Running head: VISUAL APPEAL OF WEB PAGES

Judging Web Page Visual Appeal

Gary J. Fernandes, B.A. Hons.

A thesis submitted to
the faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of
Master of Arts

Department of Psychology

Carleton University

Ottawa, Ontario

January 2003

© 2003, Gary J. Fernandes

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-79669-8

Canada

The undersigned recommend to the
Faculty of Graduate Studies and Research
Acceptance of the thesis:

"Judging Web Page Visual Appeal"

submitted by

Gary Fernandes

in Partial fulfillment of the requirements for

the degree of Master of Arts

Chair

Thesis Supervisor

Carleton University
January, 2003

Abstract

The objective of this research was to determine the properties and characteristics related to web page visual appeal. In the first of three studies, 22 participants individually viewed 100 web pages for 500 ms in sequence and rated each page for visual appeal. Based on their responses, the 25 most and 25 least appealing web pages were selected for use in Study 2 with 31 participants. Reliability of judgments was very high. After making visual appeal ratings to the 50 web pages, participants rated each web page on seven visual characteristics. Five of the 7 rated characteristics were very highly correlated with visual appeal. In Study 3, web page design experts rated the 50 web pages from Study 2 on numerous properties related to color, layout, graphics, and fonts. Properties that correlated with visual appeal are discussed. During the process of this research, a technique for assessing web page visual appeal was developed.

Table of Contents

	Page
Abstract	ii
Table of Contents	iii
List of Figures	vi
List of Tables	vii
List of Appendices	viii
Introduction	1
Method	5
General Issues for Studies 1 and 2	5
Participants	5
Apparatus	6
Hardware	6
Software	6
Materials	6
Procedure	7
Study 2	8
Study 3	9
Participants, Apparatus, Materials, and Procedure	9
Results	10
Study 1	10
Study 2	13
Reliability of Visual Appeal Ratings	13

	Page
Validation with Subsamples	16
Appeal Score Comparisons between Studies 1 & 2	20
Relation between Appeal and Other Visual Characteristics	21
Complexity and Appeal	22
Study 3	24
Inter-rater Readability	24
Relation between Visual Appeal Ratings and Individual Visual Properties	26
Relation between Visual Appeal and Combinations of Visual Properties	28
Understanding Visual Properties	29
Discussion	29
Previous Research	29
Expert Agreement on Visual Properties in Web Sites	31
Visual Appeal Assessment Technique	32
Reliability of Appeal Judgments	32
Overall Means as a Measure of Visual Appeal	33
Eye-Tracking	33
Stimulus Duration	34
Timing Accuracy	35
Scale Anchors	36
Visual Appeal and Complexity	37
A Tool for Assessing Visual Appeal	38
A Tool for Assessing First Impressions of Trust and Perceived Usability	41

	Page
Aesthetics and Usability Research	42
Usage Log Analysis	43
Cross-Cultural Studies	44
References	45
Appendix A: Task Flow for Study 1	50
Appendix B: Task Flow for Study 2	51
Appendix C: Properties rated by graphic design experts in Study 3	52
Appendix D: Squared correlations between visual appeal ratings in two test phases for participants in Study 1.	57
Appendix E: Squared correlations between visual appeal ratings across test phases for participants in Study 2.	59
Appendix F: Subsample Validation	61
Appendix G: Principal Components Analysis of Web Page Characteristics	72

List of Figures

Figure	Description	Page
1	Visual appeal rating scale	7
2	Web page characteristic rating scales	9
3	Scatter plot of mean visual appeal scores for Study 1	12
4	Frequency histogram of rating correlations between phases for Study 1	13
5	Scatter plot of mean appeal scores for Study 2	15
6	Frequency histogram of rating correlations between phases for Study 2	16
7	Frequency histogram of correlations between individuals and sample	17
8	Mean squared correlations of ranks between smaller and total samples	19
9	Scatter plot of mean appeal scores for Study 1 vs. Study 2	21
10	Scatter plot of appeal vs. complexity	23

List of Tables

Table	Description	Page
1	Squared Correlations between Properties and Visual Appeal	28
2	Coefficients for Property Variables Predicting Appeal	29

List of Appendices

Appendix	Description	Page
A	Study 1 Task Flow	50
B	Study 2 Task Flow	51
C	Study 3 Expert-Rated Properties	52
D	Study1 Test Phase Correlations	57
E	Study 2 Test Phase Correlations	59
F	Subsample Validation	61
G	Principal Components Analysis for Web Page Properties	72

Judging Web Page Visual Appeal

The purpose of this research was to identify the web design features that make web pages visually appealing and unappealing to provide information that may assist web designers to design more visually appealing web pages. Traditionally, practitioners in the field of human-computer interaction have sought to enhance the user experience by improving system productivity, focusing on system functionality and usability while ignoring affective components (Karat, 2003), although some usability practitioners have considered affective components of user interface design and their impact on the user experience (e.g., Hofmeester, Kemp, & Blankendaal, 1996; Weeks, 1997; Jordan, 1998; Kim & Moon, 1998; Jennings, 2000; Hassenzahl, Beu, & Burmester, 2001). The most salient example of usability practitioners giving attention to affective components in design is a book being prepared by Donald Norman tentatively titled *Emotion & Design* (Norman, 2002).

Interface aesthetics is one dimension of the affective component of system design. There are numerous benefits to visually appealing web pages. Aesthetic factors can influence the user's experience of a system (Jordan, 1998). Improving the user's experience and satisfaction with a web site is important because they may affect whether or not people return to the site.

Visually appealing web pages may provide users with good first impressions (Basso, Goldberg, Greenspan, & Weimer, 2001). Aesthetic factors are also important because they can be used to project a desired image to people browsing web sites. Web page visual appeal can affect web site branding potential which is "the ability of a firm to repeatedly attract people to its web site" (Simeon, 2001, p. 420). Branding potential, and

therefore visual appeal, has obvious financial implications for e-commerce sites. Also, aesthetics may influence users' perceptions of other site characteristics that influence satisfaction such as perceptions of usability (Weimer, 1995). Karvonen (2000) stresses that aesthetics may influence trust, a crucial point for e-commerce sites, because of its implications for customer loyalty (Kim & Moon, 1998). Aesthetics may influence perceived reliability of an online vendor, especially if the potential customer is unfamiliar with the vendor (Basso et al., 2001).

Historically, aesthetics has been defined in various ways. Its Greek etymon is philosophical and deals with the sensory perception of tangible things rather than cognition or thinking about immaterial things (Simpson & Weiner, 1989). A more modern definition describes aesthetics as the "study or theory of beauty" (Neufeldt & Guralnik, 1997, p. 21). In terms of products, Brunel (1998) suggests that aesthetics is about "overall beauty or attractiveness" (p. 16). Veryzer (1999) mentions the widely held notion that there is an emotional component to the appreciation of aesthetics when he observes that an aesthetic response is an affective response.

Visual appeal is the only aspect of web aesthetics that is considered in this thesis. Usability, usefulness, or any other web page characteristics are not the focus of this research. By *visual appeal* I am referring to how attractive (Brunel, 2000) a web page looks. Visually appealing refers to visual attractiveness over visual repulsiveness.

There are various factors that may influence the visual appeal of web pages. The use of color is one example. In a study conducted for his doctoral dissertation, Knutson (1998) found that participants rated web pages with color as more "visually pleasing" (p. 31) than web pages without color. Participants in a study by Lindgaard (1999) preferred

pastels rather than dark colors for web page backgrounds. Other examples of how color could influence visual appeal of web pages are the number of colors used, types of color combinations employed, and the use of color gradients.

The font types and sizes used on web pages may also affect visual appeal. Bernard, Mills, Peterson, and Storrer (2001) used semantic differential scales to determine preferences for fonts on web pages. They compared 12 fonts that were categorized as either serif, sans serif, or ornate. They tested font appearance for personality, elegance, youthfulness, and business-likeness. Participants identified the Courier and Times fonts as most businesslike, possibly making these fonts appropriate for sites such as online banking sites. In a follow-up study, Bernard, Lida, Riley, Hackler, and Janzen (2002) tested the “perceived attractiveness” of certain web page font types and sizes. Participants rated the Georgia font as most attractive while Verdana was the most preferred.

The use of pictures or imagery may also affect visual appeal of web pages. Participants in a study by Haig and Whitfield (2000) gave higher aesthetic ratings to a web site with images than a site without images. Knutson (1998) also found the presence of pictures on web pages to be related to ratings of visual appeal. Schenkman and Jonsson (2000) found that web pages that had more illustrations than text were rated as being more beautiful. They also found that ratings of web page complexity and beauty were unrelated.

Layout is obviously a factor that could affect visual appeal. For example, Tractinsky, Katz, and Ikar (2000) used layout as the sole aesthetic manipulation in a study examining the relation between actual usability, perceived usability, and aesthetic

performance of ATM displays. There were aesthetic preferences for certain layout patterns over others. Ngo, Samsudin, and Abdullah (2000) suggest that components of layout such as balance (centered screen layout), equilibrium (balance between opposing screen elements), symmetry (balanced distribution of equivalent elements about a common line), sequence (layout that facilitates eye movement from upper left to lower right), and order and complexity of web pages may affect visual appeal. These researchers provide mathematical models for maximizing aesthetic appeal on each of these aesthetic dimensions.

Knutson (1998) examined whether the presence or absence of web page background textures and the presence or absence of a 3-dimensional look affected web page visual appeal. He concluded that the presence of background textures has a positive influence on visual appeal while the influence of a three-dimensional rather than two-dimensional look on visual appeal was inconsequential.

Clearly, a variety of factors may affect the visual appeal of web pages. However, the extent to which particular factors are more important than others is unclear and how combinations of these factors affect visual appeal is also unclear. The proposed study seeks to elucidate which web design factors or combinations of factors are related to visual appeal.

Measuring web page visual appeal must take issues such as emotion and first impressions into account. The idea that emotion is an important factor in aesthetics has been discussed in the literature (e.g., Bamossy, Scammon, & Johnston, 1983; Veryzer, 1993; Veryzer, 1999). First impressions are also important to aesthetic judgments (for a discussion see Lindgaard & Dudek, in press). Veryzer (1999) has suggested that non-

conscious, and therefore non-cognitive, processes may mediate aesthetic preference. It may be important, therefore, when measuring visual appeal of web pages to take emotional, precognitive, and first impression considerations into account since these issues are relevant to aesthetic preference. For example, permitting participants only a brief period of time to view a web page before commenting on its visual appeal would allow for a first impression relatively uncontaminated by cognition about things unrelated to visual aesthetic appeal such as textual content.

This thesis is specifically concerned with the aesthetic issue of visual appeal in Internet web pages. Participants in the studies judged the visual appeal of web site home pages. A subset of highly visually appealing and unappealing web pages was compiled. Different participants made visual appeal ratings and used rating scales to comment on web page characteristics that influenced their appeal ratings. I looked for relations between visual characteristics and appeal ratings. In addition, trained designers went through the subset of highly visually appealing and unappealing web pages to identify physical design properties of the pages. I then looked for relations between these properties and appeal ratings.

Method

General issues for Studies 1 and 2

Participants

Participants were university students who were not colorblind (based on self report) and had normal vision (after correction, in some cases). They also reported that English was their first language. Twenty-two university students participated in Study 1.

Thirty-one students who did not participate in the first study participated in the second study. Each received a token monetary honorarium for participating.

Apparatus

Hardware. Each participant was tested with a computer workstation of the following configuration: 1.6GHz Athlon CPU, 256 Mbytes RAM, Matrox dual-head video card, Samsung SyncMaster 950p 19-inch monitor with a white balance calibrated at 9300° K and a gamma value of 2.1.

Software. A program created in Microsoft Visual Basic 6.0 specifically for this research was used to present stimuli and collect response data. The experimental program ran under the Microsoft Windows 2000 operating system.

Materials

The practice and test stimuli were web page screen shots of presumably varying visual appeal. Stimulus web pages were selected from a number of sources. Many web pages were solicited from members of the local HCI lab. Members were asked to submit links to web pages that they thought “looked really good or looked really bad.” They were also asked to submit pages that they did not think to be high-traffic sites; for example, <http://www.cnn.com/> or <http://www.amazon.com> would be unacceptable. Other web pages that were included were brought to my attention through email distribution lists such as UK-USABILITY. Web pages came from a variety of contexts including entertainment, e-commerce, and general information.

Screen shots of web pages were taken within an Internet Explorer 6 browser at 1024 x 768 pixel resolution in 32-bit true color. The Internet Explorer browser was selected because it is the most popular web browser (Kubaitis, 2002).

Procedure

Participants were tested individually. After reading a briefing form and signing an informed consent sheet, each participant was seated in front of a computer. They were encouraged to adjust their seating height and monitor angle to their preference.

Each participant saw the first stimulus web page for 500 ms followed by a white screen with a continuous rating scale. Participants were instructed to “rate the visual appeal of the web page” by clicking the appropriate location on the rating scale slider. The 101-point scale, shown in Figure 1, ranged from *Very Unattractive* to *Very Attractive*. The slider appeared in the middle of the rating scale at the start of each trial.

Figure 1. Visual appeal scale. Participants clicked on the line to make their ratings.

The program advanced to the next stimulus web page only after the participant had entered a response followed by a 1-s delay during which a blank, white screen was shown. This procedure was followed for all trials. Each session had a set of practice trials followed by 2 test phases (see Appendix A for a flow diagram). The 20 practice trials were presented in a fixed order for each participant. In the first test phase, each participant viewed web pages presented in a random order for each participant. In the second test phase, each participant viewed the same web pages for a second time in a newly randomized order. This served as a reliability check by enabling comparisons of responses to the same web page across the 2 test phases.

Study 2

Materials were the 25 most- and 25 least-visually appealing web pages from the first study. The only procedural difference between Studies 1 and 2 was an additional test phase in the second study which required the use of a dual-monitor setup (see Appendix B for a flow diagram). The second monitor was a 17-inch Samsung SyncMaster 753DF. During the added test phase, participants saw the 50 web pages, one at a time, on the left monitor and a form with rating scales on the right monitor. Web pages were displayed for as long as participants needed to complete the rating scales form shown in Figure 2. Participants used these scales to indicate the extent to which particular web page characteristics influenced their visual appeal ratings for each page.

Anchors for these scales were drawn from a number of sources. *Boring*, *imaginative*, *confusing*, *good use of color*, *bad use of color*, and *well designed* were taken from Tombaugh, Dillon, and Carboni (1982). These researchers did not use semantic differential scales when asking participants to identify UI characteristics; rather, participants in their study indicated which features did or did not apply. These terms were transformed into semantic differential scales for the current study as shown in Figure 2.

Figure 2. Web page characteristics. Participants made their responses by clicking on the lines.

Study 3

Participants, apparatus, materials, and procedure.

Participants were 8 visual design professionals. Seven were professional graphic designers while the remaining participant was a professor of industrial design. Requests for participation were sent out via email to local design professionals. The first to respond were selected as participants. Each professional received a token monetary honorarium for participating. A dual-monitor computer setup identical to the one employed in Study 2 was used. The materials were the same 50 web pages derived from Study 1 and used in Study 2. Another program was developed in Visual Basic for stimulus presentation and data collection. Participants were tested individually. The web page presentation order was randomly determined for each participant. Each participant saw a web page on the monitor on the left and a data collection instrument on the monitor to the right. Participants used the data collection instrument to identify web page

properties that fall into the following categories: color, layout, graphics, and fonts. (The web page properties that composed this data collection instrument were selected based on input from a professor of industrial design and a professional industrial designer.) Two graphics experts were randomly assigned to each of these 4 categories, examining all 50 web pages and identifying properties. Appendix C shows a list of visual properties and examples.

Results

Study 1

The Visual Basic computer program saved visual appeal ratings on a 101-point scale for each site. For each subject, visual appeal responses for web pages within each phase were transformed into z-scores. The mean of each web page's z-scores was calculated across the 2 phases for each participant. This mean provided a measure of visual appeal for each web page for each participant. Then, the medians of these means were calculated for each web page across participants. These medians provided a general visual appeal score for each web page. Next, web pages were ranked by their medians. The 25 least appealing and 25 most appealing web pages were set aside for use in Study 2.

A check was performed to determine if there was agreement among participants as to whether these 50 web pages were actually highly appealing or unappealing. Each web page was ranked according to its mean z-score calculated across both test phases for each participant. Then, I performed frequency counts of how often each web page fell in the top 50 and bottom 50 for each participant. The 25 most appealing web pages had to be ranked in the top 50 for at least half of the participants. The 25 least appealing web

pages had to be ranked in the bottom 50 for at least half of the participants. All the 25 most- and 25 least-appealing web pages selected for Study 2 met these criteria.

To check the reliability of participants' responses to the same web pages presented in the two test phases, bivariate correlations were calculated. For these correlations, the 100 web sites were cases. The mean raw score responses to each web page were averaged across all 22 participants for the first test phase and the mean raw scores to each web page were averaged across all participants for the second test phase. The correlation was taken between these mean visual appeal scores.

The squared¹ Pearson Product Moment correlation (r^2) was .94 ($p < .001$), indicating a strong relation between participants' mean visual appeal ratings for the same web pages across the 2 phases. The scatter plot of the data and regression line are shown in Figure 3 where each point represents a web site. Higher ratings indicate more visual appeal.

¹ Squared correlations are presented because they indicate the proportion of variance in the outcome variable associated with the predictor variable. For example, with a correlation of .97, 94 percent of the variance in the Phase 2 visual ratings can be accounted for by the visual appeal ratings in Phase 1 and vice versa.

Figure 3. Scatter plot of mean visual appeal scores for Study 1, Test Phase 1 vs. Test Phase 2 ($r^2 = .94$).

To determine the degree of reliability across test phases within participants, separate correlations were calculated between test phases for each participant. The cases were the 100 web site home pages. Figure 4 and Appendix D show the distribution of squared correlations for 22 participants. The mean of these 22 squared correlations was .64 with a standard deviation of .14. All correlations were statistically significant ($p < .001$). Only 4 of the 22 participants had squared correlations less than .50 while 11 had squared correlations greater than .70.

Figure 4. Frequency histogram of squared correlations between participants' web page appeal ratings in Phase 1 and Phase 2 of Study 1.

Study 2

Study 2 used the same visual appeal rating procedure as Study 1, but with the top 25 and bottom 25 web sites for rated visual appeal from Study 1.

Reliability of visual appeal ratings. Reliability of participants' ratings for the same web pages in Phase 1 and Phase 2 was assessed by calculating a bivariate correlation. For this correlation, the 50 web sites were cases. The mean raw score responses to each web page were calculated across all 31 participants for the first test

phase and the mean raw scores to each web page were calculated across all participants for the second test phase. The correlation was taken between these mean visual appeal scores. As in Study 1, there was a very strong relation between responses in the 2 phases ($r^2 = .97$, $p < .001$). The scatter plot and regression line are shown in Figure 5. As demonstrated by the two distinct clusters apparent in Figure 5, including the top- and bottom-25 web pages from Study 1 in Study 2 successfully eliminated judgments in the middle range.² Based on mean ratings, none of the visually appealing pages from Study 1 were rated as unappealing in Study 2 and none of the unappealing pages in Study 1 were rated as appealing in Study 2.

² Extreme clusters of data points tend to increase the squared correlation between variables. I also calculated squared correlations for appeal ratings to pages in the highly appealing and unappealing clusters across both test phases. The squared correlation for highly appealing web pages was .83 while the squared correlation for highly unappealing pages was .90.

Figure 5. Scatter plot of mean visual appeal scores for Study 2, Test Phase 1 vs. Test Phase 2 ($r^2 = .97$).

Reliability was high for individual participants. Bivariate correlations between test phases were calculated for each participant. Figure 6 and Appendix E show squared correlations for the 31 participants. The mean of the squared correlations was .71 with a standard deviation of .16. Again, all correlations were statistically significant (all $p < .001$). Only 3 of 31 participants had squared correlations below .50 while 18 had squared correlations at .70 or above.

Figure 6. Frequency histogram of squared correlations for participants' visual appeal ratings between Test Phases 1 and 2 of Study 2.

Validation with subsamples. Are overall visual appeal means (calculated across all participants over both test phases for each web page from Study 2) representative measures of web pages' visual appeal for *individuals*? I wanted to assess how well the overall visual appeal means represent the individual participant's scores that make up those means. To conduct this analysis, I calculated the correlations of visual appeal means averaged across both test phases for all 31 participants vs. the visual appeal means for each of the 31 participants. For each participant, a web page's visual appeal mean was the mean across both test phases. These squared correlations indicate the proportion

of variance in the web page means based on all participants that is shared with individual participants. The mean of these squared correlations was .65 with a range of .46 and a standard deviation of .14. Figure 7 shows the frequency distribution of the squared correlations between individual participants and the total sample. Cases for the correlations are web pages.

Figure 7. Frequency histogram of squared correlations for individual participants' visual appeal ratings with the mean appeal ratings based on all 31 participants of Study 2.

For example, five individuals had mean ratings, based on Phases 1 and 2, on the 50 web pages that had squared correlations with the total sample means for the same pages between .80 to .89. All correlations were significant at the .01 level or better.

Figure 7 shows that, of 31 squared correlations, only 4 squared correlations were below .50 while 14 were higher than .70. These findings indicate that most individuals reliably judge visual appeal and that the results are similar to those obtained with a much larger sample of 31 participants.

The squared correlation between Phases 1 and 2 for the total sample of 31 participants is .97 (Figure 5), the mean squared correlation for individuals is .71 (Figure 6) and the mean squared correlation between the total sample and individual samples is .65 (Figure 7). It must be the case that a smaller sample size, somewhere between $N = 1$ and $N = 31$, might be adequate to judge visual appeal with high reliability, perhaps as high as that obtained with a sample of 31. To get an estimate of the smallest possible sample size that would give results comparable to the total sample, I validated the visual appeal assessment technique in Study 2 in two ways.

The first way compared the mean ratings for each web page based on small, randomly selected subsamples of the data with results obtained with all 31 participants. Random samples of participants' appeal ratings from Phases 1 and 2 were taken from the complete set of 31 participants. There were subsamples from size $N = 2$ to $N = 10$. For each of these subsamples, mean appeal ratings were calculated for each web page averaged over Phase 1 and Phase 2. The web pages were then ranked based on these means. These rankings were then compared to web page rankings based on the mean appeal ratings calculated from all 31 participants across both Phases 1 and 2.

Bivariate correlations were calculated between web page ranks based on the complete sample and ranks based on the subsamples. For example, the correlation between ranks for web pages in the complete sample and web pages in the first

subsample with 10 participants was calculated. The same was done for web pages in the complete sample and the other subsamples. The mean squared correlations for all subsamples of each size are shown in Figure 8 and in Table F1 of Appendix F.

Figure 8. Mean squared correlations between subsamples and total sample for subsample sizes from 2 to 10 participants.

All squared correlations, shown in Table F1 in Appendix F, between the subsamples and the complete sample were statistically significant ($p < .001$), ranging from .74 to .94, indicating a very strong relation between each of the web page rank orders of the subsamples and the rank order based on the complete sample.

Of particular interest was whether the same web pages would be considered very appealing or unappealing in each of the subsamples as in the complete sample. Tables F2 – F5 in Appendix F show strong agreement between the visual appeal ranks of web pages for the total sample and the subsamples. Furthermore, this agreement in rank order among the ranked pages between the subsamples and the complete sample did not

decrease systematically as the sample size decreased from 10 to 5, but did decrease at smaller subsamples.

As all the tables in Appendix F show, the proportion of variance in the complete sample accounted for by the subsamples did not decrease systematically with sample size above $N = 4$. These analyses indicate that using sample sizes of as few as 5 participants may produce results similar to those of a larger sample size with 31 participants.

Appeal score comparisons between Studies 1 and 2. I wanted to determine if removing the 50 intermediate web pages from Study 1 affected visual appeal scores between Study 1 and Study 2. Correlations for mean visual appeal ratings for the 50 web pages in common between Study 1 and Study 2 were calculated. Cases were the 50 web pages used in both studies. The mean visual appeal ratings collapsed across Test Phases 1 and 2 for the two studies were the variables. There was a strong relation between visual appeal ratings in the two studies ($r^2 = .95$, $p < .001$). This relation is shown in Figure 7 where the vertical axis is mean rating in Study 1 and the horizontal axis is mean rating in Study 2.

Figure 9. Scatter plot of mean visual appeal scores, Study 1 vs. Study 2 ($r^2 = .95$).

Relation between appeal and other visual characteristics. The relation between perceptions of visual appeal and other web page characteristics was investigated. Web page visual appeal was determined by calculating the mean appeal scores for each web page. To determine the relation between a characteristic rated in the third test phase and a web page's visual appeal, Pearson Product Moment correlations were calculated, with web pages as cases, using mean web page visual appeal rating (collapsed over all participants and both test phases) and mean rating for each characteristic for that web page (collapsed over all participants).

There were very high correlations between mean rated visual appeal and mean ratings for 5 of the 7 characteristics: *interesting-boring* ($r^2 = .91$, $p < .001$), *good design-bad design* ($r^2 = .92$, $p < .001$), *good color-bad color* ($r^2 = .90$, $p < .001$), *good layout-bad*

layout ($r^2 = .88$, $p < .001$), and *imaginative-unimaginative* ($r^2 = .86$, $p < .001$). There was a moderate correlation between mean rated visual appeal and mean ratings for *clear-confusing* ($r^2 = .37$, $p < .001$) and a low correlation between mean rated visual appeal and mean ratings for *simple-complex* ($r^2 = .01$, $p > .80$).

While the previous results show how each individual visual characteristic was related to the visual appeal ratings, a multiple regression analysis was conducted to examine how the web page characteristic ratings were collectively related to visual appeal.³ Regressing mean visual appeal ratings on the 7 rated characteristics resulted in a very high squared multiple correlation, $R^2 = .94$, but it was barely higher than the largest of the bivariate correlations, $r^2 = .92$, between individual characteristics and visual appeal ratings. However, none of the partial regression coefficients in the multiple regression were statistically significant. This is not surprising, considering that the correlations among these five visual characteristics were extremely high, ranging from $r^2 = .85$ to $r^2 = .97$. The multiple regression analysis did not tell anything that was not already known from the individual correlations and will not be considered further.

Complexity and appeal. Berlyne (1971) suggested an inverted U-shaped curve as the relation between visual appeal and complexity. Data from Study 2 do not support this position. Figure 9 shows mean visual appeal ratings as a function of mean complexity rating. If Berlyne were right, the pattern of ratings would be low on the left with low complexity, increase up to some point with moderate complexity and then decrease as

³ A principal components analysis was performed to understand the inter-relations among the visual characteristics. Results of this analysis, which did not add to our understanding of the visual characteristics, are presented in Appendix G.

complexity gets even higher to the right. There is no discernible visual pattern, linear or curvilinear, in Figure 9.

Figure 10. The scatter plot of mean visual appeal vs. mean complexity is not an inverted U-shaped curve as Berlyne (1971) may have predicted.

Fitting a quadratic model when regressing mean visual appeal on mean complexity resulted in a squared correlation of only .01 ($p > .80$). Contrary to what Berlyne (1971) may have predicted, there is no apparent relation between ratings of visual appeal and complexity for web pages in Study 2.

Study 3

The purpose of Study 3 was to have eight graphic experts identify the graphical properties in the 50 web pages used in Study 2. Knowing the graphic properties for each web page, it should be possible to determine the relation between the visual appeal ratings made by the 31 participants in Study 2 and the properties as identified by the graphics experts. The objective is to understand what graphical properties underlie judgments of visual appeal.

Properties fell into 4 categories: color, layout, graphics, and font properties. Two experts rated all 50 web pages on one of the property categories. With 8 experts, all four sets of properties were evaluated twice. Thus, for each of the 50 web pages, there is a mean rating based on 31 participants from Study 2 and a mean ratings on a variety of graphic properties for that page each judged by two graphic experts from Study 3.

Inter-rater reliability. Before physical properties could be related to visual appeal ratings, the degree of inter-rater reliability for the two graphics experts for each property was determined. Graphic expert ratings of properties were made on either categorical or continuous scales. An example of a categorical scale would be *Is the background: a) a solid color, b) textured, or c) patterned*, and an example of a continuous scale would be *Rate the overall image quality* (from *high* to *low*). All graphic properties with continuous data were ratings by experts on 101-point scales⁴. Pearson product-moment correlations were used to determine the degree of inter-rater reliability between the 2 experts' ratings

⁴ These were technically 101-point interval scales. However, I decided to refer to them as continuous because these scales are interval representations of continuous scales.

for each continuous data property. There were significantly high degrees of inter-rater reliability for all 15 continuous scales, all $p < .05$, but only three correlations exceeded .70, which is generally accepted as the minimal acceptable correlation for reliability. Table C1 in Appendix C is a complete list of every property rated by design experts with statistical measures of inter-rater reliability and significance values.

For properties in categorical rather than continuous form, the Phi coefficient was used to determine the degree of inter-rater reliability between two experts. Some categorical data properties had either of 2 possible responses (e.g., *Are color gradients present in the background? Yes or No*), others had 3 (e.g., *Is the Background a Solid Color, is it Textured, or is it Patterned?*) and some had 4 (e.g., *What is the font type? Serif, Sans Serif, Script, or Custom.*). As shown in Table C1 in Appendix C, only 8 of the categorical variables had Phi coefficients of .70 or greater. Categorical font properties had conspicuously lower rates of inter-rater reliability than three other property categories. A possible reason for the low inter-rater reliability for font properties is the high number of properties that required subjective responses (e.g., *Is the font size too large, just right, or too small?*) rather than objective responses (e.g., *What is the font type? Serif, Sans Serif, Script, or Custom*) responses. Thirty-three of the 39 categorical font properties required subjective responses.

Simply having high inter-rater reliability was not enough to relate a categorical property to visual appeal. Categorical properties with insufficient variability were excluded from the analyses because variability in both predictor and outcome variables is necessary to establish a relation. For example, 46 of the 50 pages had some form of graphic image so the absence of variability makes presence or absence of an image

worthless for establishing a relation with visual appeal. Therefore, the *presence or absence of graphic images* was removed from the analysis. On the other hand, for example, the *presence or absence of photographs* had sufficient variability among the web pages to be useful in determining the relation between graphic properties and visual appeal. Based on this variability criterion, the single color property, 1 of the 2 graphics properties, and 1 of the 4 font properties that had adequate inter-rater reliability (Phi coefficients above .70) were removed.

Data for two sets of categorical properties were not collected in a manner that facilitated establishing their relation to visual appeal. These properties dealt with font types (e.g., *What is the font type? Serif, Sans Serif, Script, or Custom*) and background types (e.g., *Solid color, Textured, or Patterned*). A mean of ratings from two experts on these non-ordinal properties makes no sense. Therefore, I did not explore the relations between ratings on these properties and visual appeal.

Therefore, only 5 expert-rated properties were considered. The 2 categorical properties were *appropriateness of the number of images (too many, just the right number, too few)* and *screen dominance (more text than graphics, equal text and graphics, less text than graphics)*. The 3 continuous properties were the *rating of how carefully and discreetly display techniques are used together* (from *uncarefully/indiscreetly* to *carefully/discreetly*), the *rating of the use of fonts independently of other elements* (from *bad* to *good*), and the *rating of how well the fonts go with all the other elements* (from *bad* to *good*).

Relation between visual appeal ratings and individual visual properties. Means were calculated for the two experts for the 3 properties with continuous data for each of

the 50 web pages. Correlations between these means and the mean visual appeal scores for each web page calculated across test phases from Study 2 were then determined. All 3 continuous properties were positively correlated with rated visual appeal at statistically significant levels ($p < .005$) and are reported in Table 1. The 2 categorical properties were coded in such a way that means could be taken between the 2 expert raters. Judgments on the *appropriateness of the number of images property* were coded as *too many* = 0, *too few* = 0, and *just the right number* = 1. In this way, the presumably more positive response received the higher rating. The screen dominance property was coded as more text than graphics = 1, equal text and graphics = .5, and, less text than graphics = 0. Then, means were taken between expert ratings on these 2 categorical properties. Table 1 reports the squared correlations calculated between means on these properties and visual appeal. All were statistically significant at the .001 level.

Table 1. Squared Correlations between Properties and Visual Appeal.

Property	r^2
<i>rating of how carefully and discreetly display techniques are used together</i> (from <i>uncarefully/indiscreetly</i> to <i>carefully/discreetly</i>)	.64
<i>rating of the use of fonts independently of other elements</i> (from <i>bad</i> to <i>good</i>)	.59
<i>rating of well the fonts go with all the other elements</i> (from <i>bad</i> to <i>good</i>)	.68
Appropriateness of the number of images (too many, just the right number, too few) [categorical data]	.24
Screen dominance (more text than graphics, equal text and graphics, less text than graphics) [categorical data]	.23

Relation between visual appeal and combinations of visual properties. I wanted to determine whether visual appeal was best predicted by individual properties or by the combination of the properties and explore the combinations that gave the best predictions. A multiple regression analysis was computed by regressing visual appeal on the five properties. In contrast to correlations between individual ratings on properties with visual appeal, the linear combination of these properties was highly correlated with rated visual appeal ($R^2 = .78$, $p < .001$). This squared correlation is larger than any squared correlation calculated between single predictors and visual appeal indicating that the combination of properties is a better predictor than individual properties.

An examination of the standardized partial regression coefficients in Table 2 reveals that the *screen dominance* and *ratings of how carefully/discreetly display techniques are used together* properties, the only properties with a statistically significant standardized partial regression coefficient, β , contribute the largest unique proportions of variance in the regression equation.

Table 2. Standardized Partial Regression Coefficients, ts, and Significance Values for Property Variables Predicting Visual Appeal.

Properties	β	t	p
Rating of how carefully and discreetly display techniques are used together (from uncarefully/indiscreetly to carefully/discreetly)	.29	2.04	< .05
Rating of the use of fonts independently of each other (from bad to good)	.07	.26	> .05
Rating of how well the fonts go with all the other elements (from bad to good)	.47	1.50	> .05
Appropriateness of the number of images (too many, just the right number, too few) [categorical data]	.20	1.87	> .05
Screen dominance (more text than graphics, equal text and graphics, less text than graphics) [categorical data]	.17	2.15	< .05

Understanding visual properties. A Principal Components Analysis was conducted to determine which of the five properties were rated similarly enough to be grouped to represent single concepts. The analysis was of little use because it identified only one component and will not be mentioned further.

Discussion

Previous research

The results of these studies largely agree with results from previous research. In accordance with findings by Schenkman and Jonsson (2000), web page complexity was unrelated to appeal. As expected, the following web page characteristics were all highly related to visual appeal: degree of web page imaginativeness, how interesting a web page seemed, quality of web page design, and quality of web page color and layout usage. Degree of web page clarity was less highly related to visual appeal.

Some of the results from Study 3 are in agreement with previous research while others are not. Based on the results of Study 3, a preference for web pages with more illustrations than text was confirmed (Schenkman & Jonsson, 2000). An economic layout (careful/discreet rather than uncaredful/indiscreet use of display techniques) was also confirmed to be related to visual appeal (Ngo, Samsudin, & Abdullah, 2000).

Unfortunately, the relation of other factors to visual appeal could not be determined for various reasons. This study could not confirm the presence of images as important to high visual appeal (Haig & Whitfield, 2000; Knutson, 1998) because the majority of web pages included images so there was insufficient variability to establish a relation. The relation between the presence of photographs (rather than the presence of iconic, hand drawn, 3D computer illustrated, or decorative treatment images) and visual appeal could not be determined because of low inter-rater reliability. The same was true of the importance of symmetrical layout of materials on a page vs. asymmetrical layout of materials to visual appeal (Ngo, Samsudin, & Abdullah, 2000; Schenkman & Jonsson, 2000). For example, when the two experts failed to agree on whether a layout was symmetrical or not, it made little sense to determine correlations between the mean expert rating on symmetrical layout and the mean visual appeal score for each web page from Study 2. I could not determine the relation of various other properties to visual appeal including grid layout patterns (Borchers, Deussen, & Knorzer, 1995), or the presence or absence of a 3D look was not related to visual appeal due to low inter-rater reliability.

Expert agreement on visual properties in web sites. There were often properties that had levels of inter-rater reliability which were not statistically significant for the pairs of experts in Study 3. This was more true of the pair of experts rating font

properties than the other pairs of experts who rated color, layout, and graphic properties. The most likely reason for the lack of agreement among font raters was that items in the font category included a greater proportion of subjective responses (e.g., *Is the font size too large, just right, or too small?*) rather than objective responses (e.g., *What is the font type? Serif, Sans Serif, Script, or Custom.*) as compared with any other property category. Objective items were straightforward, mostly inventory-type questions that asked the experts to identify the presence or absence of properties. Subjective items involved experts making very subjective judgments based on their previous experience and interpretation of the question. Future studies attempting to relate web page properties to visual appeal could reduce subjective items in favor of objective items. In this case, the use of experts may be superfluous. By providing examples of different design properties usage (e.g., showing screen shots of different background types or font types), non-experts may be able to complete a web page visual property inventory similar to the one used in Study 3. Examples of design property usage were provided for experts in the current study. Certainly these examples contributed to the many instances of properties for which there were statistically significant inter-rater reliabilities. However, these examples were obviously not good enough to provide phi coefficients over .70.

In addition to use of objective criteria and concomitant examples, evaluations of visual characteristics of web pages could benefit from better categorization of response choices than I used. Quantitative rather than qualitative scales would make it easier to do analyses and increase the chances that adequate variability would be available to do meaningful analyses.

Visual appeal assessment technique

Reliability of visual appeal judgments. There was high reliability between participants' visual appeal ratings of web pages between the two test phases in Studies 1 and 2 using the visual appeal assessment technique. This reliability was impressive for two reasons. It was feared that the brief stimulus display duration would not allow participants enough time to judge visual appeal. It was also feared that an aesthetic opinion would be so capricious that it could change across test phases. Although most participants showed strong agreement between judgments in the two phases, a few participants had lower, though still statistically significant, squared correlations of responses between test phases.

It is possible that the high reliability between test phases is a result of a procedural artifact such as a demand characteristic. The procedure had participants rate the visual appeal of numerous web pages. After a brief rest, participants again rated the appeal of the same pages. It is possible that during the second test phase, participants tried to recall their responses to the same web pages they had seen previously in the first test phase and consciously provide those same responses. The interval between rating the same web page across test phases was approximately 15 minutes and reliability is thought to be inversely related to elapsed time between test sessions (Trochim, 2002). A future study could extend the time between test phases, perhaps for as long as weeks. This extended interval between test phases may reduce the demand characteristic of participants actively trying to recall their responses to web pages in the first test phase when they make judgments during the second test phase for the sake of consistency. After a long time I would expect some memory of the sites to still be possible, but I would expect memory

of the appeal judgments used for individual pages to diminish. Under those circumstances, similar responses across test phases would be based on independent judgments.

Overall means as a measure of visual appeal. I wanted to be sure that overall visual appeal means (calculated across responses in Phases 1 and 2 for all participants of Study 2) were a valid measure of visual appeal. Due to the possibility of extreme individual differences which may be expected in subjective ratings, I feared that the overall visual appeal means may not have accurately reflected some participants' ratings. Therefore, I sought to determine the relation between individuals' appeal ratings and ratings based on the overall visual appeal means calculated across all participants. I calculated the correlations between visual appeal means averaged across both test phases for all 31 participants and the visual appeal means for each of the 31 participant averaged across both phases. All squared correlations were statistically significant at $\alpha = .01$ level or much better with about half the squared correlations being above .70. Given these results, I concluded that the overall appeal means were a valid measure of visual appeal because they were representative of individual participants' appeal ratings.

Eye-tracking. What exactly do people look at in 500 ms that influences their visual appeal judgment of a web page? Data from eye-tracking systems can answer this question by revealing what a participant views, when they view it, and the duration of their gazes. In the future, Study 1 could be rerun while measuring eye movements. Researchers could examine the relation between appeal scores and web page design components that receive the most visual attention as measured by gaze duration and the sequence that they look at items on the page. Eye-tracking data could also reveal whether

participants are looking at images or text or even whether they could possibly be reading text and making judgments based on content rather than aesthetics. Eye-tracking data could also determine where participants look first on the most appealing pages.

Stimulus duration. Stimuli were displayed for 500 ms to minimize the chances that participants would read web page text and maximize the probability of participants responding to web page aesthetics. However, it could be argued that, at 500 ms, participants were viewing and processing much more than simply the aesthetic attributes of the web pages. A future study should examine how short the stimulus duration could be while still maintaining high reliability across test phases. In such a study, the time for display of web images would be varied from the current 500 ms to extremely short. This study would inform our understanding of the minimum viewing time necessary to make judgments on visual aesthetics, which is important to studies of first impressions. Extremely short stimulus durations that maintain high test phase reliability would ensure that appeal judgments are responses to visual appeal and not other characteristics such as textual content that cannot be read when stimuli are displayed for such short durations.

Veryzer (1999) suggests that affect is a primary characteristic of an aesthetic response and Karvonen suggests that people make aesthetic judgments immediately. Another modification to a study with reduced display duration would be to have participants indicate how they *feel* when viewing web pages and responding as quickly as possible to ensure a first impression response (Lindgaard & Whitfield, in press). The *Very Appealing – Very Unappealing* rating scale could be changed to have the anchors *positive* and *negative feeling* (Lindgaard & Whitfield, in press). Responses to this affect scale could be related to visual appeal data as collected in Study 2. The proposed study

could help determine how well emotional responses relate to aesthetic responses. A high correlation between responses to the affect scale and visual appeal scale might indicate that the visual appeal scale used in Study 2 was also measuring affect. Such an outcome might have important implications for first impressions aesthetic research that could further validate the methodology used in this thesis.

Timing accuracy. There can be timing inaccuracies when using software to control stimulus display duration that may affect a study's results. Display time accuracy can be problematic on computers running multi-tasking operating systems such as the one used in this research (MacInnes & Taylor, 2001). For example, the standard Windows timer is accurate to "multiples of approximately 55 ms" (Riemersma, 1999). Timing in this study, the 500 ms presentation of web page images, was controlled by the standard system clock software. An obvious limitation of an inaccurate software timer is the possible deception of thinking that a stimulus has been displayed for an exact amount of time (e.g., 500 ms) when it has actually been displayed for a greater or lesser time. Shareware timer controls such as OpenRT substitutes Windows code problematic for timing with a superior version (MacInnes & Taylor, 2001). Other issues include synchronizing stimulus presentation with the refresh rate of CRT monitors and preloading visual stimuli, all of which are addressed by PCTSCOPE (Tsai, 2001). Though these problems are probably not issues for the current study which displayed stimuli for a relatively long duration (500 ms), software for future studies which display stimuli for duration as short as 100 ms, 50 ms, or lower, should take advantage of shareware modules that correct Microsoft Windows system timing inaccuracies.

Image loading time from the hard drive into memory could also affect a study's results. For example, larger files take longer to load than smaller files. Graphic files (e.g., jpegs) with more content could conceivably take longer to load than files with less content. Therefore, loading time issues could cause systematic timing inaccuracies that allowed certain web pages to be displayed for longer durations than others could have implications for the present studies. It is possible that web pages displayed for longer times could be systematically rated as having less visual appeal. The solution is to preload the web pages from the hard drive into memory before starting the timer that controls duration of display.

Scale anchors. It is important to consider anchor placement when constructing rating scales. The rating scales used in the third test phase of Study 2 had positive anchors on the left and negative anchors on the right (e.g., *clear-confusing*, *interesting-boring*). Devasagayam (1999) conducted a study to explore whether the position of positive and negative rating scale anchors biased responses in a particular direction. Each participant in his study responded to 3 questionnaires. One questionnaire had all positive anchors on the left, the second had all positive anchors on the right, while the final one had randomized position of positive and negative anchors. Devasagayam found that questionnaires for which all positive anchors were on the left resulted in a positive bias while questionnaires that had all negative anchors on the left resulted in a negative bias. Randomized positive and negative anchors eliminated the biases. The positive bias was 7% above the mean unbiased rating while the negative bias was 8% below the mean unbiased rating. If his findings are accurate, ratings from Study 2 may be positively biased, that is, ratings of visual characteristics may be too high. Even if there were a bias,

there was high reliability and strong relations between the visual characteristics and visual appeal ratings. Additionally, constantly changing anchor direction could distract and confuse participants and possibly increase error if respondents don't carefully read the anchors for each judgment. I feel that a systematic bias is preferable to the possibility of inducing participant error by constantly changing anchor direction. Furthermore, there would be no easy way to know if an error were made because of the constantly changing direction. Therefore, I feel that future studies should maintain the anchor placement used in the present studies.

Visual appeal and complexity. The relation between simplicity or complexity and visual appeal is unclear. Study 2, consistent with Schenkman & Jonsson (2000), did not find a relation between rated web page complexity and aesthetic appeal. However, based on Karvonen's (2000) thoughts regarding the relation between beauty and simplicity, it might be expected that simple web pages would have been rated as appealing while complex pages would have been rated as unappealing. On the other hand, Berlyne (1971) suggested that the relation between appeal and complexity of a visual stimulus can be described as an inverted U curve with lowest appeal for high and low complexity and highest appeal for moderate complexity. Study 2 demonstrated no relation between visual appeal and complexity. Figure 9 shows that there was no relation, linear or curvilinear, between visual appeal and complexity. The difference between these findings and those of Berlyne (1971) could be attributed to numerous factors. One explanation for the discrepant findings could be differences in methodology. Berlyne's (1971) visual stimuli were line drawings of preexisting paintings while those in Study 2 were web pages. Additionally, visual appeal ratings in Study 2 were gathered after the stimuli were

displayed for only 500 ms. Berlyne did not impose such brief display durations. With such strong relations between other visual characteristics and visual appeal, it looks like complexity of web pages is relatively unimportant in visual appeal. Based on the findings of the present study and previous research (Schenkman & Jonsson, 2000), it may be that the relation between complexity and appeal in web pages is different than the relation between complexity and appeal in traditional art work.

A tool for assessing visual appeal

Results from this thesis suggest the possibility that the visual appeal judgment technique could be adapted to provide a practical tool to assess the visual appeal of user interfaces. Karat (2003) suggests there is a bias among usability specialists to measure objective factors such as task efficiency, task completion, and error rates rather than more subjective user interface qualities such as aesthetics. One reason for the lack of concern for aesthetic issues is the absence of an objective way to measure aesthetics. The technique used in Study 1 and 2 can be applied to measure visual appeal of web pages with objectivity.

An additional benefit of the technique is that only a small number of participants, possibly as low as 5, possibly running through a single test phase would be required. As already reported, the mean of the correlations between web page rankings based on 5 subsamples consisting of 5 participants each and the total sample of 31 participants was very high ($r^2 = .91$). These findings indicate that sample sizes as small as 5 participants may be sufficient to judge the visual appeal of web pages.

The method used for stimulus presentation and data collection in Studies 1 and 2 could be used in an industrial setting to assess the visual appeal of web pages. A client

needing to assess the visual appeal of their web page could compare it against the web pages and data collected from Study 2 as a normative dataset. The client's page could be placed in the study with the current web pages then appeal responses to it could be compared with web pages from Study 2. Results could be in the form of a point representing the visual appeal of their web site on the regression line. The point would not be interpreted in a precise fashion, but could give an estimate of approximately where on the visually appealing web site scale the client web site falls.

A software package consisting of the normative dataset, instructions on how to add a web page to the set, the program to run the visual appeal study, and a data analysis module that presents the predicted appeal score of the target web page in relation to those of the normative dataset could be developed and provided for free or for a fee.

Alternatively, a company could offer the service for a fee.

Web pages in the normative dataset might require some modification to account for different web page purposes. The stimuli from Study 2 are web pages from a variety of contexts including homepages for educational institutions, government agencies, online and offline vendors and service providers, and individuals. The technique could be refined by selecting a normative dataset of web pages of the same usage context as the client's and collecting visual appeal preference data. For example, if the client were an online vendor, it would be most appropriate to compare the visual appeal of the home page against other online vendors rather than those of government agencies or individuals.

The importance of comparing web pages from the same usage context when assessing web page visual appeal should not be ignored. Several researchers have

suggested that making aesthetic judgments about an object cannot be divorced from the object's aesthetic category or context (e.g., Karvonen, 2000; Whitfield, 2000). For example, Whitfield (2000) postulates that aesthetic judgments of an object are based on the conceptual category to which that object belongs, particularly if the category is well-defined. An example of a well-defined category would be gothic cathedrals, of which Notre Dame in Paris is an exemplar. Judgments of an object's appeal are not made in isolation but are best made through comparisons with other objects of the same well-defined category. Whitfield suggests that the *web page* category is not yet completely defined. However, it may be that subcategories within the web page category, such as e-commerce or news sites, are fairly well articulated already and people have general expectations of what they should look like. For example, Bernard (2002) found that people have expectations regarding the locations of objects on e-commerce web pages. These expectations could affect judgments of visual appeal. People may also have expectations about how web pages from other categories such as government agencies or universities should look, so separate normative data would be collected for such web sites.

An assessment tool created for industrial use need not be limited to assessing the visual appeal of web pages. The tool could also be used to assess the visual appeal of GUIs for software applications and even operating system user interfaces. The tool may be of particular relevance to user interface (UI) graphic designers who are designing a new *look and feel* for a forthcoming release of their software. Often, designers come up with competing design concepts and may have difficulty deciding which to implement. In keeping with the user-centered design approach, the tool would allow designers to let

users decide. It may also be used to compare the visual appeal of a product against those of competitors. Another benefit of this tool is that it can be used with ease. It can be easily included as part of a usability test, thus reducing the need for additional participant recruitment. S. Marsh, (personal communication, October 30, 2002) suggests that the visual appeal tool need not be limited to assessing GUI visual appeal but may also be used for physical objects such as computer-related hardware, or even images of objects such as cars, toasters, or other product conceptualizations. For example, designers often produce hand-drawn or computer-rendered design prototypes of software products. These images could easily be used with the tool to determine visual appeal preferences. Thus, the technique has the potential to become a useful addition to the usability specialist's user-centered design (UCD) toolkit that could possibly be used, for example, to conduct competitive benchmarks of product visual appeal.

A tool for assessing first impressions of trust and perceived usability

The tool presented here has the potential to become a quick and reliable instrument for assessing perceived trust or perceived usability of an interface. According to Tractinsky (1997), there are links between perceived usability and long-term system acceptance. Kim and Moon (2000) suggest that trust of e-commerce pages is thought to affect customer loyalty. Karvonen (2000, p. 87) suggests impressions of usability and trust are made "on-the-spot." Future studies could examine whether the method employed in Study 1 could be applied to studying perceived trust by substituting *very trustworthy* – *very untrustworthy* as anchors instead of visual appeal anchors. Following Phase 2, semantic differential scales could also be included: *looks cheap* – *doesn't look cheap*, *looks like they spent a lot of time designing this* – *doesn't look like they spent a lot*

of time designing this, looks well-planned – doesn't look well-planned, and has a quality look – doesn't have a quality look. Using a method similar to the third test phase of Study 2, a Principle Components Analysis and a multiple regression analysis could be employed to determine which concepts (e.g., cheap look, well-planned look, quality look) are related to trustworthiness.

The same approach could be used to get direct ratings of perceived usability by using anchors of *looks very easy to use* and *looks very difficult to use*. Characteristics that might underlie perceived usability could be included as in Phase 3. Indeed, the same sites could be submitted to different users making ratings of visual appeal, trust and perceived usability. Using the results of these studies, it would be possible to determine the relations between judgments of visual appeal, perceived trust and perceived usability by regressing mean ratings for one on the others.

Aesthetics and usability research

The visual appeal assessment technique could be used to help researchers identify levels of visual appeal for use in experiments that investigate relations between aesthetics and perceived usability. A major difficulty in experimentally investigating the relation between visual appeal and usability is producing an appeal manipulation that is not confounded with usability. Others have sought to explore the relation between aesthetics and usability (Kurosu & Kashimura, 1995; Tractinsky, 1997; Tractinsky et al. 2000) while using differences in layout as the major visual appeal manipulation. However, layout is an aesthetic factor closely linked to usability. Manipulating non-layout aesthetics factors may prove more fruitful when examining the link between aesthetics

and usability. The visual appeal technique presented in this thesis could be used to identify levels of visual appeal that do not have layout differences.

The technique from Study 1 could be used to select the stimuli for an experiment examining the relation between visual aesthetics and user satisfaction. For example, the experiment could have a 3 x 2 (3 levels of visual appeal, 2 level of usability) factorial design. Two or more sites varying in usability could be designed. Web page looks or *skins* could be designed. It would be possible to control for the potential layout usability-aesthetics confound by maintaining consistent layout while varying other characteristics such as use of images and decorative treatments. Then the variations would be evaluated using the visual appeal assessment technique. The results could be used to select pages of high, medium, and low visual appeal. Participants would complete the same usability tasks using different combinations of aesthetic appeal and usability and be asked standard user satisfaction and usability questions. Task performance would provide a measure of actual usability, and differences in responses to these questions would indicate how the different conditions affected overall satisfaction and perceived usability. Thus, the assessment technique could be used to select the stimuli for an experiment investigating the relation between visual appeal and usability while excluding the possible layout confound.

Usage log analysis

The technique could be used to determine whether the time people spend on a web page is related to the page's visual appeal. Anecdotal evidence (A. Patrick personal communication, October 30, 2002) from examinations of logs of the pages that people view suggests that they spend less time on visually unappealing web pages than

appealing pages. This anecdotal evidence also suggests that people may leave sites they find visually unappealing while lingering longer on sites they find appealing. A simple study could examine the relation between appeal and time spent on pages. For example, web site usage logs can be analyzed to identify the home pages on which people spend the most and least time. The visual appeal of these pages can then be assessed using the technique developed in this thesis. Usage logs can also be used to identify home pages with high click-through rates, where people proceeded to other pages of the site after viewing the home page. These home pages can then be assessed for visual appeal to determine the relation between visual appeal and click-through on web page links.

Cross-cultural studies.

Another application of the visual appeal technique could be to assess how users from other cultures would respond to the aesthetics of a product's user interface. The studies in this thesis included North American participants who had English as their first language. It would be surprising if people from other cultural backgrounds did not respond differently to visual appeal. This issue is faced by software companies who try to target products developed within a particular cultural context for people from other cultures. Visual appeal data about a particular product's UI gathered with North American participants could be compared to data from participants of other cultures. The tool could be used to determine the need for aesthetic modifications to product UIs ported to different cultural markets.

References

- Bamossy, G., Scammon, D. L., & Johnston, M. (1983). A Preliminary Investigation of the Reliability and Validity of an Aesthetic Judgement Test. In R. Bagozzi & A. Tybout (Eds.). *Advances in Consumer Research: Vol. 10*. (pp. 685-690). Ann Arbor, MI : Association of Consumer Research.
- Basso, A., Goldberg, D., Greenspan, S., & Weimer, D. (2001). First impressions: Emotional and cognitive factors underlying judgments of trust in e-commerce. *Proceedings of the 3rd ACM conference on Electronic Commerce* (pp. 147-143). New York: ACM Press.
- Berlyne, D. E. (1971). *Aesthetics and Psychobiology*. New York: Appleton-Century-Crofts.
- Bernard, M., Mills, M., Peterson, M., & Storrer, K. (2001). A comparison of popular online fonts: Which is best and when? *Usability News*, 3.2. Retrieved from <http://psychology.wichita.edu/surl/usabilitynews/3S/font.htm>
- Bernard, M., Lida, B., Riley, S., Hackler, T., & Janzen, K. (2002). A comparison of popular online fonts: Which size and type is best? *Usability News*, 4.1. Retrieved from <http://psychology.wichita.edu/surl/usabilitynews/41/onlinetext.htm>
- Bernard, M. (2002, July 23). Examining user expectations for the location of common e-commerce web objects. *Usability News*, 4.1. Retrieved September 1, 2002, from http://psychology.wichita.edu/surl/usabilitynews/41/web_object-ecom.htm
- Borchers, J., Deussen, O, & Knorzer, C. (1995, October). Getting it across: Layout issues for kiosk systems. *ACM SIGCHI Bulletin*, 27, 68-74.

- Brunel, F. F. (1998). The psychology of product aesthetics: Antecedents and individual differences in product evaluations. *Dissertation Abstracts International*, 59 (03), 890A. (UMI Number: 9826334)
- Cupchik, G. C., & Gebotys, R. J. (1990). Interest and pleasure as dimensions of aesthetic response. *Empirical Studies of the Arts*, 8, 1-14
- Devasagayam, P. R. (1999). The effects of randomized scales on scale-checking styles and reaction time. *Marketing Management Association*. Retrieved on September 5, 2002, from <http://www.sbaer.uca.edu/Research/1999/MMA/99mma155.htm>
- Haig, A. & Whitfield, T. W. A. (2000). Predicting the aesthetic performance of web sites: What attracts people? *Smart Graphics AAAI 2000, Spring Symposium*.
- Hassenzahl, M., A. Beu, & Burmester, M. (2001). Engineering joy. *IEEE Software*, 18, 70-76.
- Hofmeester, G. H., Kemp, J. A. M., & Blankendaal, A. C. M. (1996). Sensuality in Product Design: a Structured Approach. *CHI 1996 Electronic Proceedings*.
- Jennings, M. (2000). Theory and models for creating engaging and immersive ecommerce Websites. *Proceedings of the 2000 ACM SIGCPR Conference on Computer Personnel Research* (pp. 77-85). New York: ACM Press.
- Jordan, P. W. (1998). Human factors for pleasure in product use. *Applied Ergonomics* 29, 25-33.
- Karat J. (2003). Beyond task completion: Evaluation of affective components of use. In J. A. Jacko and A. Sears (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (pp. 1152-1164). Hillsdale, NJ: Erlbaum

- Karvonen, K. (2000). The beauty of simplicity. *Proceedings of the Conference on Universal Usability* (pp. 85-90). New York: ACM Press.
- Kim, K. and J. J. Moon (1998). Designing towards emotional usability in customer interfaces: Trustworthiness of cyber-banking system interfaces. *Interacting with Computers*, 10, 1-29.
- Knutson, J. F. (1998). The effect of the user interface design on adoption of new technology. *Dissertation Abstracts International: Section B: The Sciences and Engineering* 59(3), 1399B. (UMI Number: 9827294)
- Kubaitis, E. (2002, May). *EWS Browser Statistics for July 2002*. Retrieved June 20, 2002, from <http://www.cen.uiuc.edu/bstats/latest-month.html>
- Kurosu, M. & K. Kashimura (1995). Apparent Usability vs. Inherent Usability: Experimental analysis on the determinants of the apparent usability. In I. R. Katz, R. Mack, L. Marks, M. B. Rosson, & J. Nielsen (Eds.), *Conference Proceedings on Human Factors in Computing Systems* (pp. 292-293). New York: ACM Press.
- Lindgaard, G. (1999). Does emotional appeal determine the usability of web sites, *Proceedings CYBERG '99*, Western Australia.
- Lindgaard G., & Dudek, C. (in press). What is this evasive beast we call user satisfaction? *Interacting with Computers*.
- Lindgaard, G., & Whitfield, T. W. A. (in press). Integrating aesthetics within an evolutionary and psychological framework. *Theoretical Issues in Ergonomic Science*.
- MacInnes, J., & Taylor, T. L. (2001). Millisecond timing on PCs and Macs. *Behavior Research Methods, Instruments, & Computers*, 33, 174-178.

- Neufeldt, V., & Guralnik, D. B. (Eds.). (1997). *Webster's New World College Dictionary* (3rd ed.). New York: Macmillan.
- Ngo, D. C. L., Samsudin, A., & Abdullah, R. (2000). Aesthetic measures for assessing graphic screens. *Journal of Information Science and Engineering*, 16, 97-116.
- Norman, D. (2002, July). Emotion & Design: Attractive things work better. *Interactions*, 9, 36-42.
- Riemersma, T. (1999, July 17). *Accurate timing in Visual Basic*. Retrieved September 12, 2002, from <http://www.compuphase.com/vbtiming.htm>
- Ryle Design. (2000, July 18). *ExacTicks 1.1 Manual*. Retrieved September 12, 2002, from <http://www.ryledesign.com/download/winhrf.pdf>
- Schenkman, B. N., & Jonsson, F. U. (2000). Aesthetics and preferences of web pages. *Behaviour & Information Technology*, 19, 367-377.
- Simeon, R. (2001). Evaluating the branding potential of Web sites across borders. *Marketing Intelligence and Planning*, 19, 418-424.
- Simpson, J. A., & Weiner, E.S.C. (Eds.). (1989). *Oxford English Dictionary* (2nd ed.). New York: Oxford University Press.
- Tombaugh, J. W., R. F. Dillon, & Carboni, N. (1982). Evaluation of graphics on videotex by inexperienced users. *Proceedings for the Graphics Interface Conference*, Toronto, Ontario, Canada, pp. 339-344.
- Tractinsky, N. (1997). Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In S. Pemberton (Ed.), *Conference proceedings on Human Factors in Computing Systems* (pp. 115-122). New York: ACM Press.

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

Trochim, W. K. (2002). *Types of reliability*. Retrieved August 12, 2002, from <http://trochim.human.cornell.edu/kb/reotypes.htm>

Tsai, J.-L. (2001). A multichannel PC tachistoscope with high resolution and fast display change capability. *Behavior Research Methods, Instruments, & Computers* 2001, 33, 524-531.

Veryzer, R. W., Jr. (1993). Aesthetic response and the influence of design principles on product preferences. In L. McAlister & M. L. Rothschild (Eds.), *Advances in consumer research*, Vol. 20, pp. 224–228. Provo, UT: Association for Consumer Research.

Veryzer, R. W. (1999). A Nonconscious Processing Explanation of Consumer Response to Product Design. *Psychology & Marketing*, 16, 497-522.

Weeks, C. E. (1997). How to design an aesthetically-pleasing, ADA-compliant website. *Proceedings of the 25th SIGUCCS conference on User services: are you ready?*, 319 -326.

Weimer, J. (1995). *Research techniques in human engineering*. Englewood Cliffs, NJ. Prentice Hall.

Whitfield, T.W.A. Beyond prototypicality: towards a Categorical-Motivation model of aesthetics. *Empirical Studies of the Arts*, 18, 1-11.

Appendix A

Task flow for Study 1

Appendix B

Task flow for Study 2

Appendix C

Properties rated by graphic design experts in Study 3

Table C1 provides a list of all expert-rated web page properties with measures of inter-rater reliability between the 2 experts that rated each property. The Phi coefficient was used to determine inter-rater reliability on categorical properties while the Pearson product-moment correlation (r) was used for properties rated on continuous scales. Significance values are also reported. Figures C1 and C2 are screenshot examples as seen by the experts of a continuous and a categorical properties, respectively.

Table C1. Properties rated by graphic design experts in Study 3.

	Phi	r	p
Color Properties			
Background type: Solid, Textured, Patterned	.87		< .05
Presence or absence of color gradients	.65		< .001
Color use: Harmonious, Unharmonious	.56		< .001
Appropriateness of number of colors used: too few, just the right number, too many	.59		< .05
Color types			
Presence of soft/pastel colors	.30		< .05
Presence of hard/neon colors	.50		< .001
Presence of darker hues	.49		< .05
Presence of warm colors	.37		< .01
Presence of cold colors	.06		not sig
Ratings of the use of color independently of other elements, from <i>bad</i> to <i>good</i> [continuous data]		.68	< .001
Ratings of how well the color usage goes with all the other elements, from <i>bad</i> to <i>good</i> [continuous data]		.63	< .001
Layout Properties			
Justification: Right, center, left	.32		not sig
Layout: Presence or absence of a clearly defined grid	.42		< .01

	Phi	r	p
Screen dominance: More text than graphics, equal text and graphics, less text than graphics	.78		< .001
Symmetry: balanced or unbalanced distribution of equivalent elements about a common line	.28		not sig
Horizontal balance: maintains or does not maintain equal weighting of components on each side of the horizontal axis	.28		not sig
Vertical balance: maintains or does not maintain equal weighting of components on each side of the vertical axis	.41		< .01
Economic or uneconomic: uses few or many styles and display techniques	.49		< .01
Rating of how carefully and discreetly display techniques are used together, from <i>uncarefully/indiscreetly</i> to <i>carefully/discreetly</i> [continuous data]		.70	< .001
Rating of the layout independently of other elements, from <i>bad</i> to <i>good</i> [continuous data]		.50	< .001
Rating of how well the layout goes with all the other elements, from <i>bad</i> to <i>good</i> [continuous data]		.41	< .005
Graphic properties			
Presence or absence of 3D elements	.67		< .001
Presence or absence of overlapping elements	.27		not sig
Presence or absence of text shadows	.65		< .001
Presence or absence of image drop shadows	.40		< .01
Presence or absence of 3D images (not including photographs)	.47		< .01
Presence or absence of transparent objects	.56		< .001
Presence or absence of 3D interface elements	.58		< .001
Rating of the use of 3D elements independently of other elements, from <i>bad</i> to <i>good</i> [continuous data]		.47	< .005
Rating of how well the 3D elements go with all the other elements, from <i>bad</i> to <i>good</i> [continuous data]		.50	< .001
Presence or absence of images	.56		< .001
Presence or absence of photographs	.83		< .001
Presence or absence of clip art	.52		< .001
Presence or absence of iconic graphics	.43		< .01
Presence or absence of hand drawn images	.39		< .01
Appropriateness of number of images: Too few, just right, too many	.70		< .001
Integration: Images are integrated well with pages or images seem just slapped on to page	.48		< .01

	Phi	r	p
Rating of the overall image quality, from <i>low</i> to <i>high</i> [continuous data]		.39	< .05
Rating of the use of images independently of other elements, from <i>bad</i> to <i>good</i> [continuous data]		.48	< .001
Rating of how well the images go with the other elements, from <i>bad</i> to <i>good</i> [continuous data]		.35	< .05
Presence or absence of decorative treatments	.14		not sig
Presence or absence of lines	.45		< .01
Presence or absence of curves	1.00		< .001
Presence or absence of circles	.39		< .01
Presence or absence of rectangles	.39		< .01
Presence or absence of dots	-- ⁵		--
Presence or absence of fading	.39		< .01
Purposeful application: treatments are applied purposefully or haphazardly	.41		not sig
Rating of the use of decorative treatments independently of other elements, from <i>bad</i> to <i>good</i> [continuous data]		.38	< .005
Rating of how well the decorative treatments go with each other, from <i>not well</i> to <i>well</i> [continuous data]		.46	< .001
Font Properties			
Presence or absence of title font	-- ⁶		--
Title font type: Serif, sans serif, stylized/custom, script/cursive	1.09		< .001
Title font size: large, medium, small	.68		< .001
too large, just right, too small			< .05
relative to:			
body font size	-.09		not sig
legibility	.08		not sig
other sites	.04		not sig
High or low readability	.39		< .01
Due to:			
color	.40		< .005
font type	.31		< .05
font size	.41		< .005
font treatments (bold, italics, etc.)	.32		< .01
saturation/pixelation	.06		not sig
Presence or absence of body font	.12		not sig
Body font type: Serif, sans serif, stylized/custom,	1.40		< .001

⁵ This value could not be computed because both experts' ratings were considered constant. The experts rated only 1 web page each as having curves.

⁶ This value could not be computed because 1 expert's ratings were constant, that is, he or she rated every web page as having a title font.

script/cursive		
Body font size: large, medium, small	.37	not sig
too large, just right, too small	.73	< .005
relative to		
title font size	.25	not sig
legibility	.20	not sig
other sites	.37	< .01
High or low readability	.41	< .05
Due to:		
color	.44	< .005
font type	.20	not sig
font size	.22	not sig
font treatments (bold, italics, etc.)	.05	not sig
saturation/pixelation	-.06	not sig
Presence or absence of other font	.30	< .05
Other font type: Serif, sans serif, stylized/custom, script/cursive	.70	< .05
Other font size: large, medium, small	.54	not sig
too large, just right, too small	.41	not sig
relative to:		
the other font sizes	.16	not sig
legibility	.11	not sig
other sites	-.19	not sig
High or low readability	.07	not sig
due to		
color	.31	< .05
font type	.43	< .005
font size	.15	not sig
font treatments (bold, italics, etc.)	-.05	not sig
saturation/pixelation	-.04	not sig
Rating of how well all the fonts go together, from <i>not well</i> to <i>well</i> [continuous data]	.45	< .005
Rating of the use of fonts independently of other elements, from <i>bad</i> to <i>good</i> [continuous data]	.70	< .001
Rating of how well the fonts go with all the other elements, from <i>bad</i> to <i>good</i> [continuous data]	.72	< .001

Please rate the image quality:

Low

High

Figure C1. An example of a property coded as continuous data.

Screen Dominance

- ☐ More text than graphics
- ☒ Equal text and graphics
- ☐ Less text than graphics

FigureC2. An example of a property coded as categorical data.

Appendix D

Squared correlations between visual appeal ratings in two test phases for participants in
Study 1.

This table presents the squared correlations calculated between each participant's scores in Test Phase 1 and Test Phase 2 for Study 1. All squared correlations were highly significant ($p < .001$).

Participant	r^2
1	.77
2	.73
3	.41
4	.71
5	.57
6	.47
7	.63
8	.33
9	.40
10	.71
11	.72
12	.73
13	.63
14	.70
15	.56
16	.80
17	.73
18	.81
19	.80
20	.55

Participant	r^2
21	.53
22	.69

Appendix E

Squared correlations between visual appeal ratings across test phases for participants in Study 2.

This table presents the squared correlations calculated between each participant's scores in Test Phase 1 and Test Phase 2 for Study 2. All squared correlations were highly significant ($p < .001$).

Participant	r^2
1	.84
2	.60
3	.71
4	.31
5	.41
6	.85
7	.78
8	.78
9	.95
10	.76
11	.79
12	.81
13	.82
14	.62
15	.92
16	.64
17	.26
18	.79
19	.86
20	.83

Participant	r^2
21	.78
22	.60
23	.70
24	.70
25	.73
26	.69
27	.55
28	.70
29	.73
30	.69
31	.88

Appendix F

Subsample Validation

Web page rank comparisons were performed between the total sample and subsamples of smaller sizes ($N = 2$ to $N = 10$) randomly selected from the total sample to determine the sample size where the data for subsamples starts to differ from the data for the total sample of 31. Two subsamples were taken for the subsample sizes that were fairly stable ($N = 7$ to $N = 10$), while five subsamples were taken for less stable subsample results ($N = 2$ to $N = 6$). The means for the 50 web pages in the total sample and subsamples were calculated and ranked in terms of mean visual appeal rating. The subsample ranks were then paired with the corresponding total ranks for each page. For example, a page might be ranked first in terms of visual appeal for the total sample, but fifth for visual appeal in a particular subsample. Table F1 shows the squared correlations between ranked visual appeal responses to each web page by participants in these subsamples and rank of overall visual appeal mean responses to pages by all 31 participants. For example, the first row shows that the squared correlation between the web page ranks for the first subsample of 10 subjects and the ranks for the total sample was .93. The last column in Table F1 shows the means of the squared correlations between subsample ranks and total sample ranks for each sample size.

Table F1. Squared correlations between web page ranks in the subsamples and the total Sample with mean squared correlations calculated for each sample size.

Subsamples	r^2	Means of subsample squared correlations
N = 10, Sample 1	.93	
N = 10, Sample 2	.90	.92
N = 9, Sample 1	.94	
N = 9, Sample 2	.93	.93
N = 8, Sample 1	.94	
N = 8, Sample 2	.90	.92
N = 7, Sample 1	.92	
N = 7, Sample 2	.94	.93
N = 6, Sample 1	.92	
N = 6, Sample 2	.89	
N = 6, Sample 3	.87	
N = 6, Sample 4	.93	
N = 6, Sample 5	.90	.90
N = 5, Sample 1	.92	
N = 5, Sample 2	.92	
N = 5, Sample 3	.92	
N = 5, Sample 4	.86	
N = 5, Sample 5	.91	.91
N = 4, Sample 1	.83	
N = 4, Sample 2	.90	
N = 4, Sample 3	.90	
N = 4, Sample 4	.85	
N = 4, Sample 5	.86	.87
N = 3, Sample 1	.82	
N = 3, Sample 2	.85	
N = 3, Sample 3	.86	
N = 3, Sample 4	.85	
N = 3, Sample 5	.77	.83
N = 2, Sample 1	.74	
N = 2, Sample 2	.87	
N = 2, Sample 3	.89	
N = 2, Sample 4	.84	
N = 2, Sample 5	.82	.83
Individuals	--	.65

Table F1 shows that the relation between visual appeal responses by participants in the samples and all 31 participants begins to degrade with sample sizes of 4 participants or smaller. Squared correlations between web page rankings in the subsamples and the sample are generally above .90 for 5 or more participants. At 4 participants, the squared correlation is around .87; with sample sizes of 2 or 3 the squared correlation is around .83. The mean of the squared correlations between individual rankings drops sharply to .65. Based on these results, it seems that 5 participants may be an acceptable sample size to get web page visual appeal rankings similar to those obtained with a sample size of 31 participants.

The next four tables provide details of the data that went into the squared correlations in Table F1. Tables F2 and F3 compare visual appeal ratings for the 25 highest-ranked web pages in the complete sample with visual appeal for the corresponding web pages in a number of randomly-selected subsamples varying in size from $N=8$ to $N=2$. Tables F4 and F5 deal with similar comparisons for the lowest 25 ranked web pages.

Table F2 reports ranks of visual appeal ratings for the highest ranked web pages in the subsamples as they related to the ranks based on the complete sample for Study 2.

Table F2. Ranks for 25 Highest-Ranked Web Pages in the Total Sample and Subsamples of Different Sizes

Total sample	Web page ranks for randomly selected subsamples													
N = 31	N = 8	N = 7	N = 6	N = 5	N = 4	N = 3	N = 2							
1	1	1	1	1	1	1	1	1	1	2	1	4	1	1
2	3	3	2	2	4	3	4	2	3	1	3	13	27	2
3	2	9	10	5	5	8	6	16	15	9	6	1	3	6
4	6	4	11	3	6	9	8	13	12	3	10	5	8	5
5	8	8	8	4	15	2	2	11	10	6	4	6	2	14
6	5	17	5	9	2	11	9	4	2	14	7	3	5	4
7	7	2	4	7	3	6	11	5	5	8	5	22	13	18
8	4	5	12	6	8	16	18	9	17	7	16	8	18	7
9	9	6	3	12	16	4	3	3	6	12	21	17	22	9
10	10	15	7	13	7	20	10	7	7	15	14	7	4	10
11	16	7	6	15	24	5	15	6	8	11	18	10	7	16
12	12	12	22	11	21	10	12	22	25	4	13	24	10	11
13	13	11	19	10	12	7	13	14	16	5	19	12	23	8
14	22	14	9	16	18	15	22	12	13	13	8	16	21	17
15	14	13	17	8	13	17	16	8	4	10	2	18	14	19
16	15	20	25	14	22	18	5	25	23	19	22	2	9	3
17	23	25	16	24	9	14	7	10	9	24	11	23	15	13
18	19	21	14	25	10	12	14	18	21	26	24	9	11	20
19	20	24	15	19	11	24	21	17	14	17	9	14	20	21
20	24	10	13	21	23	21	19	20	20	21	17	19	6	24
21	17	16	21	20	14	13	20	15	11	16	15	21	17	12
22	21	22	18	23	17	22	17	21	18	23	12	15	29	22
23	18	23	20	17	20	23	23	19	19	22	25	11	26	15
24	11	18	26	18	19	19	24	26	26	18	20	20	24	23
25	25	19	24	22	26	25	26	24	24	20	23	25	12	25

Ranks were based on the mean visual appeal ratings for each web page across all participants and Phases 1 and 2 in the subsamples or total sample of Study 2. For example, the first column shows rankings of the 25 highest ranked pages for visual appeal for the total sample. The second column shows the ranks for the same web pages in column 1, but for the first subsample composed of 8 randomly selected participants.

The first 2 entries for the total sample are 1 and 2 representing the highest and second highest ranked sites for visual appeal based on the total sample. The second column shows corresponding ranks for the first subsample of 8 randomly selected subjects from the total sample. In column 2, the first two ranks are 1 and 3 indicating that the highest ranked web page for the total sample was also highest ranked site for this subsample, but the second highest ranked for the total sample was third highest ranked for this subsample.

For subsamples of size 2 through 6, of the five subsamples only two are presented in Table F2. The subsamples with the highest and lowest squared correlation of the five are shown. For example, of the five subsamples based on 6 participants shown in Table F1, subsample 3 ($r^2 = .87$) and 4 ($r^2 = .93$) are presented. Also, subsamples of size 9 and 10 are omitted from this table because there are very little differences between results for subsamples of size 8, 9 and 10.

Ideally, there would be direct correspondence between web page ranks in the subsamples and the total sample. To the extent that the ranks differ in the subsamples from those in the first column, the total sample is not representative of the subsamples.

Also, it would be ideal that out of the 50 web pages in Study 2, the 25 highest-ranked web pages in each subsample would be the same 25 highest-ranked web pages in the total sample. Ranks that represented unappealing web sites are shown in bold.

Rankings for the appealing web pages according to the total sample were generally appealing for the subsamples, even for very small subsamples, with a notable exception for one subsample with $N = 2$. In that subsample the second highest rated web page for the total sample was rated 27. That subsample based on two participants was also the

only subsample with more than one of the 25 most appealing web pages ranked among the 25 least appealing web pages with three of the top 25 pages misclassified as unappealing.

Table F3 reports the *differences* between web page ranks in the subsamples and the complete sample for Study 2.

Table F3. Differences between Subsample and Total Sample ranks for the Twenty-Five Highest-Ranked Web Pages in Phase 1, Study 2

Total Sample N = 31	Differences in web page ranks for randomly selected subsamples													
	N = 8		N = 7		N = 6		N = 5		N = 4		N = 3		N = 2	
1	0	0	0	0	0	0	0	0	0	-1	0	-3	0	0
2	-1	-3	0	0	2	-1	-2	0	-1	1	-1	11	-25	0
3	1	-7	-7	-2	2	-5	-3	13	-12	-6	-3	2	0	-3
4	-2	-2	-7	1	2	-5	-4	-9	-8	1	-6	-1	-4	-1
5	-3	3	-3	1	10	3	3	-6	-5	-1	1	-1	3	-9
6	1	-3	1	-3	-4	-5	-3	2	4	-8	-1	3	1	2
7	0	4	3	0	-4	1	-4	2	2	-1	2	15	-6	-11
8	4	-4	-4	2	0	-8	-10	-1	-9	1	-8	0	-10	1
9	0	1	6	-3	7	5	6	6	3	-3	-12	-8	-13	0
10	0	-5	3	-3	-3	-10	0	3	3	-5	-4	3	6	0
11	-5	7	5	-4	13	6	-4	5	3	0	-7	1	4	-5
12	0	-8	-10	1	9	2	0	10	-13	8	-1	12	2	1
13	0	-3	-6	3	-1	6	0	-1	-3	8	-6	1	-10	5
14	-8	7	5	-2	4	-1	-8	2	1	1	6	-2	-7	-3
15	1	4	-2	7	-2	-2	-1	7	11	5	13	-3	1	-4
16	1	2	-9	2	6	-2	11	-9	-7	-3	-6	14	7	13
17	-6	-6	1	-7	-8	3	10	7	8	-7	6	-6	2	4
18	-1	-3	4	-7	-8	6	4	0	-3	-8	-6	9	7	-2
19	-1	-6	4	0	-8	-5	-2	2	5	2	10	5	-1	-2
20	-4	7	7	-1	3	-1	1	0	0	-1	3	1	14	-4
21	4	3	0	1	-7	8	1	6	10	5	6	0	4	9
22	1	0	4	-1	-5	0	5	1	4	-1	10	7	-7	0
23	5	4	3	6	-3	0	0	4	4	1	-2	12	-3	8
24	13	7	-2	6	-5	5	0	-2	-2	6	4	4	0	1
25	0	1	1	3	-1	0	-1	1	1	5	2	0	13	0
mDiff	0	0	-.12	0	.04	0	-.04	-.12	-.16	-.04	0	0	-0.88	0
lmDiff	2.48	4.00	3.91	2.64	4.68	3.60	3.32	4.00	4.91	3.56	5.04	4.96	6.17	3.52
Range of Diffs	21	15	17	14	21	18	21	0	24	16	25	29	39	24

Note. mDiff = Mean differences from the total sample for each subsample. lmDiff = Mean absolute differences from the total sample for each subsample. Diff = range of differences from the total sample for each subsample.

Ideally, there would be no differences between rankings, indicated by differences of 0. A negative difference indicates that a site was ranked lower by the subsample than by the total sample. For example, the site that was ranked second by the total sample was ranked third by the first subsample of 8 so the entry is -1. Mean differences, shown in the third row from the bottom, were very small even for small sample sizes, from - 0.88 to 0. This very small mean difference reflects the fact that the 25 appealing web pages were broadly rated as appealing, that is, they generally fell within the set of appealing web pages (ranks at or above 25) for all subsamples. Therefore the differences were within a set of 25 pages so the positive and negative differences offset each other.

Mean *absolute* differences were calculated by summing the absolute values of all the differences and dividing by 25. Because the 25 appealing pages were generally rated as appealing by even the small subsamples, mean absolute differences in ranks more accurately indicate the differences within the 25 most appealing web pages. The mean absolute differences are also relatively small with some indication of an increase in the magnitude of differences as the subsample sizes get smaller.

Table F4 reports rankings for the 25 lowest-rated pages between the subsamples and the total sample. (Ranks were based on the mean responses to each web page across all participants in the subsamples or total sample in both test phases of Study 2). As in Table F2, there are 9 instances of web pages ranked among the complete sample's 25 lowest-ranked web pages overlapping with the subsamples' 25 highest-ranked web pages. These instances are indicated in bold fonts. The final row of Table F4 reports the squared correlation between ranks for the subsamples and the total sample.

Table F4. Ranks for 25 Lowest-Ranked Web Pages in the Total Sample and Fourteen Subsamples

Total sample	Web page ranks for randomly selected subsamples													
N = 31	N = 8	N = 7	N = 6	N = 5	N = 4	N = 3	N = 2							
26	27	28	38	29	27	26	25	37	33	35	32	38	34	26
27	35	36	23	34	30	29	27	23	22	37	38	33	35	34
28	28	27	34	33	37	27	28	36	44	31	33	45	19	30
29	30	33	27	32	31	31	34	28	31	32	31	27	33	32
30	29	32	28	28	25	34	35	29	28	27	29	30	31	31
31	32	30	30	27	36	28	29	31	29	28	28	37	28	33
32	33	34	31	36	39	32	33	34	37	42	42	36	25	36
33	34	35	32	26	32	37	32	42	46	25	27	34	36	28
34	26	26	37	30	29	35	36	33	30	29	26	43	37	29
35	36	38	29	31	34	30	30	30	34	34	36	29	32	39
36	31	31	36	37	35	39	44	43	39	33	34	35	39	38
37	39	41	35	43	38	42	38	38	38	39	37	32	43	35
38	42	46	33	39	43	33	31	27	27	36	35	39	44	41
39	40	29	44	38	46	41	43	44	43	30	30	44	16	46
40	43	40	40	44	41	38	40	39	41	38	41	42	40	40
41	38	37	39	41	33	40	37	45	40	40	43	46	42	37
42	46	42	42	42	28	36	46	41	32	45	47	26	30	27
43	37	44	45	35	42	45	42	40	36	41	46	31	46	42
44	45	43	41	46	45	43	45	35	35	46	45	28	47	45
45	41	45	43	45	44	44	41	32	42	48	44	41	45	44
46	44	39	46	40	40	46	39	46	45	43	39	40	38	43
47	47	47	47	48	49	48	48	47	47	47	48	48	49	50
48	50	48	48	50	47	47	50	49	48	50	50	49	48	48
49	48	50	49	47	50	50	49	48	49	44	40	50	50	47
50	49	49	50	49	48	49	47	50	50	49	49	47	41	49
r ²	.94	.90	.92	.94	.87	.93	.92	.86	.83	.90	.85	.77	.74	.89

Table F5, analogous to Table F3, reports the *differences* between web page ranks in the subsamples and the complete sample for Study 2.

Table F5. Differences between Subsample and Total Sample ranks for the Twenty-Five

Total Sample N = 31	Differences in web page ranks for randomly selected subsamples													
	N = 8	N = 7	N = 6	N = 5	N = 4	N = 3	N = 2							
26	-1	-6	-12	-3	1	0	1	-11	-7	-9	-6	-12	-8	0
27	-8	1	4	-7	3	-2	0	4	5	-10	-11	-6	-8	-7
28	0	-5	-6	-5	9	1	0	-8	-16	-3	-5	-17	9	-2
29	-1	2	2	-3	2	-2	-5	1	-2	-3	-2	2	-4	-3
30	1	-1	2	2	5	-4	-5	1	2	3	1	0	-1	-1
31	-1	3	1	4	5	3	2	0	2	3	3	-6	3	-2
32	-1	2	1	-4	7	0	-1	-2	-5	-10	-10	-4	7	-4
33	-1	-5	1	7	-1	-4	1	-9	-13	8	6	-1	-3	5
34	8	0	-3	4	-5	-1	-2	1	4	5	8	-9	-3	5
35	-1	6	6	4	-1	5	5	5	1	1	-1	6	3	-4
36	5	1	0	-1	-1	-3	-8	-7	-3	3	2	1	-3	-2
37	-2	-2	2	-6	1	-5	-1	-1	-1	-2	0	5	-6	2
38	-4	1	5	-1	5	5	7	11	11	2	3	-1	-6	-3
39	-1	-2	-5	1	7	-2	-4	-5	-4	9	9	-5	23	-7
40	-3	0	0	-4	1	2	0	1	-1	2	-1	-2	0	0
41	3	-2	2	0	-8	1	4	-4	1	1	-2	-5	-1	4
42	-4	6	0	0	-14	6	-4	1	10	-3	-5	16	12	15
43	6	-1	-2	8	-1	-2	1	3	7	2	-3	12	-3	1
44	-1	2	3	-2	1	1	-1	9	9	-2	-1	16	-3	-1
45	4	-1	2	0	-1	1	4	13	3	-3	1	4	0	1
46	2	1	0	6	-6	0	7	0	1	3	7	6	8	3
47	0	0	0	-1	2	-1	-1	0	0	0	-1	-1	-2	-3
48	-2	0	0	-2	-1	1	-2	-1	0	-2	-2	-1	0	0
49	1	-1	0	2	1	-1	0	1	0	5	9	-1	-1	2
50	1	1	0	1	-2	1	3	0	0	1	1	3	9	1
mDiff	0	0	0.12	0	-0.04	0	0.04	0.12	0.16	0.04	0	0	0.88	0
lmDiff	2.48	2.08	2.36	3.12	3.64	2.16	2.77	3.96	4.33	3.80	4.00	5.68	5.36	3.12
Range of Diffs	16	12	18	15	23	11	15	24	27	19	20	33	31	22

Note. mDiff = Mean differences from the total sample for each subsample. lmDiff = Mean absolute differences from the total sample for each subsample. Diff = differences from the total sample for each subsample.

Lowest-Ranked Web Pages in Phase 1, Study 2

The mean differences were very small, from 0 to .88. The mean absolute differences are generally also quite small. The mean absolute differences tended to increase slightly as subsample sizes decreased below 5 participants. These findings also add validity to the use of sample sizes with as few as 5 participants but not fewer.

Appendix G

Principal Components Analysis of Web Page Characteristics

A Principal Components Analysis with a Varimax rotation was conducted in an attempt to determine the structure underlying the web page characteristics. The results are presented in Table G1. Two components were identified.

Table G1. Summary of Items, Component Loadings, and Eigenvalues for Varimax Orthogonal Two-Component Principal Components Analysis Solution for Non-Expert-Rated Web Page Characteristics from Study 2.

Item	Principal component loading	
	1	2
1. Good-Bad Design	.99	.10
2. Interesting-Boring	.99	-.04
3. Good-Bad Color	.97	.11
4. Good-Bad Layout	.97	.14
5. Imaginative-Unimaginative	.96	-.05
6. Simple-Complex	-.14	.97
7. Clear-Confusing	.65	.70
Eigenvalues	5.27	1.38

The first component was based on five visual characteristics: *interesting-boring*, *good design-bad design*, *good color-bad color*, *good layout-bad layout*, and *imaginative-unimaginative*. The second was based on one visual characteristic: *simple-complex*. The *clear-confusing* scale loaded moderately on both components. A reliability analysis confirmed the presence of the 5-characteristic grouping with an Alpha coefficient of .97.

Five of the seven characteristics, the five that were highly correlated with each other and with visual appeal in the previous bivariate correlation analyses, clearly grouped together to measure what might be called design quality. The second component was based on the complexity characteristic. The original intent was to use the components to understand how the visual characteristics were related and then relate them to the visual appeal ratings. However, the principal component analysis did not add to our understanding of the interrelations beyond what was already known from the individual bivariate correlations, so the relation between the principal components and rated visual appeal was abandoned.