Are First Impressions about Websites Only Related to Visual Appeal?

Eleftherios Papachristos and Nikolaos Avouris

Human-Computer Interaction Group, Electrical and Computer Eng. Dept.,
University of Patras, GR-265 00 Rio Patras, Greece
{epap,avouris}@ece.upatras.gr

Abstract. This paper investigates whether immediate impression about websites influences only perceptions of attractiveness. The evaluative constructs of perceived usability, credibility and novelty were investigated alongside visual appeal in an experimental setting in which users evaluated 20 website screenshots in two phases. The websites were rated by the participants after viewing time of 500 ms in the first phase and with no time limit in the second. Within-website and within-rater consistency were examined in order to determine whether extremely short time period are enough to quickly form stable opinions about high level evaluative constructs besides visual appeal. We confirmed that quick and stable visual appeal judgments were made without the need of elaborate investigations and found evidence that this is also true for novelty. Usability and credibility judgments were found less consistent but nonetheless noteworthy.

Keywords: Webpage design, aesthetic evaluation, credibility, visual appeal, perceived usability.

1 Introduction

The importance of appropriate aesthetic web design has been clearly shown by Lindgaard et al. [1] in a series of experiments about the immediacy of first impressions. Their findings indicate that first impressions about websites can be formed during the initial 50 ms of viewing and that they are highly stable over time. In a subsequent study Tractinsky et al. [2] replicated and extended the above study providing further evidence for the immediacy and consistency of aesthetic impressions.

These results had quite an impact on the HCI community because they suggested an elevated importance for website aesthetics. However, there is an ongoing debate about the nature of such aesthetic responses regarding the involvement of cognition. According to Norman [3] the visceral response to visual stimuli is merely an affective unconscious reaction about good or bad: a “gut” feeling. Hassenzahl [4] rejects the notion of visceral beauty stating that beauty judgments are “cognitive elaborations of the initial diffuse reaction” to stimuli. In that vein of thought cognition is required for aesthetic judgment. Additionally, that initial reaction may serve as a starting point for subsequent, more complex evaluation which often involves expectation and prior

P. Campos et al. (Eds.): INTERACT 2011, Part I, LNCS 6946, pp. 489–496, 2011.

experience. However, Lindgaard's et al. [1] and Tractinsky's et al. [2] results contradict to some extend the above by showing that their subjects could provide stable aesthetic evaluations in time periods too short to discern all of the stimuli details.

If first impressions are only positive or negative feelings about stimuli as Norman [3] and Hassenzahl [4] presume, then users wouldn't be able to distinguish between a set of high-level evaluative constructs. Any judgment would be a "halo effect" or a carry-over effect of that positive or negative impression to the other construct and evaluations should be highly correlated and not independent. If however, website users have predisposed concepts such as simplicity, symmetry and familiarity associated for example to usability perceptions then it is possible that judgments are a result of those individual intuitive criteria. If that is true then first impressions are not simple assessments of positive or negative feelings toward stimuli, but a bundle of quick and intuitive evaluations of several characteristics which are particularly important to the individual user.

However, from a designer's point of view it is important to understand the implications of website users' first impressions regardless of the origins of their formation. Are first impressions only about visual appeal? And if not, what else are users able to form opinions about in split seconds? In order to investigate if website users are able to form stable judgment about several website characteristics in a glimpse of an eye we had first to identify evaluative constructs previously linked to aesthetic matters. Literature research helped us identify: perceived usability [5,6], perceived credibility (trustworthiness) [7] and novelty [8,9] as appropriate constructs for the purposes of our study.

The objectives of this study were:

1. To investigate whether the formation of impressions about other high level evaluative constructs related to aesthetic design (perceived usability, credibility and novelty) is as quick as visual appeal, and how stable they are over time.
2. To examine whether their judgments on the evaluation constructs for the websites are independent or only covariations with visual appeal.

2 Method

Forty undergraduate university students (25 male, 15 female, aged 21 – 34, mean age = 23.9) participated in the study as partial fulfillment of the requirements in a human computer interaction course. The participants evaluated screenshots of 20 hotel websites. All participants reported having previous experience with hotel websites in general but none with the specific sample selected for the study. The selected hotel websites originated from a remote to the participants destination country (New Zealand) in order to minimize the possibility of prior sample familiarity. The website selection criteria were to have a balanced sample of good, average and bad designed websites. Although the selection process was subjective, post evaluation analysis showed that participants perceived our sample as balanced according to visual appeal. Unlike the studies of Lindgaard et al. [1] and Tractinsky et al. [2] we felt that our test material should belong to the same website domain in order to minimize possible confounding factors.

In addition, we had to reduce stimuli number to avoid participants' fatigue since they were asked more questions per website. Similar to Tractinsky et al. [2] we chose to replicate only the 500 ms condition of Lindgaard's et al. [1] experiment, which has been characterized as a time period short enough to form first impressions, but not long enough to evaluate other features such as semantic content [1][2].

2.1 Procedure

After participants were informed about the purpose of the experiment, specific instructions were given about the evaluative constructs (visual appeal, perceived usability, credibility and novelty) in order to ensure a unanimous understanding of them.

The evaluation took place on an eye-tracker (Tobii T60) using a specifically developed software. In the study's first phase the test websites where displayed as screenshots for 500 ms and were followed by a screen that contained the rating scales. We used an unmarked slider (from 0 to 100) as in [1] with the appropriate description on each end for each of the aforementioned evaluation criteria. Between each rating screen and each website screenshot a delay screen lasting for 1sec was placed. The delay screen contained a crosshair exactly in the middle of the screen in order to ensure that all users had the same viewing staring point. The software presented to each participant the website screenshots in a completely randomized order. There was no time limit while viewing the evaluation screen.

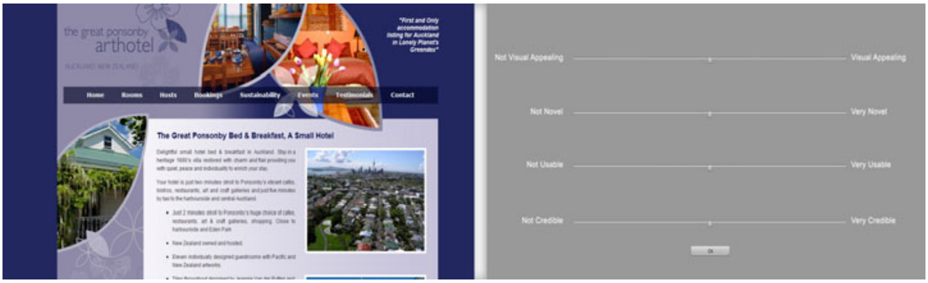

Fig. 1. Representative website (on the left), rating scales screen (on the right). A delay screen appeared between websites and ratings scales in each phase.

The second experimental phase was identical to the first one with the difference that there was no limit in displaying time. In this phase participants were asked to evaluate the same websites again on the same evaluation criteria after they viewed each website for as long as they wished. Screenshots were displayed in a new randomized order for each participant. The whole procedure lasted approximately 30 minutes for each participant.

3 Analysis

Since, 40 participants evaluated 20 webpages there was a total of 800 evaluations for each construct. As a first step of the analysis we examined the frequency distributions of user evaluations for each construct individually. Our concern was to examine whether user evaluations revealed sample skewness in a particular construct which might limit result generalizability [2,6]. The examination showed quasi- normal distributions for all constructs, which means that most evaluations for the entire sample were around the middle of the scale and fewer at the extremes. In figure 2 mean rating's for the entire sample in both phases are displayed, all evaluations were more favorable in the second phase except visual appeal. However, visual appeal was the only construct with significant difference (t(19) = 2.09, p = .05) between the two phases.

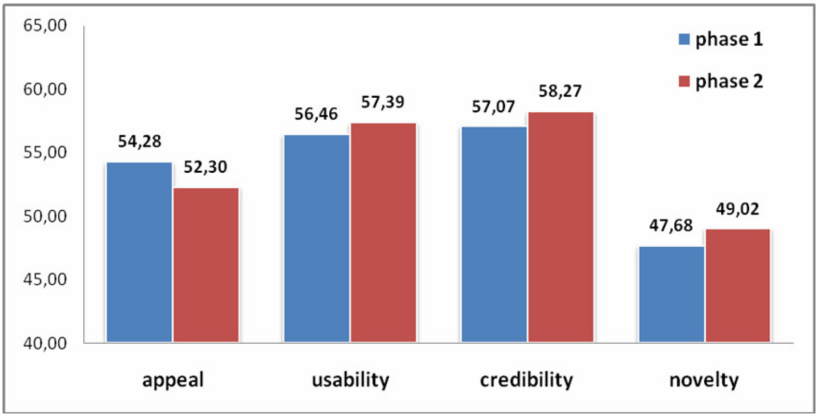

Fig. 2. Average evaluation of the entire sample in both phases

In a subsequent analysis we examined our data per website and looked at the correlations of average scores between the two phases for each construct. As it can be seen in table 1 the lowest correlations are for perceptions of usability (r = 0.64) with only 41.4% of the explained variance shared through the phases. The average ratings for the other constructs were highly correlated with explained shared variance ranging from 64.8% to 90.4%. Correlation of novelty perceptions were even higher (r = 0.951, p < .001) than of visual appeal. All correlations were significant and relatively high, indicating consistency for user evaluations between very short and long viewing periods when averaging over stimuli as in [1,2].

As a next step we looked into within-participants consistency by calculating the between phases correlations of each construct for each participant individually. In both [1] and [2] this analysis resulted in lower correlations than these aggregated over stimuli. Within-rater reliability, however, was notably lower in [2] (ranging from -0.09 to 0.9 with average correlation of 0.55) than in [1]. Our analysis yielded similar to [2] results indicating large variation in participant consistency (table 2). Participant reliability in visual appeal ratings ranged from r = .03 to r = .90 with an average of r = .521. The correlations of 13 participants fell below r = .50 and of 27 above. In total, 70% of the correlations were significant. Participant consistency was somewhat lower for the other constructs.

Table 1. Correlations of average scores

	Correlation	Sig.
Visual appeal	.864	.001
Per. Usability	.644	.002
Credibility	.805	.001
Novelty	.951	.001

Finally we investigated the between-construct relationship of websites mean ratings for each phase independently. In order for the constructs to have been judged independently we had to rule out that the evaluations were a result of simple covariation effects. If attractive websites were evaluated by participants high and the unattractive low on all constructs then the between construct correlation would be high. The ability of participants to differentiate between constructs, especially at the 500 ms condition, could indicate that evaluations are not simply a product of a positive or negative first impression. As depicted in table 3 credibility is positively correlated to visual appeal and to perceptions of usability in both experimental phases. The influence of visual appeal on perceived usability (r = .45*) wears off in the second phase (r = .19) which could indicate that more elaborate investigation is needed by participants in order to form perceptions of usability.

Table 2. Within participants correlations

	Mean Correlation	Range	Sig. Cor.
V. Appeal	.521	0.03 - 0.90	70%
P. Usability	.261	0.01 - 0.92	22.5%
Credibility	.332	0.03 – 0.83	35%
Novelty	.503	0.10 – 0.90	62.5%

The interesting result, however, is that novelty is significantly negative correlated to perceived usability and to some extent to credibility, but is positively correlated to visual appeal (significant only in second phase r=.49). It seems that novelty perceptions, which were proven relatively consistent both in within participant and in average rating, mediate the other evaluations. As shown in [8,9] slightly above average novelty perceptions are associated with attractiveness, while extreme deviation from the norms results to confusion and therefore low perceived usability.

These results were a first indication that participants could differentiate at least novelty perception from a positive or negative first impression that was formed in split seconds. However average between – construct correlations alone is not enough to indicate independence of perception. For that reason we examined if particular websites received differing scores for all or some of the evaluative constructs. For example, finding some websites rated highly in perceived usability and at the same time low in visual appeal could indicate that the evaluative constructs were judged independently from each other.

Table 3. Correlations between constructs Phase A and B

Phase A	V. Appeal	P. Usability	Credibility	Novelty
V. Appeal	1	.448*	.580**	.284
P. Usability		1	.859**	-.608**
Credibility			1	-.395
Novelty				1
Phase B	V. Appeal	P. Usability	Credibility	Novelty
V. Appeal	1	.191	.484**	.489*
P. Usability		1	.713**	-.470*
Credibility			1	-.250
Novelty				1

For the 500ms condition aggregated over website ratings showed that the seven most appealing websites were also rated high in perceived usability and credibility but received only moderate novelty ratings. Six of the seven less appealing websites were rated low or average on novelty but high in perceived usability. Results from the second phase were very similar regarding the above trend, except from some minor changes in the ranking order of the websites. Although, on average none of the websites scored at the two extremes (largest difference was visual appeal=58.9 and perceived usability = 28.7) we found large divergence of certain constructs in individual ratings. Averaged over designs constructs scores were used as within subject's variables in one way repeated measures ANOVA. The analysis revealed that construct differences were greater in the 500 ms (F(1,19) = 8.82, p < .008) than in the no time limit condition (F(1,19) = 5.62, p < .028). Post Hoc comparison showed that novelty ratings were significantly different from all other constructs in the first phase while only credibility and visual appeal differed significantly in the second. Although, constructs differences seem to vary between the two phases, the aforementioned results serve as first indicators of construct independence. However, further studies are required in order to fully understand their relationships.

4 Discussion

The above findings are indications that users form quickly reliable judgments about various websites characteristics. We found evidence that the formation of novelty perceptions in split seconds is particularly stable over time. It is certain that participants used inference and reflection while confronted with the rating scales since no time limit was imposed. Any kind of experimental setting can't avoid tempering with the natural circumstances in which judgments about websites are made. Participants have to formulate their opinion or give ratings on a scale which interferes with the natural process in which websites are viewed, judged and used. However judgments made during extremely short and long exposure shared high explained variance which means that similar conclusions are made between having only glimpse and after rigorous examination. In addition the judgments participants were able to make in this study are very different from simple reactions of liking or disliking.

On the other hand we found considerable differences in participants' ability to rate the websites under the experimental conditions. Three of them had no consistent rating in any of the evaluative constructs, most had only in one or two and only seven had significant correlations in all of them. The explanation for this could be that certain participants had strong, predefined notions about some constructs or strong likes or dislikes about design characteristics easily identified in the 500 ms condition (color, form, background texture). It is also possible that some participants had the ability to identify more visual attributes during the same timeframe than others.

The reliabilities concerning credibility and especially perceived usability were noticeably lower. Still the correlations reported aren't atypical in research in which human judgment process is involved. Although, we feel that usability judgments are more moderated by novelty we have to further investigate other alternative visual factors such us symmetry, complexity and order which have been previously linked with perceptions of usability

In addition, we confirmed that average visual appeal evaluations of web pages are very consistent. Furthermore, within participant consistency was considerably lower than [1] and similar to [2]. An explanation for that could be that Lindgaard [1] used a polarized sample; half the websites were "ugly" and half "beautiful". In addition, in experimental phases 2 and 3 of their study, a subset of the initial sample was used after keeping only the websites that were rated on the extremes by users in phase 1. In our and Tractinsky's [2] studies the sample was indented to be balanced in terms of attractiveness-beauty by following a quasi - normal distribution. As previously indicated by several studies [10,11] and clearly demonstrated by Tractinsky et al. [2] in the same context (website evaluation) extreme ratings are more easily generated by participants. Apparently, participants need more time to evaluate close to average stimuli since more elaboration is needed to identify flaws or positive characteristics before forming a final opinion.

5 Conclusion

The present study was able to replicate findings of [1,2] regarding the consistency of visual appeal evaluations of websites between extremely short and long exposure. Our aim was to extent previous research and to investigate the consistency of additional evaluative constructs related to website aesthetics. We found indications that participants were able to provide stable ratings for novelty and to some extent for credibility and perceived usability. Our findings support the initial hypothesis that besides attractiveness other aesthetic responses are also able to be made by website user in the first critical split seconds of first viewing.

As future work we indent to analyze the eye-tracking data gathered during the experiment in order to examine what participants were able to focus on during the 500 ms period. Also, implicit measures such as response latencies for each evaluative construct could, as in [2], further validate our results. Finally, we intend to investigate the relation of low level constructs such us symmetry, order, complexity balance and contrast to the high level constructs investigated in this study. Such an investigation could help identify which visual attributes have a stronger influence to certain aesthetic impressions and which are more stable during time.

References

1. Lindgaard, G., Fernandes, G.J., Dudek, C., Brownet, J.: Attention web designers: You have 50 milliseconds to make a good first impression! Behaviour and Information Technology 25(2), 115–126 (2006)
2. Tractinsky, N., Cokhavi, A., Kirschenbaum, M., Sharfi, T.: Evaluating the consistency of immediate aesthetic perceptions of web pages. International Journal of Human - Computer Studies 64(11), 1071–1083 (2006)
3. Norman, D.A.: Emotional Design: Why We Love (or Hate) Everyday Thinks. Basic Books, New York (2004)
4. Hassenzahl, M.: The interplay of beauty, goodness, and usability in interactive products. Human-Computer Interaction 19(4), 319–349 (2004)
5. Lavie, T., Tractinsky, N.: Assessing dimensions of perceived visual aesthetics of web sites. International Journal of Human-Computer Studies 60(3), 269–298 (2004)
6. Norman, D.A.: Introduction to this special section on beauty, goodness, and usability. Human-Computer Interaction 19(4), 311–318 (2004)
7. Fogg, B.G., Soohoo, C., Danielson, D., Marable, L., Stanford, J., Tauber, E.: How do people evaluate a Web site's credibility? Results from a large study. Persuasive Technology Lab, Stanford University, http://www.consumerwebwatch.org/news/report3_credibilityre-search/stanfordPTL_TOC.htm
8. Coates, D.: Watches tell more than time: product design, information and the quest for elegance. McGraw-Hill, London (2003)
9. Papachristos, E., Avouris, N.: The Subjective and Objective Nature of Website Aesthetic Impressions. In: Gross, T., et al. (eds.) INTERACT 2009, Part I. LNCS, vol. 5726, pp. 119–122. Springer, Heidelberg (2009)
10. Bassili, J.N.: The how and why of response latency measurement in telephone surveys. Jossey-Bass Publishers, San Francisco (1996)
11. Pham, M.T., Cohen, J.B., Pracejus, J.W., Hughes, G.D.: Affect monitoring and the primacy of feelings in judgment. Journal of Consumer Research 28, 167–188 (2001)