assess the quality of printed education material [Bernier 1996]. For the history study we selected and adapted nine items, directly related to clarity of learning objectives, level of detail, quality of content, learning potential, delivery of up-to-date information.

Overall preferences were collected by group discussion in the Astronomy study and by a post-evaluation questionnaire, asking users to express their choice on a dichotomous question, for overall preference in a number of different scenarios. Indirect measures of preference were obtained by questionnaire comparisons.

Design

The results reported in this paper are based on a 2 * 2 design contrasting *UI-style* by *website*. UI style (Menu-based vs. Metaphor-based) was the main experimental factor in both studies and was originally manipulated within participants. Website (Astronomy vs. History) is used here as a post-hoc between-participants factor. A group of participants evaluated the Astronomy websites [study 1: Sutcliffe and De Angeli 2005] and another group the History website [study 2: De Angeli et al. 2006]. It has to be noted that this design is not orthogonal, as the two websites were completely independent (different design, look-and-feel, content) and the experiments were performed at different times. Nevertheless, the similarities between the testing procedures and the UI-style are strong enough to allow the comparison.

4.3. Results

The results are summarised in 7 sections for usability evaluation, ratings of aesthetics, information quality, engagement, the memorability test, and finally the participants' overall preference. All scales showed high reliability (Cronbach's alpha > .80); therefore comparisons are based on the average of individual items.

Task Performance

Information-retrieval performance was extremely accurate in both studies, with only 4% of 748 tasks resulting in inaccurate or wrong information (6% of 300 for the Astronomy site; 3% of 448 for the History site). There were no significant differences in task performance between the design versions in both websites.

Usability

The menu-based version caused fewer usability errors and was rated as better for usability than the interactive-metaphor version for both sites. The subjective evaluation of usability was analysed by a 2 * 2 ANOVA with website (2) as between-participants factor and UI-style (2) as within-participants factor. Results showed a significant difference for UI-style $F_{(1,48)} = 25.50$; $p < .001$ and no significant difference between the websites, nor significant interactions. Participants in the metaphor-based condition consistently perceived the usability of the site as significantly worse (M = 3.63, SE = .20) than in the menu-based design (M = 4.99, SE = .17).

Objective usability measures followed the same pattern. Overall 147 problems were reported in the history site (47 in the menu-based and 100 in the metaphor-based condition) and 163 in the astronomy site (58 in the menu-based and 105 in the metaphor-based condition). A 2 (UI-style) * 2 (website) mixed-model ANOVA on problem severity showed a clear effect of UI-style $F_{(1,304)} = 27.84$, $p < .001$, no difference between websites, and no significant interaction. Overall, the problems associated with the metaphor style were perceived as more severe (M = 3.69 SE = .9) than the problems associated with the menu-based style (M = 2.91, SE = .12).

In the metaphor-based version of the Astronomy website, usability problems were frequently associated with specific design features. The most troublesome design features were the solar system and planets metaphor, which was responsible for 22% of the total errors, followed by the cockpit metaphor with 7%. The more common observed problems were the lack of a caption for the planet name in the solar system metaphor (N = 14); the difficulty of selecting a planet (10); and the obscure interaction in the cockpit menu (8). The most common complaints in the menu-based version were inconsistent use of the back and home button (N = 6) and the need for scrolling (N = 7).

A similar pattern emerged within the History website. Menu / navigation problems in the metaphor version mainly occurred in the innovation section and the telescope metaphor (N = 20). In contrast, the menu / navigation issues in the menu-based interface were more general and differentiated, such as broken or missing links (N = 8) and the need for scrolling (N = 6).

Aesthetics

The measures of classical and expressive aesthetics were entered as dependent variables in an ANOVA with aesthetic dimension (2) and UI-style (2) as within-participants factors and website (2) as between-participants factor. Both the within-subjects effects were significant, UI-style $F_{(1, 47)} = 8.66$, $p < .01$ and aesthetic-dimension $F_{(1,47)} = p < .05$. These results however must take into account a very strong 2-way interaction UI-style * website, $F_{(1,47)} = 78.94$, $p < .001$. All other sources of variances were not statistically significant.

Further analyses elucidated the nature of the interaction. Specifically, a t-test indicated that expressive aesthetics was evaluated significantly better in the metaphor-based version than in the menu-based conditions $t(49) = 7.25$, $p <.001$. In contrast, no significant effect between the two UI-styles emerged in the evaluation of classical aesthetics $t(49) = -187.$, $p = .07$ Average values are illustrated in Figure 4.

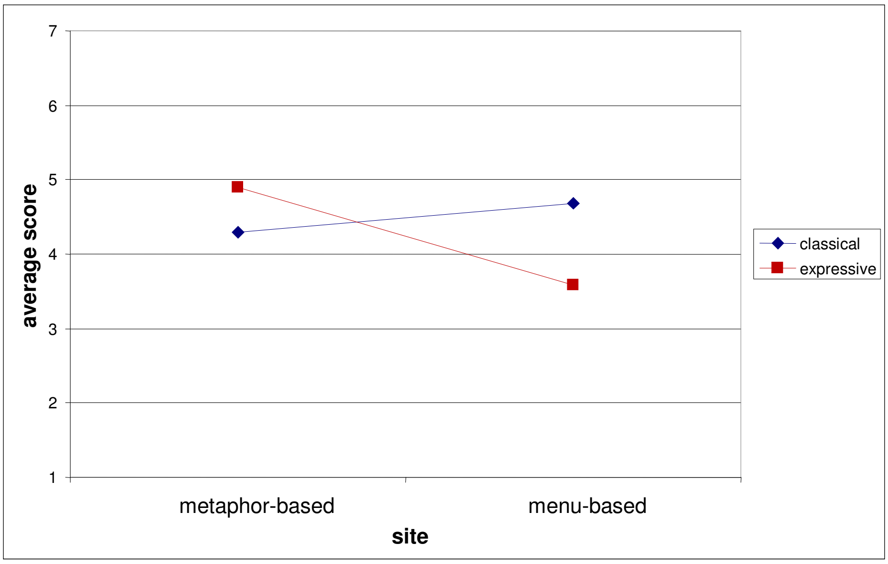


Figure 4, Average evaluation scores of classical and expressive aesthetics as a function of UI-style.

Engagement

Participants' assessment of engagement favoured the metaphor-based design in both sites. A 2*2 mixed-model ANOVA with UI-style (2) as within-subjects factor and website (2) as between-subjects factor showed that UI-style significantly affected the perception of engagement $F_{(1,48)} = 10.40$; $p < .01$. There were no effects between sites and no interactions. The metaphor-based version of both sites was perceived as more engaging (M= 4.70; SE = .19) than the menu-based interface (M = 3.97; SE = .18). In the history site's post-test questionnaire 19 / 24 users picked the metaphor-based interface as the most engaging.

Service and Information Quality

Service quality, measured by Lavie and Tractinksy's [2004] inventory, was consistently better for the menu-based version of both sites (M = 4.67, SE = .17) than for the metaphor-based version (M = 4.0, SE = .15). The 2*2 mixed-model ANOVA with UI-style as within-participants factor and website as between-participants factor showed a significant effect for UI-style $F_{(1,47)} = 12.69$; $p < .01$. All other sources of variance returned no significant results.

For the History website we also measured information quality by adapting a subset of items from the Bernier Instructional Design Scale (BIDS). The menu-version was more favourably ranked (M = 4.87, SE= .21) than the metaphor-based version (M = 4.35, SE = .20), t(23) = 1.1, $p < .05$). In the post-test dichotomous preference question, 75% of the sample also indicated the menu-based style as their favourite interface for content.

Memory

Overall, participants reported 657 items in free recall memory in the evaluation of both websites (289 for the Astronomy site and 368 for the History site). Memory items were clustered into reports referring to the user interface and content memories. In both websites, relatively more content memories were elicited after using the menu based UI-style (42% within UI-style) than after using the metaphor-based site (25%), $\chi^2 = 21.05$ $p < .001$.

An ANOVA was run to analyse possible differences in the valency of individual memory items as a function of type (2: UI vs. content), UI-style (2: flash vs. html) and websites (2: Astronomy vs. History). Valency scores were transformed into Z-values, as the scales used in the two experiments were different (1-5 for the Astronomy website and 1-7 in the History site). Scores were re-coded in the Astronomy data-base so that higher numbers reflected positive evaluations, consistently with the scale used in the History study. The ANOVA returned a significant main effect for UI-style $F_{(7,617)}= 4.81$; $p < .05$, and a significant interaction Type *Web-sites $F_{(7,617)} = 7.07$; $p < .01$. The menu-based style tended to elicit more positive memory than the metaphor based one. The interaction is due to a difference in the evaluation of content and UI elements in the Astronomy website (content was better evaluated) which is not evident in the History website.

Overall Preferences

During the group discussion of the Astronomy site, all participants agreed that the metaphor version was worse for usability and many participants pointed out problems with the design. Nevertheless, when asked which version they personally preferred and might be willing to use again in the future, all but one subject voted for the metaphor-based style. Their preference was explained by an increase in engagement and pleasure experienced during the interaction, even though this came at the cost of an increased workload to understand interaction procedures and solve usability problems. The participants were attracted by the planet animation on the front page which strongly appealed to their curiosity even though it was clearly recognised as the worst usability feature of the design. However, when asked to assess which version was more suitable for teaching secondary school pupils, the participants were evenly divided between the two interaction styles. The main reasons cited in favour of the menu-based style were simplicity and clarity; while the reasons in favour of the metaphor-based style were engagement by interaction and motivation.

Overall preference for the version of the history site was tied, with 14 participants favouring each interface. When asked to justify their decision, nine participants who preferred the menu version explicitly referred to negative features of the metaphor version; in contrast, only two people explicitly referred to negative features of the menu-based version when stating their preference for the metaphor-based design. The reasons driving participants' preferences were very different for each design. The most common reason for preferring the metaphor-based interface made explicit reference to a *more engaging* (N=8*)* and *more interactive* (7) style. Only two participants declared it was *easier to use*. On the other hand, all but one of the participants who preferred the menu-based style made explicit reference to usability issues. Participants' preferences for the design style changed drastically according to the target population and the scenario of use. A clear majority agreed that the metaphor-based style was better for children interacting at home (leisure time), whereas less agreement was found when the system was meant to be used in a classroom environment (formal education). Similarly, the metaphor-based style was deemed inappropriate for more mature and knowledgeable target populations (see table II).

Table II. Participants' preferences for the History and Astronomy (underlined) sites for scenarios with different target populations.

Scenario of use	**Menu-based** %	**Metaphor-based** %	**Binomial test** p
Children (7-10), school	20	71	.05
Children (7-10), home	14	86	.001
Children (10-14), school	32	68	n.s
Children (10-14), home	21	79	.01
Teenagers	43	57	n.s
University students	85	15	.001
Arts experts	82	18	.001
Elderly	89	11	.001
General use	4	96	.001
Education	52	48	n.s

In a supplementary investigation into which criteria predicted overall preference, a binary logistic regression was conducted on the data from the history site. The forward stepwise method based on likelihood ratio was applied. The analysis is similar to a linear regression but is better suited to a model where the dependent variable is dichotomous. It predicts whether an event will or will not occur and identifies the variables useful in making that prediction. The dependent variable was the overall user preference, as expressed by their choice in the post-test questionnaire (metaphor-based vs. menu-based). We selected three main covariates for the model, corresponding to the evaluation dimensions which strongly differentiated the two interaction styles (usability, expressive aesthetic and information quality). Engagement was discarded as it is highly correlated to expressive aesthetics in both UI-styles (r = .89 and r = .71) and the sample was not large enough to reliably accommodate four factors. The model predictors corresponded to the within-participants differences on the evaluation criteria among the two UI-styles (e.g. usability_predictor = usability_menu-usability_metaphor).

Results suggest that usability was an important predictor of final evaluation (Nagelkerke R Square = .50). Together with expressive aesthetics it explained 88% of the final preferences (total

Nagelkerke R Square = .77). The model significantly fit the data ($\chi^2_{(3)}$ = 22.97, *p* < .001), and it becomes less reliable if information quality is added. From these results, we can conclude that differences in evaluation of usability and expressive aesthetics are good predictors of overall preferences, whereas information quality is not.

To further investigate these results we ran a series of mixed model ANOVAs with UI-style (2) as the within-participants factor and Preference (2, metaphor-based vs. menu-based) as the between-participants factor. The four dimensions which were found to significantly differentiate the two UI-styles (usability and information quality, which favours the menu-based style; and expressive aesthetics and engagement which favour the metaphor-based style) were tested as dependent variables.

In all the analyses we found a significant difference for interaction style (*p* < .001) showing that participants are more negative in evaluating weak attributes of the style they did not prefer. Examples of this effect are illustrated in Figure 5 and 6. Figure 5 reports average values for content and usability as a function of preferred UI-style (x axis). People who preferred the menu-based style were much more severe in evaluating usability and content of the metaphor-based style and more positive in evaluating these dimensions of their favourite style. Similarly, people who preferred the metaphor-based style were much more negative in evaluating both expressive aesthetics and engagement of the menu-based style (Figures 5 and 6).

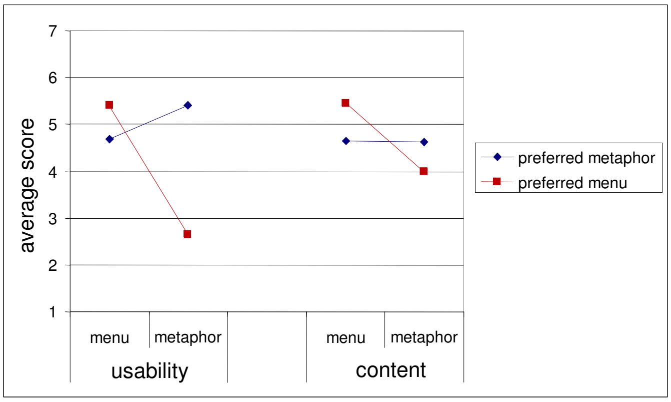


Fig. 5: Average values for content and usability as a function of preferred interaction style

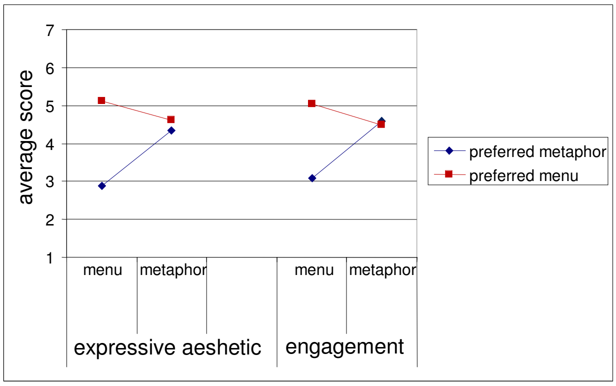


Fig. 6: Average values for expressive aesthetics and engagement as a function of preferred interaction style.

4.4. Summary

The results of both experiments are summarised in table III. From these data it appears that there is a strong association between the design style and users' perception of aesthetics. Design versions with more animation, interactive metaphors and graphics were consistently rated to be more aesthetic and engaging. However, the aesthetic designs exhibited worse usability on both objective and subjective measures.

Table III. Summary of the differences between Menu-based and metaphor-based UI-styles (+ denotes better version, - worse, = no difference between the versions).

Measures	**Menu-based**	**Metaphor-based**
Usability	+	-
Aesthetic classic	=	=
Aesthetic expressive	-	+
Information quality	+	-
Engagement	-	+
Memory	Content-based	Interface-based
Memory valency	+	-
Overall preference Astronomy	**-**	**+**
Overall preference History	**=**	**=**

An interesting effect emerged when it came to the users' evaluation of the site content (which was identical in both versions). For both websites, the menu-based version was rated more favourably for information and service quality. We suggest that good usability may have a positive influence on participants' judgement of content, probably because it facilitates access to content. In contrast, the overall preference for the versions was either for the metaphor version (Astronomy site) or evenly divided (History site). The inconsistency in overall judgement between the sites can be explained by a difference in the evaluation setting. The astronomy experiment collected overall preferences in a group setting, which is likely to be biased by polarisation – the tendency of the average response of group member to become more extreme towards the initially preferred pole after group discussion [Turner 1987].

The regression analysis on participants' individual preferences collected in experiment 2 indicated that usability and expressive aesthetics are important predictors of overall preference. The relative importance assigned to each of these criteria was affected by individual differences. People who preferred the menu-based style were more sensitive to variations in the usability dimension, whereas people who preferred the metaphor-based style were more sensitive towards variations in the expressive aesthetics dimension. A conjecture for future research is that there may be sub-groups of aesthetically sensitive and non-aesthetically sensitive participants in user populations.

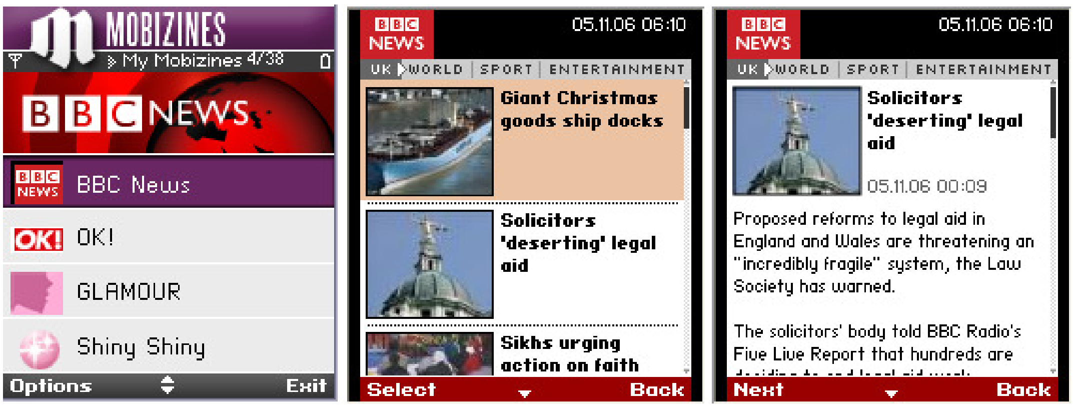
In both sites, participants' judgement showed strong framing effects from the task scenarios. Judgement could be radically biased according to how serious the intended use was. More serious use favoured the more usable but less aesthetically pleasing versions. This agrees with ADM theory, which predicts strong effects on judgement of the task and decision criticality.

5. STUDY 3: AESTHETICS, CONTENT, AND CUSTOMISATION

This section reports an experiment on a mobile content service, with two manipulations: aesthetics of the user-interface and content-fit through customisation. Aesthetics was manipulated between subjects and compared two different user interface styles: one was colourful and animated, and the other implemented a plain design. Both interface designs provided access to the same content-streams. For the content-fit manipulation, both variants were tested with customised and generic versions. The customised version allowed the users to configure the system to subscribe to a personal selection of content-streams of their choice to be available on the device. In the generic version all participants were given the generally most popular content-streams, with no configuration. A final manipulation tested mobile / non-mobile usage context within subjects to investigate mobility as a context factor.

5.1. Materials

Mobizines, a commercially available mobile content service (http://www.mobizines.com), was used; it allows users to subscribe to a set of content-streams (magazines) which can then be browsed on their mobile phones. The application's home screen (Figure 7a) shows a scrollable list of subscribed magazines available for browsing on the phone. Each magazine consists of a list of headlines with corresponding thumbnail images (Figure 7b). Clicking on the headline allows you to read the full story (Figure 7c). In the commercial version, new editions of the subscribed magazines are regularly updated and pushed to the phone. For this study, the content-streams were frozen so there were no differences in the content viewed by different participants.



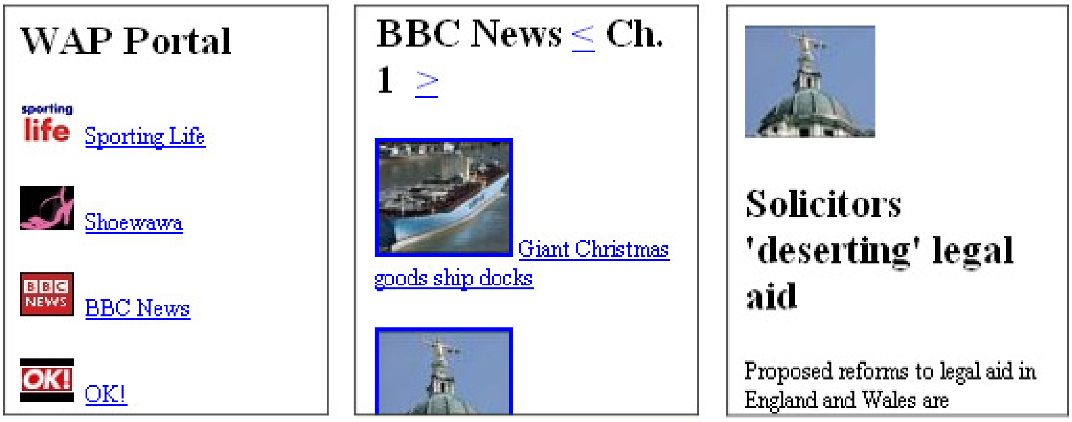
Proposed reforms to legal aid in England and Wales are threatening an "incredibly fragile" system, the Law Society has warned.

The solicitors' body told BBC Radio's Five Live Report that hundreds are

(a) (b) (c)

Figure 7: Mobizines Interface; (a) home screen; (b) magazine with list of headlines; (c) story page

The commercial Mobizines design provides a graphically enhanced interface (colourful with animations, e.g. an animated BBC News logo at the top of Figure 7a). This was used for the high-aesthetics interface condition. We created a low-aesthetics version of the application by implementing the Mobizines content-streams as offline WAP pages stored on the phone with a plain hypertext style interface (see Figure 8). Images that were part of the content, as well as logos of the magazines, were preserved in both variants of the application. The content between the two aesthetic versions remained identical. All applications were run completely offline, so network-latency was not an issue.



WAP Portal

Sporting Life

Shoewawa

BBC News

OK!

BBC News < Ch. 1 >

Giant Christmas goods ship docks

Solicitors 'deserting' legal aid

Proposed reforms to legal aid in England and Wales are

(a) (b) (c)

Figure 8: WAP Interface; (a) home screen; (b) magazine with list of headlines; (c) story page

The magazines available in the tested applications were based on the commercially available content for Mobizines in November 2006. For the content-fit manipulation we created a generic and a customised version of the application. A survey among 469 Manchester University students elicited the 10 most popular of the commercially available magazines. For the generic version, the application was configured with these 10 magazines available for browsing on the phone. For the customised version, users could configure the application with an external program to select their personal favourite 10 magazines out of a list of 31 available streams. These were then loaded for browsing in the actual application (Mobizines / WAP). In total, 4 variants were tested, contrasting aesthetics by content-fit (Figure 9).

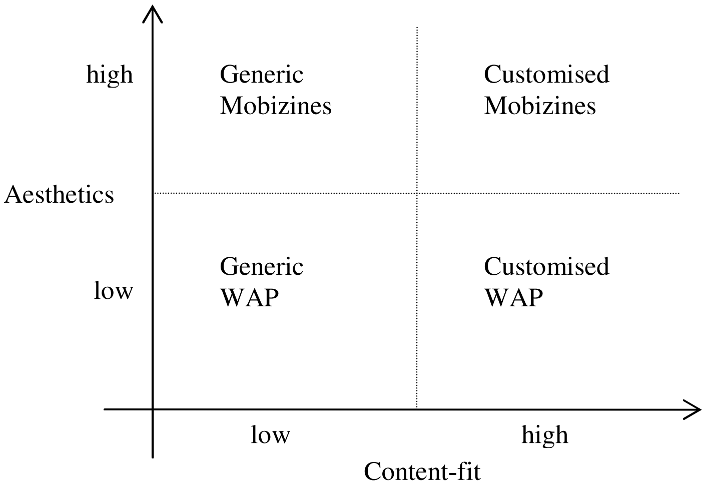


Figure 9: Application variants classified by experimental manipulation

5.2. Methods

Participants

Before the experiments, a sample of 468 students of the University of Manchester filled in a survey reporting their interest in all content-streams available for Mobizines on a 7-point Likert scale, from which the 10 most popular magazines were elicited. A subset of 40 students (age 18-29, mean = 21.58, SD = 2.66; 19 female, 21 male) was randomly selected from the survey population for the actual study. All participants had previous experience with operating mobile phones in general. None of them had any previous knowledge of the tested applications. Participants were randomly assigned to test either the WAP-interface version (N = 20) or the Mobizines-interface version (N = 20).

Procedure

1. In a pre-test phase, the participants were given a choice scenario to elicit general preferences for customisation. All participants indicated general preference between two versions of a mobile content service: Version A was a customised version, where users selected their topics, but needed to configure the application. Version B was a generic version of the system, where the most popular topics had been automatically selected, and no configuration was necessary. Participants were not told about the origins of the generic list or the details of the configuration procedure. All participants then had to indicate their expected cost of configuration for the customised version (6 items on a 7-point Likert scale), and the expected return on investment from customising (3 items on a 7-point Likert scale).

2. Participants were asked to perform a series of reading tasks in a 2 * 2 design with 4 conditions: customised- vs. generic-version and mobile vs. idle. In the mobile conditions they were instructed to move continuously between two locations in the experimental room, to mimic task performance while walking. The reading task consisted of navigating to a specific magazine as instructed and to read *at least* the first paragraph of any story in that magazine. Participants were explicitly allowed to read more than the first paragraph of the story if they wanted to. This was repeated for 5 magazines per condition, so that each available magazine was visited once; the order was randomised. Reading time and any usability errors were recorded, and the participants rated their experienced cost of configuration (6 items) and return on investment (3 items) on 7-point Likert scales. Perceived quality of content, usability, look and feel, and customisation, as well as overall preference rating were measured for both the customised and generic versions.

3. After they finished the tasks for all conditions, participants evaluated the application with Lavie and Tractinsky's [2004] questionnaire inventories for classic and expressive aesthetics, engagement, and service quality.

4. Finally, participants were asked to make a series of preference choices: (a) preference between the customised and generic version; (b) preference between the Mobizines and the WAP interface. For this, they were briefly shown the interface variant that they hadn't used and advised that there were no differences (e.g. in content or functionality) apart from the interface-style; (c) a trade-off question in which participants had to sacrifice one of their preferences from the previous two questions. The question was based on their previous answers, for example if a participant had indicated preference for

customisation and for the Mobizines interface, then the choice given would be between customised WAP and generic Mobizines.

Design

The experiment is based on a 2 * 2 * 2 design with interface-style (2: Mobizines vs. WAP) as between-subjects factor and content-fit (2: customised vs. generic) and mobility (2: idle vs. mobile) as within-subjects factors.

5.3. Results

The results are summarised in 6 sections for customisation-choice, task-performance, ratings of usability, aesthetics, content, and participants' overall preference. All scales showed high reliability (Cronbach's alpha > 0.84); therefore, comparisons are based on the average of individual items.

Customisation Choice

The top ten most popular content-streams elicited from the pre-study survey (468 students) are listed in table IV. These were chosen as available content streams for the generic version of the application.

The study's participants as a group (N=40) generally mirrored the overall picture, with only "T3" and "TV Genius" changing between the pre-study's survey population's favourites and the study's participants' choices for the customised version. The individual preferences among the study's participants varied between users, as the number of stories that overlapped between the generic selection and a participant's individual choice, ranged from 3 (low overlap, i.e. very individual selection) to 8 (high overlap, i.e. nearly identical to generic version) and was normally distributed (mean = 5.55; SD = 1.47).

Table IV. Choice of content streams, top ten most popular streams by mean ratings from survey and the most 10 most frequently selected streams by the participants of the study.

Student Survey (N = 468)	Mean Interest Rating	Study's Participants (N = 40)	N Chosen by
BBC News	5.56	BBC News	37.00
Total Film	4.32	Total Film	34.00
Time Out	4.28	Time Out	32.00
V2 Music	3.67	V2Music	23.00
Popworld	3.29	Tech Digest	21.00
Tech Digest	3.15	Glamour	20.00
TV Genius	3.11	Popworld	19.00
ITV Guide	3.11	T3	15.00
ITV Mobile	3.04	ITV Mobile	15.00
Glamour	3.01	TV Guide	14.00

In the pre-test scenario, 75% of the participants indicated a general preference for a customised version. An ANOVA with preferred-version (2: customised vs. generic) as within-subjects factor and expected cost-of-configuration as dependent variable showed a significant effect ($F_{(1,38)} = 7.25$, $p < 0.05$). Participants who would generally prefer a customised version associated configuration with a significantly lower cost (mean = 3.16; SD = .89) than those who would generally prefer a generic version (mean = 4.11; SD = 1.21).

The average time for configuration of the actual application was 1min 42sec (SD = 45.3sec). In the debrief interview all participants indicated that the cost of customising was very low for this application, consistent with their rating of perceived cost of customising in the questionnaire (mean = 1.82; SD = .83 on a 7-point scale).

Task Performance

Operating the devices was very easy. Hence, most of the task-time (circa 95%) was spent reading. Time was analysed by a 2 * 2 * 2 mixed-model ANOVA with content-fit (2) and mobility (2) as within-subjects factors, and interface-style (2) as between-subjects factor. Task-time differed significantly with content-fit ($F_{(1,35)} = 10.10$; $p < .01$) and mobility ($F_{(1,35)} = 13.62$; $p < .001$) as factors, with no interactions.

The participants read for significantly longer durations (in average 17% longer) in the customised version, which suggest they were more engaged with their own content (table V). People also spent more time reading in the idle version than in the mobile version, which we interpret as a workload effect, since it was more difficult to read while walking about. In this condition, it appears that the participants just read the bare minimum to conform to the experimental task (minimum first paragraph), whereas in the idle condition they read more than the first paragraph.

Table V. Task completion times per condition in seconds.

	content-fit	mobility	Mean	SD
Time	customised	idle	223.53	78.41
	customised	mobile	212.24	77.39
	generic	idle	200.49	62.55
	generic	mobile	169.49	52.45

Usability

Perceived usability was analysed by a 2 * 2 mixed model ANOVA with interface-style (2) as between-subjects factor and content-fit (2) as within-subjects factor. A significant effect of content-fit emerged ($F_{(1,38)} = 3.99$; $p = .05$). No other sources of variance were significant. The customised version is perceived as more usable (mean = 6.13; SD = .69) than the generic version (mean = 5.88; SD = .85).

In the de-brief interview all participants indicated that the usability was generally very high, and no usability errors were recorded, which confirms the high ratings in the subjective usability evaluation.

Look and Feel

The measure of classical-aesthetics, expressive aesthetics and pleasurable interaction as assessed by the perceived aesthetics scale [Lavie and Tractinsky 2004] were entered as dependent variables into a MANOVA with interface-style (2) as between-subjects factor. The multivariate effect of interface style was significant $F_{(3,35)} = 5.76$; $p < .01$, showing that participants were more positive in evaluating the Mobizines interface. All simple effects were also significant, namely $F_{(1,37)} = 6.05$; $p < .05$ for classical aesthetics, $F_{(1,38)} = 18.81$; $p < .001$ for expressive aesthetics and $F_{(1,38)} = 7.00$; $p < .05$ for pleasurable interaction.

Seven items of a Likert scale (appeal on emotional level, evokes positive feelings, pleasurable to look at, pleasurable interacting with it, interesting personality, has design features I like, conveys positive image) were used to measure participants' evaluation of the interface look and feel after all experimental conditions. A 2 * 2 ANOVA with interface-style (2) as between-subjects factor and content-fit (2) as within-subjects factor showed a significant main effect for content-fit ($F_{(1,36)} = 17.73$; $p < .001$) and for interface-style ($F_{(1,36)} = 14.03$; $p < .01$), with no interaction. The customised version was preferred (mean = 4.70; SD = 1.24) to the generic version (mean = 4.32; SD = 1.18), and the Mobizines interface was perceived as better (mean = 5.09; SD = 1.27) than the WAP interface (mean = 3.86; SD = .80).

Content-Fit and Customisation

The content-fit manipulation was assessed with a 4-item content-quality scale and a 4-item scale to assess customisation. The customised version was rated as being significantly more customised ($F_{(1,26)} = 46.69$; $p < .001$) and as providing better content ($F_{(1,38)} = 30.44$; $p < .001$), as assessed by two 2 * 2 mixed-model ANOVAs with interface-style (2) as between-subjects factor and content-fit (2) as within-subjects factor, and content-quality / customisation as dependant variables.

Lavie and Tractinsky's [2004] service quality scale was only assessed between subjects (WAP- vs. Mobizines interface) and showed no significant differences.

Overall Preference

At the end of the study, participants were asked to indicate their preference between the customised and the generic version of the system. They were also briefly shown the interface variant that they had not used and asked to indicate their favourite one. Finally, they were given a trade-off question in which participants had to sacrifice one of their preferences from the previous two questions.

In the first overall-preference choice question, all participants preferred the customised version over the generic version. Participants indicated that configuring the application had provided them with content that was more relevant and interesting to them, while the cost of configuration was minimal, so they perceived customisation to be considerable gain in content-quality for comparatively neglectable cost. Between the two interface styles, 92.5% people chose the Mobizines interface (only 3 people did not - 7.5%), indicating that they preferred the Mobizines' look-and-feel to the plain WAP style. The 3 participants who preferred the WAP style cited simplicity of the user interface as their reason.

Thus, there is a clear preference for both the customised application version with the better content-fit as well as for the Mobizines interface with the higher aesthetics. The final trade-off choice question produced a 50:50 split between people who would sacrifice customisation for the more aesthetic version and people who would sacrifice the aesthetics to keep customisation.

Participants' answers in the trade-off question were used as a between-subjects factor (trade-off: Aesthetics vs. Customisation) in two ANOVAs with customisation and look and feel scales as dependent variables. Content-fit (2: customised vs. generic) and interface-style (2: Mobizines vs. WAP) were also entered in the factorial design.

The analysis on customisation indicated that participants who preferred customisation over aesthetics in the trade-off question judged the effect of the customisation more positively ($F_{(1,22)} = 6.15$, $p < .05$) than the participants who would sacrifice customisation for the more-aesthetic interface. The analysis on look and feel returned a significant main effect for interface-style ($F_{(1,33)} = 21.67$; $p < .001$), trade-off preference ($F_{(1,33)} = 6.15$; $p < .05$) and a significant interaction interface-style * trade-off preference ($F_{(1,33)} = 6.38$; $p < .05$).

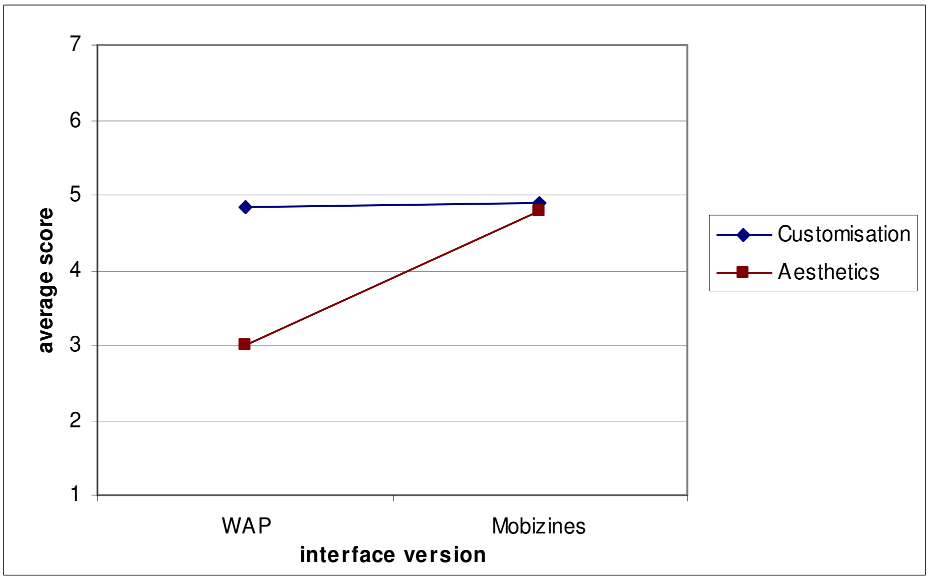


Fig. 10: Mean rating of look-and-feel of WAP and Mobizines interfaces grouped by participants' trade-off preferences (Aesthetics / Customisation)

Further analyses of the interaction (Figure 10), showed that the judgement differs for the WAP-interface version, while the ratings for the Mobizines-interface version are nearly the same. Participants who chose aesthetics over customisation in the trade-off question rate the WAP interface significantly worse than the participants who would sacrifice a more aesthetic interface to keep customisation. It seems that users who chose aesthetics in the trade-off question seem to be more sensitive to the worse aesthetics of the WAP version.

5.4. Summary

The results of the study are summarised in table VI. The Mobizines-interface is rated as more aesthetic and more engaging than the WAP interface. The customised version is rated as being more customised (manipulation-check) and having better content. Participants spent more time reading in the customised condition (reading several paragraphs as compared to the minimum requirement of 1 paragraph). It is interesting that the customised version is also rated higher in look and feel and usability. There seems to be a bias in quality-judgement favouring the overall preferred version in all quality-dimensions which reflects results from previous studies (see section 4 and [Hartmann et al. 2007]). It has to be noted, however, that the effect is only evident in the within-subjects comparison, when participants actually experienced both versions. It did not emerge in the between-subjects comparison (Mobizines vs. WAP).

Table VI. Results summary for the Mobizines study (+ higher; - lower; = no sig. diff.).

	aesthetics manipulation		content-fit manipulation	
Measures	**Mobizines**	**WAP**	**customised**	**generic**
Aesthetic classic	+	-	n/a	n/a
Aesthetic expressive	+	-	n/a	n/a
Engagement	+	-	n/a	n/a
Usability	=	=	= (+)	= (-)
Look and Feel	+	-	+	-
Customisation	=	=	+	-
Content	=	=	+	-
Overall Preference	**+**	**-**	**+**	**-**

Overall preference for the more aesthetic Mobizines interface (92.5%) and the customised version with better content (100%) were clear. The results of the trade-off question, where participants had to sacrifice either their favourite interface-style to keep customisation or sacrifice customisation for their favourite interface-style, showed a 50:50 split, consistent with results from previous studies (see section 4). The relative importance assigned to each of these criteria was affected by individual differences. People who chose to keep aesthetics in favour of customisation were more sensitive to variations in the look and feel dimension, whereas people who preferred customisation were more sensitive towards variations in the customisation / content dimension.

Model of Quality Judgement

The findings of the Mobizines, History and Astronomy site studies and a further comparison between three academic websites from Stanford University which had similar content but different UI styles [Hartmann et al. 2007] are summarised in table VII

Table VII. Summary of the studies.

	Findings supported by			
	Astronomy	History	Stanford	Mobizines
Metaphor / Animated Design better Aesthetics	√	√	√	√
Metaphor / Animated Design preferred overall	√	=	√	√
Halo effect usability +ve content quality	√	√	X	X
Menu Design better usability	√	√	N/A	N/A
Framing effect by scenario	√	√	√	N/A

From the experiments we believe that the process of judgement follows the model illustrated in Figure 11. This elaborates the Judgement and Decision Making process in the initial theoretical framework (see Figure 1) by proposing 3 stages in user judgement following the sequence of user experience from initial encounter with the application to hands on interaction. On initial encounter users assess the application according to their goals and the task domain. This stage will correspond to searching and locating a web site or purchasing a software product. Next the user's goals and task influence the selection of the decision making criteria. For example in applications where serious use with more critical outcomes is expected content and usability will be favoured. Alternatively, for less serious use (e.g. entertainment, games), aesthetics and engagement will be favoured. The dominant criterion or criteria become intentions against which the user's experience will be judged. The decision making process is iterative as users modify their opinions as their experience progresses. Experience is evaluated resulting in positive or adverse attitudes and memory which feedback on the criteria, for example in the History and Astronomy websites memory was more favoured for the menu version with better usability. However the feedback involves a complex interaction between the dominant and non-dominant criteria, task and user background that influences overall preference, as illustrated by the participants' judgement when split into their preference groups in the History site and Mobizines studies. One example of the interaction between the criteria was the finding that design versions with better usability promoted high quantities and more positive content memory while more engaging designs promoted better memory of UI details and interactive metaphors.

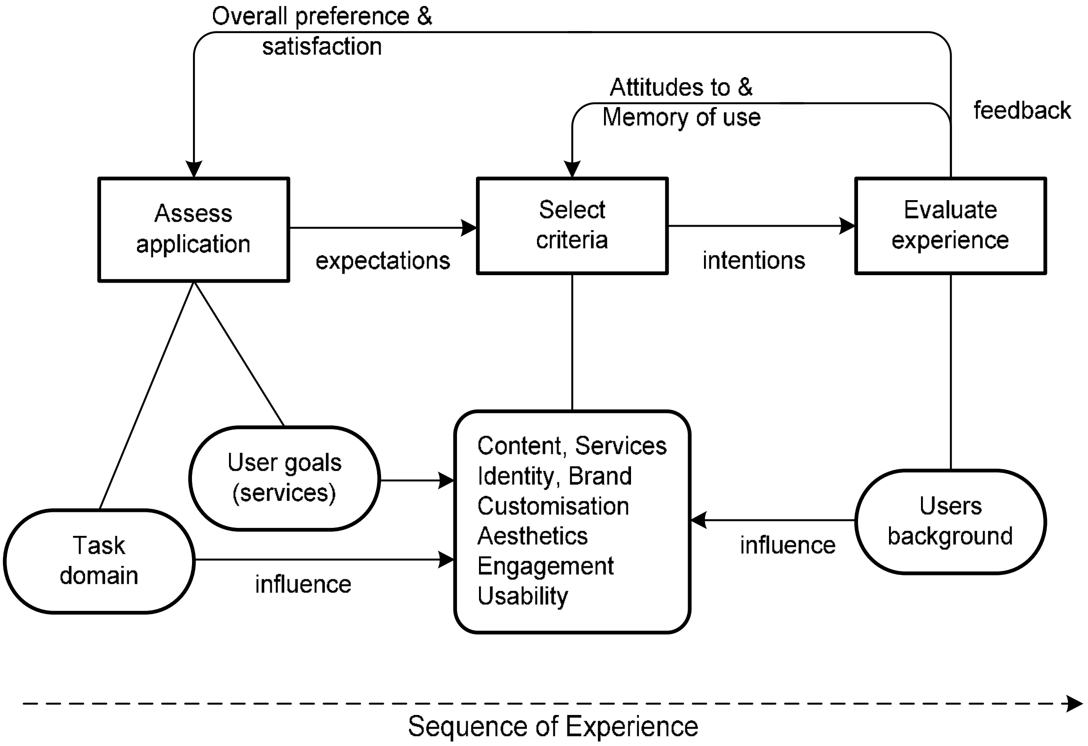


Fig. 11. Model of the users' decision making process for UI quality assessment.

The task framing effects demonstrate that people can recast their choice and overall preferences, as demonstrated by the radical changes in users' judgement between the design styles we found in all the studies when explicitly different scenarios of use were given. Interaction between the criteria, or the halo effect, is also influenced by the task and criticality of the application. If serious usability-oriented criteria are invoked then content is judged more favourably, with overall preference; if less serious use is selected, then positive aesthetics appear to override poor usability to influence the overall preference. Customisation is also task-dependent and if it has been selected as important, it appears to have a positive influence on content, as demonstrated in the Mobizines study. If selected, customisation also appears to dominate over aesthetics in determining overall preference from the trade-off choices made by users between the Mobizines versions. User background also interacts with the task framing effect, making some users less likely to change their minds, as we found when design students' preference remained loyal to a Design School website [Hartmann et al. 2007].

In the regression analyses, aesthetics / engagement and usability were the major factors accounting for users' quality judgements and determining overall preferences; while content was less important although this criterion was the same for both design versions in our experiments. The direction of the influences between the criteria: usability to content, aesthetics to overall preference, are supported by users' comments and qualitative evidence; however, we can not be sure of the causal sequence. It may be that overall preference is produced first and then the quality criteria are retro-fitted to justify the decision, as demonstrated by the task framing when people recast their preferences and discard evidence which does not support their more holistic judgement. Alternatively, we believe the users' judgement of experience in light of the selected criteria determines overall preference, but further time series experiments are necessary establish this conjecture. In terms of ADM theory, users appear to use a trade-off strategy by weighting different attributes of the design, with the weightings being driven by the task, and to an extent by the users' background.

6. DISCUSSION AND CONCLUSIONS

Our results show that the link between aesthetics and usability reported by Tractinsky's studies [Tractinsky 1997; Tractinsky et al. 2000] is more complex that than the claim that "what is beautiful is usable", although our findings partially support the reported correlation between perceived usability and aesthetics [Tractinsky 1997; Tractinsky et al. 2000; Kim et al. 2003]. In the websites we studied, the more graphical metaphor-based designs were perceived as having better expressive aesthetics using Lavie and Tractinsky's [2004] scale, yet these designs had worse objective and perceived usability. We believe that positively perceived aesthetics shows a halo effect that overrides users' poor usability experience to influence overall preference towards the more engaging, metaphor-based designs. This suggests that aesthetics could be an important determinant of user satisfaction and system acceptability, overcoming poor usability experience and even positively influencing content. This agrees with Tractinsky's experiments; however, Hassenzahl [2004] produced contradictory results in his study on MP3 skins, with the observation that the more beautiful designs did not have better perceived usability. An explanation for these apparent contradictions may lie in different interpretation of Hassenzahl's concept of 'pragmatic' product qualities. Since pragmatic qualities are related to the users' goal we expect that these correspond to our services / content factor, which we found to be positively influenced by good usability. We agree with Hassenzahl that aesthetics is more closely related to judgements of product identity, and may become the dominant factor when content of competing products is similar, even leading users to forgive adverse usability experiences.

Even though the metaphor designs had superior expressive aesthetics, there was evidence of poor graphical design in the usability measures, particularly in the first two experiments. The metaphor designs were rated better on engagement and were preferred when the framing scenario pointed to more playful applications; in contrast, the more conservative menu-based versions were preferred for more serious scenarios of use. Our explanation of these apparent contradictions is that users' judgement was influenced by their expectations of use, and users are selective with their experience when making choices. In the absence of task framing, a more general attitude prevails reflecting the overall design concept, favouring the metaphor design versions with engaging interaction based, we conjecture, on the users' default assumption that the History and Astronomy websites should be entertaining. When the scenario of use is more serious, users' attitudes are based more directly on usability experience and design features, reflected in the adverse usability and classical aesthetics ratings for the metaphor designs. The changes in judgement can be explained by adaptive strategies suggested in ADM theory. In less critical decisions, users choose designs by filtering choices based on the general impression of aesthetics and engagement; in contrast, for critical serious use, it appears that usability experience is actively considered and general impressions discarded. In serious use, usability can have a positive halo effect on content. In less critical choice, aesthetics has a halo effect in overriding the adverse usability influences, in agreement with Tractinsky's findings [Tractinsky et al. 2000] that more aesthetic ATM designs were also perceived to be more usable. However, in Tractinsky's experiments the usability effect was small (mainly response times), whereas our findings demonstrate the strength of the effect. The history and astronomy metaphor designs experienced severe problems and much worse usability ratings.

Users' memory also supports this interpretation for adaptive decision strategies depending on the task context. The metaphor designs were remembered for their interaction and multimedia appeal and for usability problems, while information content was remembered for the more serious menu design. While we interpret users' judgements in terms of attitude (i.e. valenced memory), one could also view their judgements as emotional reactions to the designs at the visceral (general) and reflective (detailed) level [Norman 2004], although we had little evidence of affective reaction in our qualitative data.

When we asked users to rank the importance of the factors in our framework for general product choice, their order was content > usability > aesthetics > customisation; however in the same study aesthetics became the dominant factor in less serious scenarios of use, while usability and content were dominant for more serious usage scenarios [Hartmann et al 2007], consistent with the first two studies reported in this paper. The importance rankings indicate that contents and services, and implicitly utility, are the dominant factors. Aesthetics and usability vied for second and third place in the importance rankings [Hartmann et al. 2007]. This study suggests that the relative importance of these two factors is sensitive to the users' goals or tasks and the users' background. When we tested separate groups of technology and design students, the designers were more resistant to changing preferences away from the more aesthetically design site in all scenarios [Hartmann et al. 2007]. When customisation was manipulated to become the dominant factor, this not surprisingly also positively influenced content since content was the subject matter of the customised designs. Our manipulation of customisation provided a low cost / high reward trade off in agreement with the principles for personalisation proposed by Blom and Monk [2004], so the users' preference for the customised version was not surprising. Similarly the aesthetic manipulation influenced users' preferences; however, the trade off choice indicated that our participant population seemed to divide between aesthetics or customisation as the 'dominant' variable, as manifest in the adverse judgement of the non-favoured version on look and feel. Since we also found this effect in the history site (i.e. metaphor preference participants were more critical of aesthetics on menu versions), we conjecture that we may need another component in our theory - an individual trait or experiential predisposition towards judgement criteria, in particular aesthetics. In their classic study, Dion et al. [1972] showed that more positive traits were ascribed to attractive individuals, compared to less attractive ones. An interesting question for further research is the relative strength of halo effects from different variables. From participants' comments in our studies we speculate that the concept of interaction (via metaphors, animations, pop-ups, etc.) might make the "feel-good factor" more important than the "look" of visually aesthetic design.

We have made a small advance in measures of aesthetics by introducing related phenomena of interaction and engagement with a rigorous evaluation methodology. We showed that users have a strong preference for customisation, although the trade-off between aesthetic preference and customised content showed a mixed result, suggesting difference preferences within a user population. The model of user judgement we proposed to account for these contextual effects is consistent with Hassenzahl's studies and theory [2002, 2003, 2004] of hedonic and pragmatic judgement of product quality which also posits a contextualising phase when users' expectations are set by the product genre and context of use. Our study has exposed conflicting opinions held by our users, combined with the doubts raised by Hassenzahl [2004] concerning the link between aesthetics and usability indicates that the measures of classical and expressive aesthetics [Lavie and Tractinsky 2004] may have to be re-assessed.

The quality judgement framework we propose gives a more comprehensive view of design quality and extends a range of subjective measures of emotive and aesthetic factors [Hassenzahl 2000; Hallnäs and Redström 2002]. One implication for our findings and model of quality judgement is to reinforce the well known advice "know your audience" [Mullet and Sano 1995; Spool et al. 1999; Lynch and Horton 2002] but to refine it as "know your audience's preferences and expectations". We argue that design priorities for aesthetics, usability, content or other components in our attractiveness framework should be matched to the user profile and application domain. Another implication is that targets and goals in user experience evaluation methods will need to be set according to the users' expectations, while investment decision about whether to improve usability or concentrate on aesthetic design should consider the product concept and application domain. In conclusion, we have demonstrated that while aesthetics is an important component of design quality, perception of aesthetics is susceptible to the user's background and task. Usability is important, but good aesthetic design can overcome some usability problems. The strength of the halo effect needs further research; for example, if usability problems are severe the halo effect from favourable aesthetics may evaporate if users lose trust in the system. In our future work we intend to extend and refine the evaluation approach we have developed to further understand the relative strengths of different influences on users' overall quality judgement, and investigate the interactions between them. One direction is to elaborate the attractiveness heuristics we used in previous studies [Sutcliffe 2002a, 2002b] that attempt to link general impressions of aesthetics to assessment of specific interactive design features. Another area for further research is extending the concept of attractiveness to encompass interaction, and engagement. The designs we evaluated showed that interactive metaphors were engaging; however, engagement may well be influenced by other factors such as the flow of interaction [Csikszentmihalyi 2002] and the sense of presence and immersion experienced by the user. We will refine our evaluation instruments to measure the relative contribution of interactive as well as presentation aesthetics in the future.

REFERENCES

BERNIER, M. J. 1996. Establishing the psychometric properties of a scale for evaluating quality in printed educational materials. *Patient Education and Counseling*, 283-299.

BLOCH, P. 1995. Seeking the ideal form: Product design and consumer response. *Journal of Marketing 59*, 16-29.

BLOM, J., AND MONK, A. 2003. Theory of personalization of appearance: Why users personalize their PCs and mobile phones. *Human-Computer Interaction 18*(3), 193-228.

CSIKSZENTMIHALYI, M. 2002. *Flow: The classic work on how to achieve happiness* (Revised ed.). Rider, London, UK.

DE ANGELI, A., LYNCH, P., AND JOHNSON, G.I. 2002. Pleasure versus efficiency in user interfaces: Towards an involvement framework. In *Pleasure with products: Beyond usability.* W. S. GREEN, AND P. W. JORDAN Eds. Taylor and Francis, London, UK, 97-111.

DE ANGELI, A., SUTCLIFFE, A.G., AND HARTMANN, J. 2006. Interaction, usability and aesthetics: What influences users' preferences? In *Proceedings* of *DIS 06*, *Designing Interactive Systems.* ACM Press, New York.

DION, K., BERSCHEID, E., AND WALSTER, E. 1972. What is beautiful is good. *Journal of Personality and Social Psychology 24*, 285-290.

DJAJADININGRAT, J.P., OVERBEEKE, C. J., AND WENSVEEN, S. A. G. 2000. Augmenting fun and beauty: A pamphlet. In *Proceedings of Designing Augmented Reality Environments, DARE 2000, Elsinore, Denmark*, 131-134.

HALLNÄS, L., AND REDSTRÖM, J. 2002. From use to presence: On the expression of aesthetics of everyday computational things. *ACM Transactions on Computer-Human Interaction*, *9*(2), 106-124.

HARTMANN, J., SUTCLIFFE, A.G., AND DE ANGELI, A. 2007. Investigating attractiveness in web user interfaces. In *Proceedings of CHI-07*, *Conference on Human Factors in Computing Systems,* San Jose, CA, USA. ACM Press, New York, NY, USA.

HASSENZAHL, M. 2002. The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction 13*, 479-497.

HASSENZAHL, M. 2003. The thing and I: Understanding the relationship between user and product. In *Funology: From usability to enjoyment* M. BLYTHE, C. OVERBEEKE, A.F. MONK, AND P.C. WRIGHT Eds. Kluwer, Dortrecht, 31-42.

HASSENZAHL, M. 2004. The interplay of beauty, goodness and usability in interactive products. *Human-Computer Interaction*, *19*(4), 319-349.

HASSENZAHL, M., PLATZ, A., BURMESTER, M., AND LEHNER, K. 2000. Hedonic and ergonomic quality aspects determine a software's appeal. In T. TURNER, G. SZWILLUS, M. CZERWINSKI, AND F. PATERNO, Eds. In *Proceedings* of *CHI 2000, Conference on Human Factors in Computing Systems,* The Hague 1-6 April 2000. ACM Press, New York, NY, USA, 201-208

IBM. 2000. Ease of use: Design principles. http://www.ibm.com/ibm/easy/eou_ext.nsf/Publish/6. (20 November 2000).

ISO. 1997. *ISO 9241: Ergonomic requirements for office systems with visual display terminals (VDTs).* International Standards Organisation, Geneva, Switzerland.

IVORY, M., AND HEARST, M. 2001. The state of the art in automated usability evaluation of user interfaces. *ACM Computing Surveys*, *33*(4), 173-197.

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.

LINDGAARD, G., AND DUDEK, C. 2003. What is this evasive beast we call user satisfaction? *Interacting with Computers*, *15*(3), 429-452.

LYNCH, P.J., AND HORTON, S. 2001. *Web style guidelines* (2nd ed.). Yale University Press, New Haven CT.

MCCARTHY, J., AND WRIGHT, P. 2005. *Technology as experience*. MIT Press, Cambridge, MA, USA.

MEINERS, M.L., AND SHEPOSH, J.P. 1977. Beauty or brains: Which image for your mate? *Personality and Social Psychology 3*, 262-265.

MERRILEES, B., AND FRY, M.L. 2002. Corporate branding: A framework for e-retailers. *Corporate Reputation Review 5*, 213-225.

MULLET, K., AND SANO, D. 1995. *Designing visual interfaces: Communication oriented techniques*. SunSoft Press, Englewood Cliffs, NJ, USA.

NIELSEN, J. 1993. *Usability engineering*. Academic Press, Boston, MA, USA.

NIELSEN, J. 2000. *Designing web usability: The practice of simplicity*. New Riders, Indianapolis, IN, USA.

NIELSEN, J., AND MOLICH, R. 1990. Heuristic evaluation of user interfaces. *SIGCHI Bulletin* (April: special issue), 249-256.

NORMAN, D.A. 2004. *Emotional design: Why we love (or hate) everyday things*. Basic Books, New York, NY, USA.

OVERBEEKE, K.C., DJAJADININGRAT, J.P., HUMMELS, C.C.M., AND WENSVEEN, S.A.G. 2002. Chapter 7. In *Pleasure with products: Beyond usability*, W.S. GREEN, AND P.W. JORDAN Eds. Taylor and Francis, London, UK, 97-111.

PARK, S., CHOI, D., AND KIM, J. 2004. Critical factors for the aesthetic fidelity of web pages: Empirical studies with professional web designers and users. *Interacting with Computers*, *16*(2), 351-376.

PAYNE, J.W., BETTMAN, J.R., AND JOHNSON, E.J. 1993. *The adaptive decision maker*. Cambridge University Press, Cambridge, UK.

PERLMAN, G. 1997. *Practical usability evaluation*. In Proceedings of *CHI 97, Conference on Human Factors in Computing Systems*, Atlanta GA 22-27 May 1997. ACM Press, New York, NY, USA, 168-169

PETERSEN, M.G., IVERSEN, O.S., KROGH, P.G., AND LUDVIGSEN, M. 2004. Aesthetic interaction: A pragmatist's aesthetics of interactive systems. In *Proceedings of DIS 2004 Conference on Designing Interactive Systems: processes, practices, methods, and techniques*. ACM Press, New York, NY, USA.

REEVES, B., AND NASS, C. 1996. *The media equation: How people treat computers, television and new media like real people and places*. CLSI/Cambridge University Press, Stanford CA/Cambridge, USA.

SCHALLER, D.T., ALLISON-BUNNELL, S., CHOW, A., MARTY, P., & HEO, M. 2004. To FlashFlash or not to FlashFlash? Usability and user engagement of HTML vs. FlashFlash. *Proceedings: Museums and the Web 2004 International Conference. Available on-line at http://www.eduweb.com/ToFlashFlashornot.pdf.*

SHUSTERMAN, R. 1992. *Pragmatist aesthetics, living beauty, rethinking art*. Blackwell, Oxford, UK.

SPOOL, J.M., SCANLON, T., SNYDER, C., SCHROEDER, W., & DE ANGELO, T. 1999. *Web site usability: A designer's guide*. Morgan Kaufmann, San Francisco, CA, USA.

SUTCLIFFE, A.G. 2002a. Assessing the reliability of heuristic evaluation for website attractiveness and usability. Proceedings of *HICSS-35: Hawaii International Conference on System Sciences,* Hawaii, 7-10 January 2002. IEEE Computer Society Press, Los Alamitos, CA, USA, 1838-1847.

SUTCLIFFE, A.G. 2002b. Heuristic evaluation of website attractiveness and usability. In *Proceedings: 8th Workshop on Design, Specification and Verification of Interactive Systems, Glasgow 13-15 June 2001*. Springer-Verlag, Berlin, Germany, 188-199.

SUTCLIFFE, A.G., AND DE ANGELI, A. 2005. Assessing interaction styles in web user interfaces. In Proceedings of *INTERACT 2005, Human Computer Interaction,* Rome, 2005. Springer Verlag, Berlin, Germany, 405-417.

TRACTINSKY, N. 1997. Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. In S. PEMBERTON (Ed.), *Human Factors in Computing Systems: CHI 97 Conference Proceedings, Atlanta GA 22-27 May 1997*, ACM Press, New York, NY, USA, 115-122.

TRACTINSKY, N., SHOVAL-KATZ, A., AND IKAR, D. 2000. What is beautiful is usable. *Interacting with Computers*, *13*(2), 127-145.

TRACTINSKY, N. AND ZMIRI, D. 2006, Exploring attributes of skins as potential antecedents of emotion in HCI. In *Aesthetic computing*. P. FISHWICK Ed. MIT Press, Cambridge, MA, 405-521.

TURNER, J. C. (1987). Rediscovering the social group: A self-categorization theory. Basil Blackwell, Oxford, UK.