

Aesthetics and credibility in web site design

David Robins^{a,*}, Jason Holmes^{b,1}

^a Kent State University, Information Architecture and Knowledge Management, 316 Library, Kent, OH 44242, United States

^b Kent State University, School of Library and Information Science, 314 Library, Kent, OH 44242, United States

Received 18 September 2006; received in revised form 12 February 2007; accepted 13 February 2007

Available online 12 April 2007

Abstract

Web sites often provide the first impression of an organization. For many organizations, web sites are crucial to ensure sales or to procure services within. When a person opens a web site, the first impression is probably made in a few seconds, and the user will either stay or move on to the next site on the basis of many factors. One of the factors that may influence users to stay or go is the page aesthetics. Another reason may involve a user's judgment about the site's credibility. This study explores the possible link between page aesthetics and a user's judgment of the site's credibility. Our findings indicate that when the same content is presented using different levels of aesthetic treatment, the content with a higher aesthetic treatment was judged as having higher credibility. We call this the *amelioration effect* of visual design and aesthetics on content credibility. Our study suggests that this effect is operational within the first few seconds in which a user views a web page. Given the same content, a higher aesthetic treatment will increase perceived credibility.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Credibility; Aesthetics; Web design; Human perception and judgment

1. Introduction

Web sites often provide the first impression of an organization. For many organizations, web sites are crucial to ensure sales or to procure services within. When a person opens a web site, the first impression is probably made in a few seconds, and the user will either stay or move on to the next site on the basis of many factors. Two of these factors are page aesthetics and a user's judgment about the site's credibility. This study explores the possible link between these two concepts.

Credibility judgment of a web site's information and overall content is a critical issue for those presenting information or selling products online. If a web site is not perceived as credible, it is unlikely it will be used. There is keen competition for goods, services and information on the web; therefore, it is vital for corporate survival to project credibility in a web presence. Because the web is a visual medium, the first credibility cues

* Corresponding author. Tel.: +1 330 672 5852; fax: +1 330 672 7965.

E-mail addresses: drobins@kent.edu (D. Robins), jholmes@slis.kent.edu (J. Holmes).

¹ Tel.: +1 330 672 0007; fax: +1 330 672 7965.

are perceived very quickly. Before any reading or other cognitive processes take place, preconscious judgments based upon visual design elements are already made.

Aesthetic treatments on web sites often require considerable time and expense; therefore, it is important to have some guidelines for designers. The guidelines are not meant to stifle creativity but only to define parameters in which designers' creativity can contribute to the overall effectiveness of a web site.

This research focuses on how users perceive two types of aesthetic treatments applied to the same content in a given scenario: low aesthetic treatment and high aesthetic treatment. A low aesthetic treatment (LAT) is one in which content is simply placed on a web site without professional graphic design. There may be graphical elements and some page layout meant to help the reader comprehend the content, but the elements and layout are crudely implemented. Our hypothesis is that this type of treatment creates a "low-budget" impression in the user, and a concomitant feeling that the content in the site is not credible.

On the other hand, a high aesthetic treatment (HAT) presents a professional look and feel appropriate to the organization it represents. Sites employing HAT employ principles of layout to enhance communication and strategically and professionally use color and graphics to build brand and concept. The pages in these sites convey professionalism and care in how they are presented. These pages should immediately invoke confidence, enjoyment, or some other positive emotion within users that makes them want to stay on the site. Our hypothesis is that this type of design will create a lucid impression of the site's intentions and invoke in users a feeling that the content in the site is credible.

The term, "credibility," is used here to describe the extent to which users trust the informational content on a web site. Credibility is increasingly important with regard to web-based content because virtually anyone can publish information on the web. This situation is what [Warnick \(2004\)](#) calls an "authorless" environment: traditional notions of the reliability or trustworthiness of "source" (what one is reading or hearing) break down and other criteria for credibility must be employed. Prior to this sort of publishing "anarchy," readers could be reasonably certain about the veracity of facts, opinions, and other trust variables because most publications had gone through some sort of editorial process in order to be published. This is not to say that all traditionally published materials were completely accurate and trustworthy, but, in general, more was known about the process by which the information arrived in print.

2. Research questions

In order to explore the impact of aesthetics on credibility judgments, we ask the following questions:

1. How quickly do people make credibility judgments when visiting a web site?
2. To what extent does a HAT produce high credibility evaluations?
3. To what extent does a HAT produce a positive increment of credibility evaluations?

The first research question (RQ1) is intended simply to produce a description of how quickly subjects are able to formulate a credibility judgment when confronted with a web page. Research question 2 (RQ2) addresses the extent to which ratings for HAT are high and ratings for LAT are low. Research question 3 (RQ3) asks whether a HAT produces positive increments of credibility judgments, even if both HAT and LAT are rated negatively. The difference between these two questions is important to this study. RQ2 will show whether users will react positively to HAT and negatively to LAT whereas RQ3 deals with the amelioration effect of a HAT on credibility judgments.

3. Related research

Credibility, for the purpose of this research, is limited to the believability or trustworthiness of information found in the World Wide Web. The focus on web-based information is important because such information is often freely contributed to publicly available space without being subjected to peer-review or editorial processes that, in general, improve its veracity. This leaves a greater burden of credibility judgment on users.

Credibility as a theoretical construct has been discussed in the field of Communication under the name of "source credibility theory" (SCT). SCT will serve as the overall framework for this study, but recent research

in the field of Human–Computer Interaction (HCI) and under the broad umbrella of Information Science has adopted and expanded on source credibility to explain interaction with information and interaction with information systems.

Below is presented an overview of SCT and a discussion of recent hypotheses of emotional design by Norman (2004), with research on authority and reputation in Information Science. A careful reading of this literature has shown that credibility judgments may occur on different levels of perception and criteria, classified here as *visceral* and *cognitive*. These criteria are framed below on the basis of available selected literature.

3.1. Foundations of credibility theory and related literature

Hovland and Weiss (1951) were the first to produce empirical evidence that the believability of a message is strongly influenced by its source. Their study demonstrated that the same content presented by two different sources (a known expert, “trustworthy,” and one of a more dubious, or “untrustworthy,” nature) was perceived on different ends of a credibility continuum. Over time, however, people tend to forget the nature of the source and remember only the content of the message.

Berlo, Lemert, and Mertz (1969) extended Hovland and Weiss’s (1951) work by identifying “safety,” “qualification,” “dynamism,” and “sociability” as factors related to source credibility. Although sociability did not have strong factor loadings in their analysis, the other concepts did. Of the remaining three factors, it is notable that two of the three, safety and dynamism, appear to be more visceral-type reactions, whereas qualification can be verified through analytical processes. Similarly, four factors emerged from another study (Whitehead, 1968), three of which corresponded to expertise, trustworthiness, and dynamism. The fourth factor that loaded on speaker characteristics such as “open-minded,” “objective,” and “impartial” was not as strong as the other three, but did suggest that source credibility might include criteria other than the accepted three.

SCT reveals that people tend to evaluate the credibility of communication primarily on the basis of the communicator’s expertise, trustworthiness and dynamism, and, to a lesser extent, on various other criteria. If this report were to focus on spoken communication, we would be primarily interested in dynamism. In our case, web users similarly make judgments on factors such as “authority” (Rieh, 2002) (similar to expertise) and “reputation” (Toms & Taves, 2004) (similar to trustworthiness).

In web-based communication, the closest link to dynamism would be unwritten factors such as design and aesthetics. Researchers at the Stanford Persuasive Technology Lab, who have conducted extensive studies on the phenomenon of web credibility (Fogg et al., 2001) were surprised to find the extent to which the dynamism of a web site mattered to users. Comments elicited from users were categorized and the largest category was “design and look,” indicated by 46.1% of respondents. The next highest category was of a similar nature: 28.5% indicated that the “information design” of a site contributed to their credibility judgments. So, nearly 75% of respondents reported making credibility judgments on the basis of content presentation rather than evaluation of the content’s/creator’s authority, trustworthiness, reputation, or expertise.

A conceptually related report by Princeton Survey Research Associates International (2005) for Consumer Reports Watch had somewhat different findings. This report dealt with trust and credibility of various types of web sites, but from the standpoint of consumer safety issues, such as people’s feelings about using credit cards online, in addition to the believability of online information. The study found that consumers were quite concerned about the authenticity of information on web sites based on criteria such as safety of personal information, trust, identifiability of sources and ease of navigation.

The differences between the Consumer Reports study and the Stanford study were focus and method. The Consumer Reports study focused on trust and safety issues associated with web use and used a survey to gather their information. The Stanford study focused on credibility and employed a web-based survey, and elicited comments from users rather than relying strictly on a survey. Elicitation of comments allowed the researchers to hear directly from users what criteria they used – to hear it in their own words. Although surveys are useful and appropriate data gathering techniques, they are limited in that users respond to statements usually without any other input of their own.

So far it has been shown that credibility judgments result from a number of different processes. Sometimes people must consciously determine if a source is trustworthy or expert, and other times dynamic or visual

properties influence judgments at a preconscious/precognitive level. In some cases, these preconscious judgments may be the result of experience and expertise on the part of the user; in other cases, people may make judgments based on unrelated experience (such as when they see an individual for the first time and make a snap judgment about that person's character) (Gladwell, 2005).

For this study, it is useful to borrow terminology from Norman (2004), who breaks down reactions to design in three experience levels: visceral, behavioral and reflective. Visceral experience in design is an immediate powerful reaction to a design. In describing various brands of bottled water bottle designs, he asks,

How does one brand of water distinguish itself from another? Packaging is one answer, distinctive packaging that, in the case of water, means bottle design. Glass, plastic, whatever the material, the design becomes the product. This is bottling that appeals to the powerful visceral level of emotion, that causes an immediate visceral reaction: "Wow, yes, I like it, I want it." It is, as one design explained to me, the "wow" factor (pp. 64–65).

The behavioral level is experienced during the use of a design. The presentation of the experience is less important than the ease and practicality of use. Whereas visceral design tries to immediately capture the user's attention, behavioral design seeks to hold the user through ease of use and ease of learning. It may be, however, that users will transcend the behavioral level and use objects that do not perform well because of some emotional attachment to the object. This represents the reflective level, and design in this area is highly analytic and cognitive. It represents an attempt to make a design better by incorporating the experience of users and their knowledge of goals and objectives of the product or service. For example, a web site designer may have listened to criticism of the navigation system, or may have experienced problems first hand, and is now contemplating how to make the navigation system of the site more efficient.

Norman, Ortony, and Russell (2003) bolster the importance of emotion and enjoyment in people's interactions with objects in everyday life. Tractinsky, Katz, and Ikar (2000) found that high aesthetic treatments on ATM interfaces positively influenced users' perceptions of the devices' usability. This confirms notions put forth by Dion, Berscheid, and Walster (1972) who found that people who are considered physically attractive are more likely to be perceived as better mates, more successful, more competent, and overall more desirable.

In the following two sections, we frame prior research on credibility into cognitive and visceral categories with regard to research objectives in the studies. Norman's (2004) design levels provided this framework and it is our object in this study to focus on visceral criteria for credibility judgments.

3.2. Cognitive criteria

The factors related to expertise and trustworthiness identified by SCT relate to what we will call "cognitive" criteria for the credibility judgments of web-based resources. Cognitive criteria would be a user's perception of the authority, reputation, or expertise of the information's source. Wilson (1983), referring to authority, influence and credibility, states simply,

All I know of the world beyond the narrow range of my own personal experience is what others have told me. It is all hearsay. But I do not count all hearsay as equally reliable. Some people know what they are talking about and others do not. Those who do are my cognitive authorities (p. 13).

This statement, in essence, defines the cognitive criteria about which we are talking here. Wilson contends that we build up our beliefs in what constitutes cognitive authority throughout our everyday private lives, educational systems, professions, the various "small worlds" (p. 148) in which we exist and belief systems.

In their review of the literature in a discussion of authority, Rieh and Belkin (1998) identified concepts that impact authority judgments: accuracy, currency, novelty, relevance, completeness, validity, reliability, and comprehensiveness. Rieh (2002) focused on the authority of web-based information. The results of her think-aloud protocols indicated that her users, across search topics, used authority-based criteria such as the name of an organization, the URL of a source, or the content found on a web site to determine credibility.

Often, information must be measured against standards known by the user, and thus credibility criteria may require time and analysis for a user to employ. Similarly, Toms and Taves (2004) directed their study to find how people assessed a web site's reputation and whether various search engines tended to retrieve sites with better reputations. Reputation, as defined in their paper, can be objectively assessed on criteria such as authority, recommendations, and ranking. Among their findings, they demonstrated a correlation between whether people would recommend a site or return to it themselves and their assessment of a site's reputation. This finding is consistent with Norman's (2004) behavioral design level: for people to be willing to recommend or reuse a resource they must find it reputable and usable.

Also concentrating on cognitive criteria, Tillotson (2002) developed a questionnaire that found that undergraduates were able to use sophisticated credibility criteria. These students claimed to use criteria such as author expertise and URLs from reputable organizations. Liu's (2004) questionnaire also elicited a relationship between criteria such as "information content," "authorship," "layout and structure," and "website and usage". The lattermost has to do with a site's reputation and its number of site users. Liu points out that although content that has a positive visual experience is only slightly more likely to induce a positive perception of credibility. "Documents having a nice layout or containing credentials of the author(s) have a positive impact on credibility".

3.3. *Visceral criteria*

The thrust of this research is to focus on "visceral" criteria, an area in which few studies have been done. Viscerally-based credibility judgments occur without conscious analytical cognitive processes. They are based on highly subjective reactions to stimuli presented when a user views a web site. For example, the combination of colors, layout, overall aesthetic treatment, fonts, use of bulleted lists, or presentation of tabular data may impact a person's credibility judgment. These judgments will be more difficult for users to explain, and they relate to such factors as dynamism, trustworthiness (if based on intangible factors such as first impressions), and sociability.

In any case, viscerally influenced criteria are primarily visual and not cognitive, so the impact of the visual experience is closer to that described by Maturana and Varela (1980) in his scenario of the autopoietic process of a frog's reaction to seeing a fly within reach. He determined there was no cognitive process by which the frog decided to snap up the fly. Rather, the flick of the tongue prompted by the sight of the fly is an action facilitated at the level of the nervous system and not the level of brain thought processes. Similarly, Gladwell (2005) summarized research on rapid cognition that begins to explain how people can make quick judgments that often are correct. Lindgaard, Fernandes, Dudek, and Brown (2006) found that significant judgments about the acceptability of a web site are made within 50 ms. This is certainly not enough time for cognitive processes to occur in an analytical or reflective manner. They also demonstrated that "visual appeal" was the prime determiner of a positive reaction to a web site. Wathan and Burkell (2002) presented a similar notion in their model of the credibility judgment process. They identified "surface credibility" (visceral) and "message credibility" (cognitive): the former addresses appearance issues that are quickly processed, and the latter requires further analysis to evaluate more objective criteria such as expertise and accuracy.

4. Research design

This study focuses on visceral credibility judgments rather than on judgments that take place at the cognitive level. To that end, an experiment was constructed to elicit precognitive responses to a set of stimuli. A variety of data collection and analysis techniques were used in order to address the research questions posed for this study. The overall intentions were to determine the extent to which people make credibility judgments when first viewing a web site, and whether such judgments varied as a function of the design/aesthetic treatment of the site.

These aims required a means of determining the direction of the judgment (credible or not credible) and the magnitude of the judgment. The study also required stimuli (web sites) of varying design/aesthetic treatment that subjects could judge. The design of the web sites needed to be carefully controlled so that judgments could be compared across subjects for the same stimuli. Consequently, it was decided that a judgment of the effects

of design/aesthetic treatment would be more informative if the same content was presented with different designs.

Three steps needed to be completed in order to carry out the study: select stimuli, select subjects and procedural design.

4.1. *Selecting the stimuli*

It was decided that the stimuli should be of the same subject matter (we chose “web accessibility” for no particular reason) so that subjects could get into a mindset of judging credibility within a single domain. Our intention was to compile a set of web pages from this domain. This was done with the foreknowledge that we would eventually ask subjects to rate credibility on the basis of whether they would use the information on the web sites (stimuli) to write a paper for a class.

A Google search on the terms “web accessibility” produced a large group of results. Stimuli were chosen from the results on the basis of moderate to high aesthetic treatment in the visual design. The determination of whether an aesthetic treatment was moderate to high was made on the basis of researcher judgment, since there is no objective measure of the degree of this characteristic. The important thing to consider, however, is that the designs found on the retrieved web sites were starting points for the study. As will be explained below, we stripped the pages of their aesthetic treatment to create a page with the same content but less aesthetic treatment than the original.

Most of these sites were informational in nature or were the sites of consultants offering services in the area of accessible web design. The next task was to make two versions of each site selected: one left as it was found on the web, and one with reduced aesthetic design. To do this, each of the 21 (to give us a large number of stimuli—the uneven number of sites chosen has no meaning) landing pages selected were saved on a local computer, opened in an html editor, and stripped of its visual enhancements. None of the content was altered, only the visual design. This left us with 21 pairs of pages (i.e., 42 pages total). To make these pages consistent, each of the 42 pages was opened in a web browser and saved as an image file that could not be altered as easily as an html file. The image was contextualized by also showing the browser window to give subjects the feeling they were browsing the web. Of the web sites chosen for stimuli, eight were “.com,” seven were “.org,” three were “.gov,” three were “.edu,” and one was “.net.” Table 1 shows examples of two sets of stimuli and it can be seen that the original (high) had much of the aesthetic design stripped (low).

Finally, the images were arranged in random order. We were concerned that a fatigue effect might skew the results, so we made two stimulus sets to show subjects: one of 42 from the original randomization, and the other a reverse of the original randomization. Odd-numbered subjects were shown the original randomized set, and even-numbered subjects were shown the reversal of the randomized set.

4.2. *Subjects*

Twenty subjects (14 Females, 6 Males) were chosen for this study from a pool of Library and Information Science (LIS) graduate students, although six were undergraduates from a variety of majors. Since web accessibility had only been covered in one unit within one LIS elective course, it was assumed that the subjects would not have much knowledge about web accessibility.

Selecting 20 subjects for the study allowed for a high degree of statistical power. The 20 subjects made a total of 840 credibility judgments (20 subjects \times 42 stimuli). The power analysis was performed on the assumption that we would administer *t*-tests of means of two samples (credibility judgments on a set of high aesthetic stimuli and on a set of low aesthetic stimuli). An acceptable power statistic is Power = >0.80 on a range of zero to one, and our power statistic was Power = 0.99.

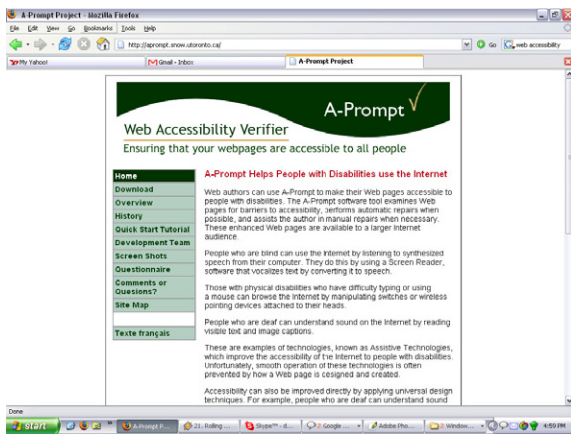
4.3. *Procedure*

Subjects were shown each of 42 images in sequence and asked to quickly judge the site’s credibility. We shuffled the order of the images so that the pairs of differently treated content would not be shown side-by-side.

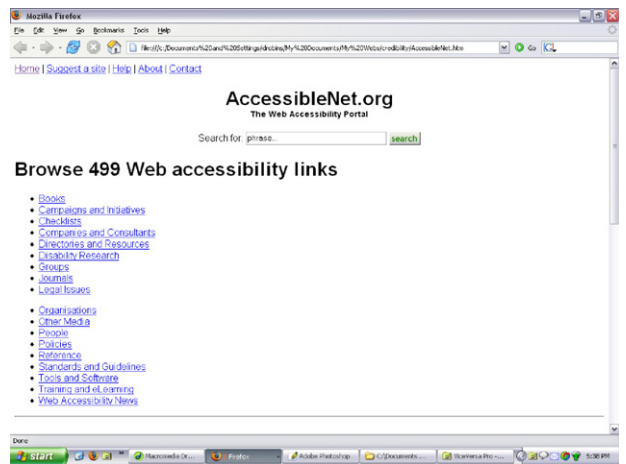
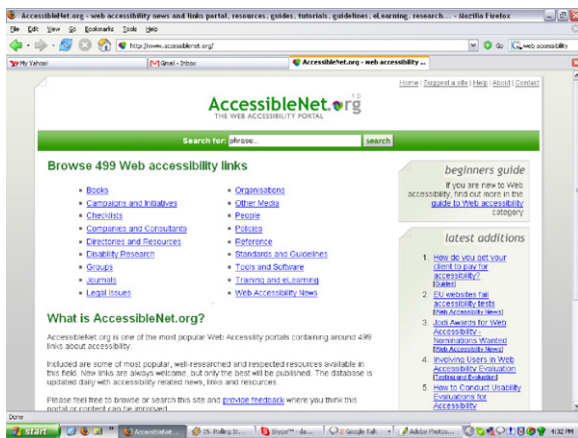
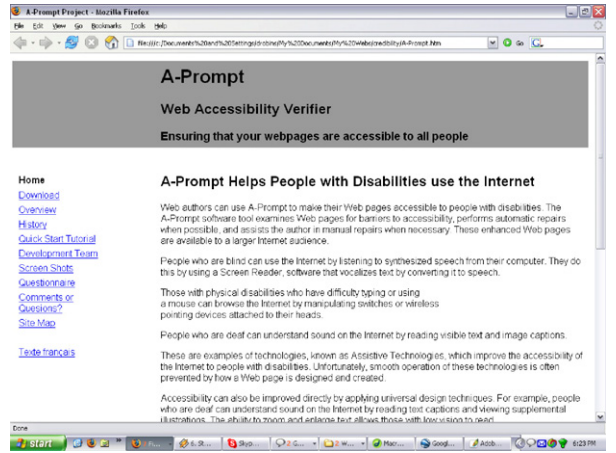
Table 1
Stimuli examples

Example of two stimuli – high and low aesthetic versions

High aesthetic



Low aesthetic



We did not tell subjects that the purpose of the study was to judge credibility on the basis of visual design, only that they were to judge each site's credibility on first impressions. As mentioned above, two sets of the stimuli (set 1 and set 2) were created, one in reverse order of presentation from the other. Set 1 was presented to odd-numbered subjects and set 2 to even-numbered subjects. This step was taken to control for fatigue and ordering effects among subjects, although the total test time was very short: the average session for each subject was approximately 3 min 38 s—average response time was 3.2 s for 42 stimuli, plus 42 cross hair images for 2 s each. A cross hair image is an image with a white background and a vertical and horizontal line crossed in the center of the image. This image is displayed between stimuli in order to get the user to refocus on the center of the screen and be ready for the next stimulus. Stimulus images remained visible until subjects indicated their credibility judgments.

Subjects indicated credibility judgments by moving a dial to the right for a positive judgment and to the left for a negative judgment. The dial was programmed to register judgments on a 14 point scale (1 through 7 (right direction) for positive judgments and -1 through -7 (left direction) for negative judgments). The dial device itself was built with 14 programmable positions on the dial, each of which could be assigned a value.

Before the images were shown to subjects, the users were assured of their anonymity and told to assume they were students in a course and had been assigned to write a paper on “web accessibility.” They were told

that the images that would appear on the screen were results from a Google search on that topic. In addition, they were read the following:

What we want you to do is to use your first hint of judgment and turn the dial to the right to indicate more credible, and to the left to indicate less credibility. For example, if you feel the information on the site looks credible or non-credible, turn the dial in the corresponding direction. A turn all the way to the right indicates maximum credibility and a turn all the way to the left indicates maximum non-credibility. You may turn the dial anywhere in between to indicate any degree of credibility.

Remember, you are judging the credibility of the web site based on your first impression.

The computer on which the study was performed was able to collect screen capture (video) with eye tracking (the results of which are not reported in this paper), the time for each judgment, and the values registered by the dial. Each image was displayed until the user moved the dial. The value that was assigned as the magnitude estimate for each image was the last in the direction turned by the subject before he or she allowed the dial to return to the center position. Each image display was separated by a neutral image with a cross hair in the middle displayed for 2 s before the next web site image was displayed. This was done so that the subjects would have a break between each “real” image, and have their eyes in a neutral, centered position for the next “site”.

In addition to the stimulus/response data, we collected a small amount of demographic data (school status, age range, and sex) and administered a brief questionnaire to determine subjects’ opinions about how certain visual cues impact their feelings about credibility on a web site. We administered this only after the stimuli had been shown to them.

4.4. Limitations

As with any experimental procedure, there are limitations in our study. First, the sample of subjects taken was a convenience sample rather than a random sample. This means that our results may not be generalizable outside the population of LIS Graduate students used in the study.

Second (and related to the first), we stated that LIS students would not be very knowledgeable about web accessibility, and this may be a point of contention. If it turns out that they are knowledgeable, it is possible they would know what resources are credible and what ones are not. However, even if they are aware of what resources are available, the research design we have presented here shows them the resource twice: once with its normal aesthetic treatment and once as a stripped down version. These are shown in random order and along with 20 other pairs. Subjects are instructed to make the judgments at their first feeling one way or another. We also chose stimuli that, for the most part, are not well known or associated with well-known organizations. If, on the other hand, students are not knowledgeable, then our assertion that their lack of knowledge will not prejudice their judgments holds.

A third point regarding the limitations of our study might be that LIS students are trained, or at least more likely, to read to determine credibility. However, the fact that LIS students were our subjects makes our case for adequate credibility judgments stronger. That is, even people who are inclined to read and judge at a cognitive level, as shown below, made visceral judgments that varied in accordance with changes in aesthetic treatment.

5. Results

Analysis of the data collected was conducted using a combination of Microsoft Excel and SPSS. This section presents the analysis results by overall observations and research questions.

Table 2 shows that, in general, the higher the aesthetic quality of a site, the higher the ranking of credibility. Furthermore, a *t*-test of the average ratings of the high and low aesthetic groups shows that the difference in these means is statistically significant ($p < 0.001$). Although the rankings were done on a scale of -7 to $+7$ (zero included), and the raw difference was not large, the value is still statistically significant. The small raw difference in such a large potential range could be explained by the fact that the differences in aesthetic treatments in the high and low set were not so great as to elicit strong enough reactions to take advantage

Table 2
Average credibility score and time for HAT and LAT

	HAT	LAT	Overall mean	<i>t</i>	<i>p</i>
Credibility rating	1.05	−0.55	0.25	4.282018	0.000113
Time taken to judge credibility	3.49 s	3.36 s	3.42 s	0.786	0.437

Note: * indicates significant difference between means at $\alpha = 0.05$ level.

of the full range of +7 to −7 on the dial. The relatively large standard deviation also indicates that there was not a strong division in judgment between the high and low aesthetic stimuli. However, the fact that subjects showed a tendency to rank credibility positively with higher aesthetic treatment, and the fact that the difference was statistically significant, indicates the need for further study. Generally speaking, a HAT is associated with a higher credibility ranking.

5.1. RQ1: How quickly do people make credibility judgments when visiting a web site?

Judgment rating time was measured from the time the stimulus appeared on the screen to the moment the subject turned the dial to indicate their ranking. The average time for this action to take place was 3.42 s. Overall, the 21 HAT stimuli averaged 3.49 s and the 21 LAT stimuli averaged 3.36 s. A *t*-test revealed this is not a statistically significant difference ($t = 0.78$, $p = 0.44$). Table 3 shows there were no significant differences within the HAT and LAT in the same set.

Three of our subjects were extreme outliers who carefully viewed each image to the point of reading the text. Their average response time was 9.79 s (6.18, 9.47 and 13.70 s). Not only are they statistical outliers, but their long average response time is due to reading of the content presented instead of evaluating credibility based on their initial visceral impression of credibility. When they are excluded from the overall calculation of the mean, the average response time per stimulus was 2.30 s, nearly a full second faster. A *t*-test on the different mean response time with and without outliers, however, revealed no statistically significant difference ($t = 1.47$, $p = 0.15$).

Table 3
Average time for HAT and LAT by image set

Image set	Time (s)		<i>t</i>	<i>p</i>
	Mean HAT	Mean LAT		
1	3.14	2.85	1.028	0.317
2	2.91	2.73	0.612	0.548
3	4.30	2.46	1.389	0.181
4	3.64	3.06	0.732	0.473
5	3.31	2.94	0.543	0.594
6	3.74	3.26	0.362	0.722
7	4.47	3.79	0.615	0.546
8	4.19	2.59	1.658	0.114
9	4.51	3.12	1.564	0.134
10	2.85	3.43	−1.064	0.301
11	3.05	2.88	0.487	0.632
12	3.57	3.36	0.545	0.592
13	3.61	3.96	−0.605	0.553
14	3.50	4.18	−0.650	0.523
15	2.61	3.74	−1.365	0.188
16	3.40	3.87	−0.710	0.486
17	2.82	3.50	−1.590	0.128
18	3.48	3.02	0.730	0.475
19	3.52	4.63	−1.288	0.213
20	3.50	3.22	0.706	0.489
21	3.17	3.92	−1.780	0.091

Note: * indicates significant difference between means at $\alpha = 0.05$ level.

5.2. RQ2: Does a HAT produce high credibility evaluations?

The second research question asks about the impact of aesthetics on credibility judgments. Specifically, it asks whether a HAT results in a positive (i.e., above zero rating) credibility judgment. Conversely, we expect a LAT to produce negative ratings for credibility. Therefore, everything else being equal, we expected to see positive credibility ratings for sites with a HAT and negative for those with a LAT. *T*-tests on mean judgments across subjects for each pair of stimuli determined there was a statistically significant difference in the ratings on the high and low aesthetic stimuli in seven of the pairs ($p < 0.05$) (shown in Table 4). Average rating scores for each image set are shown in Fig. 1.

Table 4
Average credibility score and differentials for HAT and LAT

Image set	Score		Score differential	<i>t</i>	<i>p</i>
	Mean HAT	Mean LAT			
1	2	-1.7	3.7	3.588	0.002*
2	0.5	-0.6	1.1	1.000	0.330
3	0.3	-2.05	2.35	2.283	0.034*
4	2.25	-1.9	4.15	3.983	0.001*
5	0.85	-1.4	2.25	1.723	0.101
6	1.75	0.75	1	0.960	0.349
7	1.45	-1.05	2.5	3.455	0.003*
8	1.5	1.1	0.4	0.396	0.696
9	0.1	-0.75	0.85	1.039	0.312
10	1.7	1.1	0.6	0.634	0.534
11	1.85	-1	2.85	3.707	0.001*
12	0.55	-0.55	1.1	1.273	0.218
13	2.3	2.35	-0.05	-0.050	0.960
14	1.4	-0.8	2.2	1.519	0.145
15	3.55	-0.32	3.23	3.728	0.001*
16	0.25	-0.5	0.75	0.510	0.616
17	-2.2	-0.1	-2.1	-2.840	0.010*
18	1.65	-0.4	2.05	1.741	0.098
19	0.75	-0.75	1.5	1.189	0.249
20	-0.85	-1.8	0.95	0.884	0.388
21	-0.1	-1.55	1.45	1.250	0.226

* Indicates significant difference between means at $\alpha = 0.05$ level.

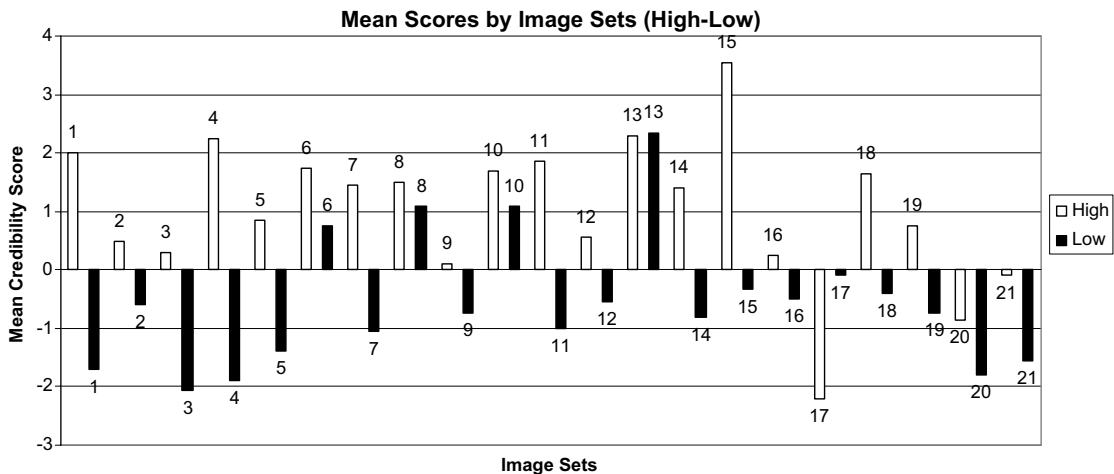


Fig. 1. Average scores for high aesthetic (white) and low aesthetic (black) stimuli.

In 14 of the 21 (67%) image sets, we found positive ratings for a HAT and negative ratings for a LAT. In three cases (17, 20 and 21), both the HATs and LATs produced negative ratings and in four cases (6, 8, 10 and 13), both the HATs and LATs produced positive ratings. Therefore, seven image sets did not meet expectations. In all but two (13 and 17, see Table 4) of those image sets, however, the HATs were rated higher than the LATs. Image set 17 will be discussed later in the results section of this paper and it is shown in Table 5.

In essence, this research question asks whether people make a simple “up or down” vote on the impact of aesthetic treatment. Since a majority of the image pairs were rated in this way, it is possible to say there is a trend toward our expectations, though not an overwhelming one. In fact, it is this finding that gives more strength to our findings with regard to RQ3. In the next section, we will further discuss the role of aesthetic treatment, not as something that produces a simple “good or bad” factor in credibility judgments, but as one that produces incremental effects at the visceral level of judgment.

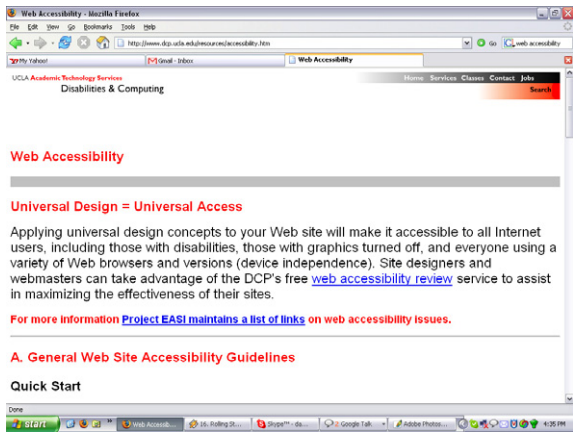
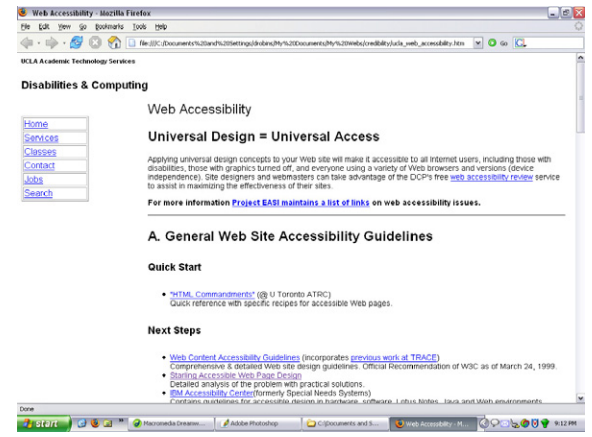
5.3. RQ3: Does a HAT produce a positive increment of credibility?

It is demonstrated in Fig. 1 that image sets 6, 8, 10, 13, 17, 20 and 21 did not meet the expectations that a HAT would produce a positive credibility rating and a LAT would produce a negative credibility rating. This question addresses whether the addition of aesthetic treatment improves the perception of credibility over the lower aesthetic version of the same content. The HAT stimuli did generally produce a higher incremental credibility score. The average score differential was 1.6.

Image sets 1, 3, 4, 7, 11 and 17 showed statistically significant differences in rating scores between the high and low aesthetic treatments of identical content (see Table 4). Specifically, in all but set 17, the differences were as expected (HATs produced higher credibility ratings than LATs). It seems that, in general, while a HAT may not necessarily produce a positive rating and a LAT may not necessarily produce a negative rating, a HAT will produce a higher rating than a LAT at the visceral level of experience.

There were two stimulus sets in which the LAT had the higher credibility rating. Set 13 had a differential of –0.05, indicating the LAT was scored as very slightly more credible than the HAT. Also of interest about set 13 is that it had equally high ratings for both low and high aesthetic treatments. The high aesthetic version contained graphics (logo and a state government seal) and another image which were stripped out of the low aesthetic version, making the latter quite plain by comparison. Therefore, we can speculate that subjects responded to cues indicating a government web site that existed in text and not graphics. For example, headings contained words such as “department of,” “office of,” and “Section 508”. These cues may have been quickly noticed and processed by subjects as credibility indicators.

Table 5
Image set 17

High aesthetic	Low aesthetic
	

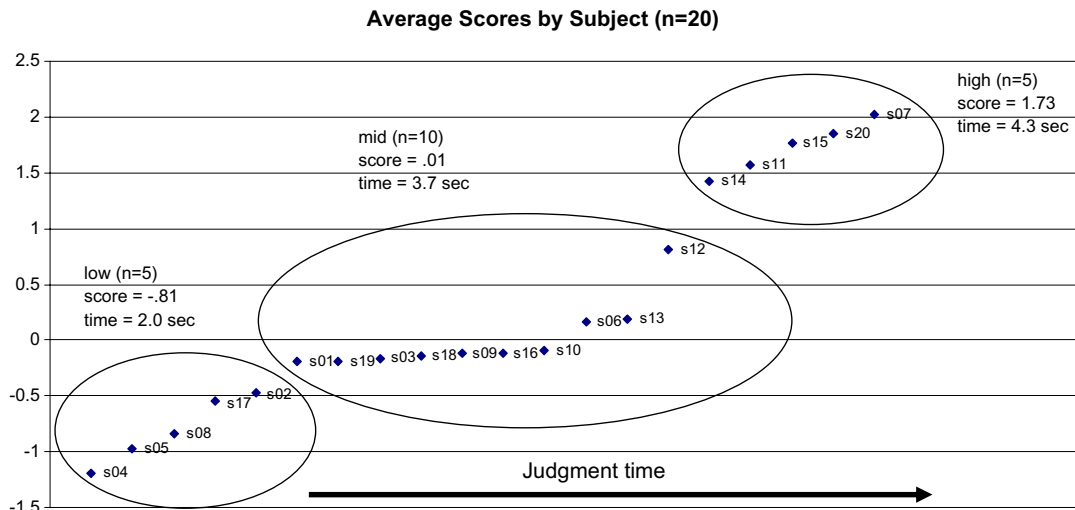


Fig. 2. Average rating score by subject.

On the other hand, set 17 had a differential of -2.1 , indicating the subjects viewed the LAT as much more credible than the HAT. The high and low scorings were statistically different from each other, but were reversed from what we expected. In fact, the high aesthetic image for set 17 had the lowest rating for credibility in the entire study. None of the low aesthetic images in the stimulus set had a lower credibility rating than 17's high aesthetic image. Table 5 shows the stimuli for image set 17. Evidently, the design of this site did not produce a credible representation of the content presented, unless some other factor influenced credibility judgment on this image set. This may be a case where the researchers actually improved the appearance of the HAT web site by making a simpler, clearer design in the LAT.

5.4. Subject groupings

Interestingly, subjects clustered into three groups with respect to rating scores. Five of the subjects tended to give overall high credibility ratings, five subjects tended toward low ratings and 10 subjects tended to hover near zero (see Fig. 2).

Those giving lower overall ratings were much quicker to do so (averaging 2.00 s to make their judgments), while those who gave higher overall ratings were slower to make judgments (averaging 4.3 s). The middle group had a mid-range average judgment time of 3.7 s. In future studies, it would be worth pursuing the notion that longer viewing results in a higher credibility judgment. There is no logical or empirical evidence (other than the present study) to support that notion, but few studies have been done in this area.

6. Discussion

The most important finding of this study was that regardless of how low a credibility rating was given to a web site, some type of aesthetic treatment increased the rating for the same content in 19 out of 21 cases (90%). In other words, when the same content is presented using different levels of aesthetic treatment, the content with a higher aesthetic treatment was judged as having higher credibility. We call this the *amelioration effect* of visual design and aesthetics on content credibility. Our study suggests that this effect is operational within the first few seconds in which a user views a web page. Given the same content, a higher aesthetic treatment will increase perceived credibility. The amelioration effect is probably limited to the visceral level of judgment and may diminish as people move into cognitive levels of judgment as their experience with the content broadens. As more time is spent with the content, more factors affecting credibility are experienced and employed in making judgments (e.g., source and authority).

As in many arenas, first impressions are crucial for web page content. Regardless of the quality or credibility of content, a poorly designed or aesthetically unappealing web page will likely produce a negative impression of credibility. In an environment such as the World Wide Web, where there are billions of documents and thousands of pages on a given topic, it is critical to present information in such a way that it does not produce a negative visceral judgment before the viewer even has a chance to engage the content at the cognitive level. People are quick to abandon a site and move on to one of any number of competing options. Lack of perceived credibility is surely one of the reasons for this behavior.

The longer people view a webpage, the greater the opportunity people have to find credibility in the content. They may, in fact, find credibility on further observation. If, at the visceral level, the design of a web site suggests non-credible information, the viewer might not stay with the site long enough for the content credibility to be perceived and judged at the cognitive level. Our study suggests that the aesthetic treatment will ameliorate this effect.

The time it took for subjects to respond with a credibility judgment (3.42 s on average) occurs during a time period that is consistent with a visceral response. It is unlikely that any type of complex cognitive analysis and evaluation can occur during that time. This is especially true if we consider the fact that without three outliers, the average response time would have been 2.30 s. We suspect that the subjects responsible for these outliers were engaged in cognitive behavior with regard to the credibility judgment, rather than reacting at the visceral level. In fact, we observed that they were reading the information on the stimuli before they ever reported a judgment.

Credibility rating scores are particularly difficult to interpret as it is unknown precisely what processes are at work at the visceral level to produce these judgments. Therefore, we can only speculate as to the cause of any particular rating. The overall credibility rating score averages were statistically significant when comparing the means of ratings on high and low aesthetic treatments. What we cannot tell from these data is what visual, content, design or other elements of these stimuli underlie these judgments. Any speculation in this area would result in suggestions for further research.

Ultimately, it will be useful to better explain why aesthetics has an impact on credibility judgments. The research described in this study was an attempt to find out whether varying the aesthetic treatment of informational content would impact credibility judgments regarding that information. In advertising, it has long been known that how a product is presented visually will impact sales, particularly with respect to branding (Vanden Bergh & Katz, 1999). However, while considerable effort has been placed on branding traditional consumer products such as automobiles, beverages, and computers, little is known about how to brand information resources. This study represents an early attempt to explore the relationship of the visual presentation of information and its perceived credibility. As it was shown that there is a relationship of that sort, further research could deal with the how and why.

7. Conclusion

This study investigated the relationship between aesthetics and credibility of information on web sites. In general, our findings were consistent with our expectations that high aesthetic treatment would produce high judgments of credibility. Our findings are similar to those of Tractinsky et al. (2000) in that positive reactions followed exposure to interfaces with higher aesthetic treatment. Although they were not investigating credibility, the end result is comparable. Liu (2004) indicated that design has only a minimal impact on credibility, but that assertion was based solely on a survey. Our data are from interactions with stimuli. Our study suggests that there is, at least on the statistical level, a significant interaction between design and credibility.

The importance of the findings presented here is that design has impact beyond decoration. Clearly there is some correlation between aesthetics/design and credibility judgments. We could say that aesthetics alone do not create a credible web site, but a higher level of aesthetic treatment incrementally increases the level of credibility. The question remains concerning exactly what features, elements or configurations of features and elements of design impact credibility judgment and in what way. It is a common (if latent) assumption that all serious web sites wish to be perceived as credible, believable, and trustworthy. This report is the first in a series meant to isolate these features so that designers can project the image necessary to support their aim, whether it be commercial, informational or educational. Our focus has been on a sort of “instant credibility,” or a

visceral-level judgment. Without that, it is possible users will not stay and use information on a site that may well be credible, but is not perceived as such.

Acknowledgements

We are indebted to Dr. Richard Rubin and Janna Korzenko for comments related to data collection and reading earlier drafts. We would also like to thank Dr. Stan Weirden for his help with source credibility theory.

References

- Berlo, D. K., Lemert, J. B., & Mertz, R. J. (1969). Dimensions for evaluating the acceptability of message sources. *The Public Opinion Quarterly*, 33(4), 563–576.
- Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, 24(5), 283–290.
- Fogg, B. J., Marshall, J., Kameda, T., Solomon, J., Ragnekar, A., Boyd, J., et al. (2001). Web credibility research: a method for online experiments and early study results. In *Proceedings of CHI 2001 extended abstracts on human factors in computing systems* (pp. 61–68).
- Gladwell, M. (2005). *Blink: The power of thinking without thinking*. New York: Little, Brown.
- Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *The Public Opinion Quarterly*, 15(4), 635–650.
- Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention web designers: you have 50 milliseconds to make a good first impression! *Behaviour & Information Technology*, 25(2), 115–126.
- Liu, Z. (2004). Perceptions of credibility of scholarly information on the web. *Information Processing and Management*, 40, 1027–1038.
- Maturana, H., & Varela, F. (1980). Autopoiesis and cognition: the realization of the living. In R. S. Cohen & M. W. Wartofsky (Eds.), *Boston studies in the philosophy of science* (Vol. 42). Dordrecht, Holland: D. Reidel Publishing Co.
- Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things*. New York: Basic Books.
- Norman, D. A., Ortony, A., & Russell, D. M. (2003). Affect and machine design: lessons for the development of autonomous machines. *IBM Systems Journal*, 42(1), 38–44.
- Princeton Survey Research Associates International (2005). Global expertise, local service. <<http://www.psra.com/globalexpertise.shtml>> Retrieved 14.09.2006.
- Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 55(8), 743–753.
- Rieh, S. Y., & Belkin, N. J. (1998). Understanding judgment of information quality & cognitive authority in the WWW. *Proceedings of the American Society for Information Science*, 61, 279–289.
- Tillotson, J. (2002). Web site evaluation: a survey of undergraduates. *Online Information Review*, 26, 392–403.
- Toms, E. G., & Taves, A. R. (2004). Measuring user perceptions of web site reputation. *Information Processing and Management*, 40, 291–317.
- Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, 13, 127–145.
- Vanden Bergh, B. G., & Katz, H. (1999). *Advertising principles: Choice challenge change*. Lincolnwood, IL: Lincoln.
- Warnick, B. (2004). Online ethos: Source credibility in an “authorless” environment. *American Behavioral Scientist*, 48(2), 256–265.
- Wathan, C. N., & Burkell, J. (2002). Believe it or not: factors influencing credibility on the web. *Journal of the American Society for Information Science and Technology*, 53(2), 134–144.
- Whitehead, J. L. (1968). Factors of source credibility. *Quarterly Journal of Speech*, 54, 59–63.
- Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Westport, CT: Greenwood Press.