THE EFFECT OF WEB PAGE TEXT-BACKGROUND COLOR COMBINATIONS ON RETENTION AND PERCEIVED READABILITY, AESTHETICS AND BEHAVIORAL INTENTION

Richard H. Hall
University of Missouri, Rolla
rhall@umr.edu

Patrick Hanna
Matrikon Corporation
patrick.hanna@matrikon.com

Abstract

The purpose of this experiment was to examine the effect of different web page text/background color combinations on users' retention and subjective perception. One-hundred and thirty-six participants were randomly assigned to one of four groups: black text on a white background (BW), white on black (WB), light blue on dark blue (B), or teal on black (TB). They then studied two Web pages, with a color combination consistent with their group assignment. One page contained information on the Neuron and the other consisted of information on a fictitious TV/DVD player. After studying each page they completed a quiz and survey. Analysis of the data indicated: a) Retention scores did not differ as a function of text/background color combinations; b) BW and B pages received the highest ratings for readability, and TB the lowest; c) B pages received the highest ratings for the aesthetic qualities; d) BW pages were perceived as most "professional"; e) Subjective readability ratings significantly predicted retention; and f) Users view "professionalism" as more strongly related to readability than aesthetics. Taken together, these results indicate that the relationship between font/background color combinations and outcomes is complex and often inconsistent with web guidelines posed by "web gurus".

Introduction

Web Design Guidelines and Font-Background Color Combinations

The flexibility of the World Wide Web has made it very simple for developers to create text and background combinations of a variety of differing colors, not to mention background textures. Luckily the use of textured backgrounds has, for the most part, come and gone, most likely driven by popular demand (and empirical evidence, (Hill and Scharff 1999)). However, a myriad of different text-background color combinations still proliferate.

Web design guidelines often include recommendations for appropriate color combinations, many of which recommend high contrast between text and background with particular emphasis on the traditional black on white. "Web gurus" are quick to make definitive statements about design and readable text, as exemplified by Jakob Nielsen:

> Use colors with high contrast between the text and the background. Optimal legibility requires black text on white background (so-called positive text). White text on a black background (negative text) is almost as good. Although the contrast ratio is the same as for positive text, the inverted color scheme throws people off a little and slows their reading slightly. Legibility suffers much more for color schemes that make the text any lighter than pure black, especially if the background is made any darker than pure white (Nielsen 2000, p 125).

Unfortunately, Nielson does not offer any references for this statement. In fact, an examination of the small amount of research that exists on this topic indicates that the relationship between text-background color combinations and readability is not at all clear-cut.

Pre-WWW Research

Most of the research on readability of text on a computer screen pre-dates the World Wide Web and, thus, was conducted with monitors that were less effective in terms of luminance and luminance contrast, which turn out to be important factors in mediating the effect of font/background color combinations (Bouma 1980; Mills and Weldon 1987). However, this research provides a useful background, and, as we will see, most results are similar to more recent research involving the Web.

One of the most consistent findings is that the effects of colors on readability are not consistent (Radl 1980). For example, one study failed to find any significant difference among 24 different color combinations on performance with a text search task (Pace 1984). On the other hand, regardless of the specific color combination, higher levels of contrast appear to lead to better readability (Bruce and Foster 1982; Radl 1980).

WWW Research

There are very few empirical studies of readability of web pages, based on font/background colors (Hill and Scharff 1997). One exception is a series of two experiments conducted by Hill and Scharff (Hill and Scharff 1997; 1999).

In the most recent study (1999) Hill and Scharff varied the background texture, color, and saturation/lightness of a given page. Participants were required to search for specific objects within the page and reaction time for completion of the search was thought to be indicative of readability. In this study they used only black text, but varied background colors (blue, gray, and yellow). They found a significant main effect for color with better performance for the gray and yellow backgrounds than with the blue. This finding is consistent with the finding mentioned above that higher contrast conditions lead to better performance.

In an earlier study (Hill and Scharff, 1997) six color combinations were varied in addition to font type and word style (italicized vs. plain). Participants searched web sites to find a target word and, again, reaction time represented readability. A main effect for color was found with the best performance for green text on a yellow background and the worst for red on green. Though this finding appears to be consistent with the high contrast effect, it should be noted that black on white was one of the 6 combinations tested, and performance was better for green text on the yellow background. The finding that performance with Black on White was not as good as a chromatic color combination is inconsistent with the contrast effect and clearly inconsistent with Nielson's recommendation in the quote above. The results were consistent with the pre-web research, however, in that the effect of color on readability was not straightforward. In fact the main effect for color can actually be explained more accurately via a significant two-way (color X font) interaction. More specifically, the better performance for green on yellow was due to performance with Times New Roman font, while the performance was much worse for this color combination when Arial font was used.

The 1997 study also included a comparison of gray and white backgrounds, which was motivated by the fact that most web browsers at the time had gray backgrounds as a default. Due to the contrast effect one would expect that a white background would result in better readability. Therefore, they replicated the method of the first experiment with the exception that only black text and three different background colors (light gray, dark gray, and white) were used. Surprisingly, they found better performance with the gray backgrounds than with the white background, a finding, again, inconsistent with the contrast effect. (Ironically, despite these findings, the default background in web browsers these days is, of course, white.)

As a preliminary activity for the 1997 study, a web survey was conducted in which students were asked to rate the readability of 20 different color combinations on a 6-point Likert scale that ranged from "terrible" to "excellent". The following trends were found: a) black text on white background was rated higher than any other combinations; b) color combinations that included black were rated higher than those that did not; and c) combinations that consisted of darker text on a lighter background were rated more readable than their inverse (e.g., blue on white was rated higher than white on blue). These results demonstrate two interesting points about users' subject ratings of text-background color combinations on web pages. First, users tend to rate more familiar combinations as more readable. Second, subjective ratings are not necessarily consistent with performance. For example, though black on white was perceived as the most readable, it was not found to be the most readable in the empirical study that followed the survey.

Research Goals

All of the studies cited above use basic measures of readability, which usually consist of some variation on a single-word-search task. Though this is informative with respect to basic processing, it does not address higher-level outcomes of readability such as retention. Retention is a very important factor for the large number of information-based web sites that exist. It is, of course,

an important factor for e-learning applications, since the user's goal is usually to retain the information beyond the time the page is being read. This also applies to information included in e-commerce sites, since the users' tasks are often facilitated when they can retain information from page to page. Therefore, measures of higher level processing, such as retention, are an important next step in examining the impact of text-background color combinations.

A second important factor, which has been examined very little in previous research on these color combinations, is their effect on subjective perception. One exception is the Hill and Scharff study cited above in which participants responded to an on-line survey with subjective ratings of readability. Other important subjective perceptual factors that have not been examined in previous studies are aesthetic judgments and behavioral intention. Experts such as Nielsen have long expressed the importance of design simplicity and de-emphasized the importance of aesthetics as a component in usable designs (Nielsen 2000). However, Web design, like most design endeavors is a balance between the functional and aesthetic. Factors such as aesthetically pleasing color combinations can play an important role in generating positive affect, which may be particularly important for a commercial web site where a company is trying to encourage users to associate a given company brand with positive feelings. Leaders in the HCI filed, such as Don Norman, have recently focused on the need to consider aesthetics and emotion in design (Norman 2002). Aesthetic factors may serve to affect behavioral intention, which could presumably lead to behaviors that would be especially important for commercial sites, in particular purchasing.

In summary, the overriding goal of this research was to examine the effect of different text-background color combinations on users' retention and subjective responses for web pages with two types of content (academic/educational and commercial).

The specific research questions addressed were:

1. How does retention differ as a function of font-background color combination?
2. How do subjective ratings of readability differ as a function of color combination?
3. How do subjective ratings of aesthetics differ as a function of color combination?
4. How do subjective ratings of behavioral intention differ as a function of color combination? (only applied to the commercial content page).
5. To what extent do subjective ratings predict retention? (included behavioral intention with the commercial content)
6. What is the relationship among subjective ratings? (included behavioral intention with commercial content)

Method

Participants

One hundred and thirty-six students enrolled in General Psychology classes at a the University of Missouri – Rolla participated in this experiment as partial fulfillment of a research participation requirement for the class.

Materials

Stimulus Materials: Web Pages

Two different web pages were used as stimulus material for this experiment. One of these web pages covered information that is used in an introductory level neuroscience class and covered information on the Neuron. The other page advertised the "Hallaview 3000", which was a fictional TV/DVD player. This content was created from information gathered from a number of technology and entertainment web sites. The passages were relatively short; the Neuron page consisted of 338 words and the Hallaview page was 279 words.

Four different font-background color combinations were used for each of these sites: black text on white background (BW); white text on black background (WB); light blue text on dark blue background (B); and teal text on black background (TB) . The hexagonal codes for these colors were: black (000000); white (FFFFFF); light blue (DED9FB); dark blue (000066); teal (00FFFF).

Outcome Measures

A ten question, multiple-choice quiz was developed covering information on both web pages (Neuron and Hallaview). In addition, surveys were developed for both of the web pages. Students responded to questions on a 10-point Likert scale with 1 labeled "strongly disagree" and 10 labeled "strongly agree". Both surveys included the following five items:

1. The color combination made the text easy to read.
2. The color combination made the text easy to study.
3. I found the color combination pleasing to look at.
4. I found the color combination stimulating to the eye.
5. I found the color combination to be professional looking.

The following two items were also added to the Hallaview survey:

1. If I had available funds, I would like to buy this product.
2. The color combination made me want to buy this product.

Procedure

This experiment took place in ten experimental sessions, made up of groups of 10 – 30 students over the course of two semesters. For each session, students were randomly assigned to one of four-color conditions: BW, WB, B, or TB (see section on web pages above for description of colors). When students arrived, an introductory web site was displayed on their computers with written directions. The entire experiment was on-line and time was strictly controlled, so that students did not proceed to the first study page until told to do so. They then viewed the page for ten minutes, after which they were required to go to the quiz/questionnaire page for 10 minutes, etc. The content areas were counterbalanced so that, in every other experimental session, students studied the commercial page first, while in the other sessions, they studied the educational page first. The experimental session schedule is displayed in Table 1.

Table 1. Experimental Session Schedule

time	activity
0 - :10	Introduction, Consent
:10 – :20	Study Content 1
:20 – :30	Quiz & Questionnaire 1
:30 - :40	Study Content 2
:40 - :50	Quiz & Questionnaire 2

Results

Question 1: Retention

In order to assess the impact of font color on retention a series of two one-way between-subjects analyses of variance (ANOVA) were computed with experimental group (BW vs. WB vs. B vs. TB) serving as the independent variable and quiz score as the dependent variable. Neither of these ANOVAs were statistically significant.

Question 2: Readability Rating

In order to assess the impact of subjective readability ratings, a series of two Multivariate Analyses of Variance (MANOVA) were computed with experimental group serving as the independent variable and survey questions 1 and 2 ("easy to read" and "easy to study") for the neuron page serving as multiple dependent variables in one MANOVA and the same questions for the Hallaview page serving as the dependent variables in the other MANOVA.

The Neuron MANOVA was statistically signification $\Lambda 6,262) = .91$, $p < .05$. Due to the significant MANOVA, this was followed by univariate analyses, which consisted of two between-subjects ANOVAs with group as the independent variable and survey questions 1 and 2 each serving as the dependent variable for one ANOVA. Both the "easy to read", $F(3,135) = 4.23$, $p < .01$, and the "easy to study", $F(3,135) = 3.32$, $p < .05$, ANOVAs were statistically significant. For both ANOVAs, Tukey's post hoc tests

found that the BW and B groups significantly outscored the TB group, while the WB group did not significantly differ from other groups. The means are displayed in Figure 1.

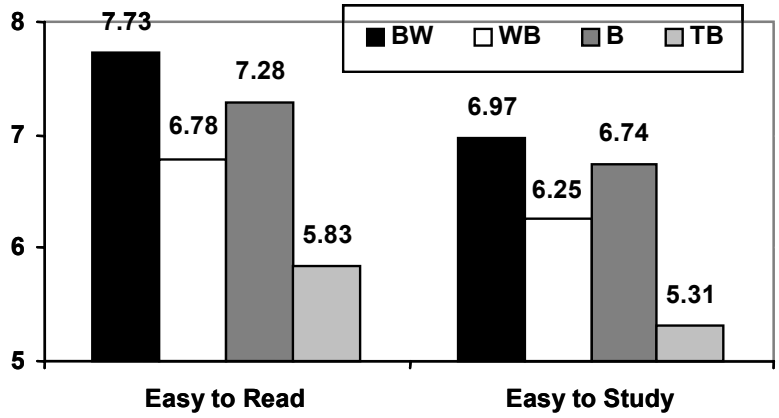

Figure 1. Subjective Readability Ratings as a Function of Experimental Group for the Neuron Page

The Hallaview MANOVA was also statistically significant $\Lambda(6,262)=.87$, $p < .01$. Due to the significant MANOVA, this was again followed by univariate ANOVAs with experimental group as the independent variable and the two readability ratings each serving as a dependent variable in one of the ANOVAS. Both the "easy to read" $F(3, 132) = 6.05$, $p < .001$ and "easy to study" $F(3,132) = 2.77$, $p < .05$ ANOVAs were statistically significant. A Tukey post hoc test for the "easy to read" ANOVA found that the BW and B groups scored significantly higher than the TB group, while the WB group did not significantly differ from other groups. For the "easy to study" ANOVA the BW group scored significantly higher than the TB group and no other post-hoc comparisons were significant.

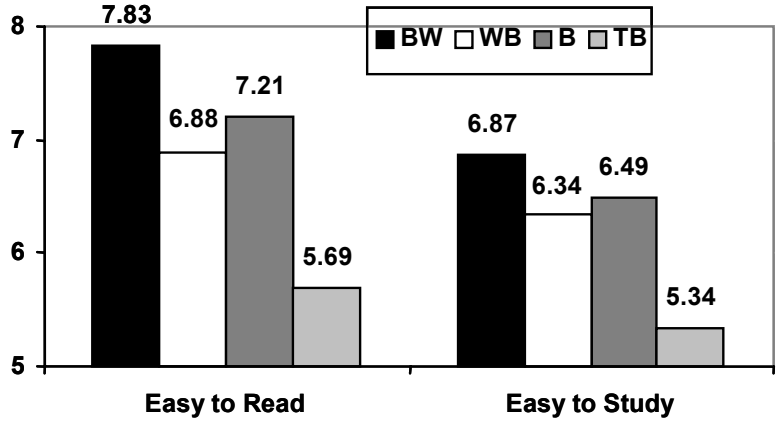

Figure 2. Subjective Readability Ratings as a Function of Experimental Group for the Hallaview Page

Question 3: Aesthetic Rating

In order to address the question regarding aesthetic ratings, two MANOVAs were computed with experimental group as the independent variable and ratings for questions three "pleasing to look at", four "stimulating to the eye", and five "professional looking" for the Neuron as the dependent variables in one MANOVA and these same questions for the Hallaview page as the dependent variables in the other. The neuron MANOVA was significant $\Lambda(9,316) = .62$, $p = .001$. This was followed by three one-way ANOVAs with experimental group as the independent variable and each of the three aesthetic ratings as the dependent variable in one of the ANOVAs. The "pleasing to look at", $F(3,132)=15.73$, $p = .06$, and "stimulating to the eye", $F(3,132)=2.60$, $p = .055$, ANOVAs were both marginally significant. The "professional looking" ANOVA was significant $F(3,132) = 16.23$, $p < .0001$. Tukey's post hoc tests, for the "professional" ANOVA found that the BW group had significantly higher ratings than

all other groups, the WB group had significantly higher ratings than the TB group, and no other mean comparisons were significant. The means for aesthetic ratings of the neuron page are displayed in Figure 3.

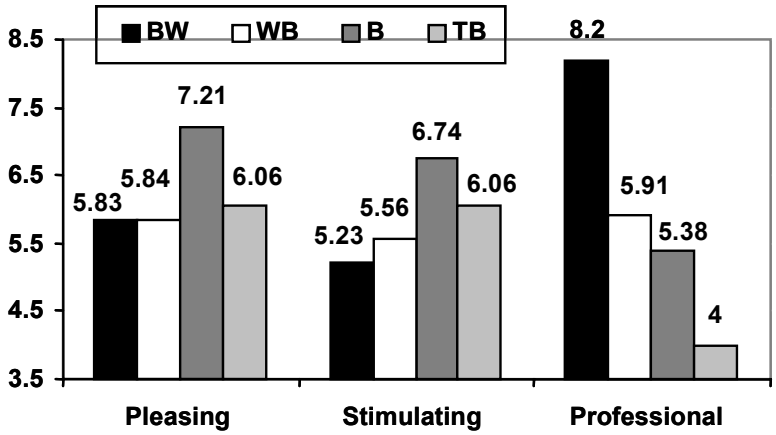

Figure 3. Subjective Aesthetic Ratings as a Function of Experimental Group for the Neuron Page

The Hallaview MANOVA for aesthetic ratings was also significant $\Lambda(9,314) = .59$, $p = .001$. This was followed by a series of ANOVAS, as above, with experimental group as the independent variable and each of the aesthetic ratings as a dependent variable in one of the ANOVAs. Only the "professional looking" ANOVA was statistically significant $F(3,132) = 18.32$, $p < .001$. For this ANOVA, Tukey's post hoc tests found that the BW group ($M = 8.27$) had significantly higher ratings than all other groups, the WB ($M = 6.10$) group had significantly higher ratings than the TB ($M = 4.06$) group, while no other mean comparisons were statistically significant. (For the B group, $M = 5.05$)

Question 4: Behavioral Intention

In order to examine the effects of behavioral intention as a function of experimental group a MANOVA was computed with experimental group as the independent variable and ratings on questions 6 ("I would like to buy this product") and questions 7 ("The color combination made me want to buy") as multiple dependent variables. These questions were only asked in the Hallaview questionnaire, so only one MANOVA was computed. This MANOVA was not statistically significant.

Question 5: Subjective Ratings Predicting Retention

In order to examine the degree to which subjective ratings predicted retention, two multiple regression equations were computed. In the first, the five neuron subjective ratings served as predictor variables with neuron quiz score as the criterion. In the second, the seven subjective ratings for the Hallaview page were the predictor variables and quiz score was the criterion. In each case, variables were entered in a step-wise fashion, with the criteria for entry set at p = .05. In the neuron equation, "easy to read" was significantly predictive of quiz score $t = 3.19$, $p < .01$. No other predictors met the criterion for entry. In the second equation, "easy to study" was significantly predictive of quiz score $t = 2.02$, $p < .05$ and no other predictors met the criterion for entry.

Question 6: Relationship Among Subjective Ratings

The examination of the relationship among the subjective ratings was driven by the assumption that the ratings for both pages would be related according to the categories posed: "readability", "aesthetics", and "behavioral intention" (the latter only for the "Hallaview" page). Therefore, two principal components factor analyses were computed, one for each content area, forcing two factors in the neuron analysis (readability and aesthetics), and three in the Hallaview analysis (readability, aesthetics, and behavioral intention). For both factor analyses a Varimax rotation was used in order to maximize independence among factors.

In the neuron factor analysis, the two factors accounted for 86% of the variance, with Eigen values of 3.31 and .98 for the two factors. In the Hallaview analysis, the three factors account for 78.5% of the variance with Eigen values for the three factors of 3.4, 1.18, & .91. The variable loadings for the rotated solutions are presented in Tables 2 and 3. In both cases, the factors loaded most strongly on the anticipated factor, with the exception of "professional looking", which loaded most strongly on the "readability" factor in both cases.

Table 2. Factor Loadings for Neuron Analysis (Rotated Solution)

Variable	Factor	
	Readability	**Aesthetics**
Easy to Read	**.77**	*(.52)*
Easy to Study	**.72**	*(.55)*
Pleasing to look at	*(.27)*	**.91**
Stimulating to the eye	*(.12)*	**.93**
Professional Looking	**.91**	*(.01)*

Table 3. Factor Loadings for Hallaview Analysis (Rotated Solution)

Variable	Factor		
	Readability	**Aesthetics**	**Behavior**
Easy to Read	**.77**	*(.49)*	*(-.02)*
Easy to Study	**.73**	*(.52)*	*(.06)*
Pleasing to look at	*(.22)*	**.88**	*(.16)*
Stimulating to the eye	*(.15)*	**.86**	*(.22)*
Professional Looking	**.84**	*(-.02)*	*(.23)*
Like to buy	*(-.03)*	*(.16)*	**.85**
Colors made me want to buy	*(.26)*	*(.13)*	**.78**

Discussion

With respect to the first experimental question, retention scores did not differ as a function of experimental group. This finding indicates that, at least based on retention, the importance of web page text-background color combinations do not make as much difference as web "gurus" would have us believe. Across both types of information, quiz scores did not significantly differ, whether the students had studied the traditional black text on a white background; it's inverse, white on black; the more "jarring" teal on black, or the low contrast light blue on dark blue. While it is important to be cautious when generalizing results from a controlled experiment such as this to every day Web behavior, the results do imply that web "guidelines" or "principles" often stated as fact, turn out to be more flexible when examined empirically. The equivocal and inconsistent nature of the effect of color combinations on objective outcomes is consistent with the empirical research presented in the introduction, which used lower level processing measures of readability (Bruce and Foster 1982; Hill and Scharff 1997; Pace 1984; Radl 1980).

Many group differences emerged with respect to subject ratings (experimental questions 2-4). In terms of readability, not surprisingly, users rated the black on white page as more readable than the teal on black page. More surprising was the fact that the light blue on dark blue page ratings were largely equivalent to the black and white page across three of the four readability ratings analyses that were conducted (see Figures 1 and 2). This is surprising first, because of the low contrast of this combination and second, because of the light text and dark background. The latter is inconsistent with the trend that dark text on light backgrounds was preferred in the study cited in the introduction (Hill and Scharff 1997). One possible explanation for this effect was that users viewed this light/dark blue page as aesthetically pleasing and this influenced ratings of readability as well. In fact, in terms of aesthetic ratings, at least with the Neuron information, the light/dark blue page faired the best, with respect to ratings for "pleasing to look at" and "stimulating to the eye" (figures 3 and 4). It's not surprising that users would rate a page with chromatic colors as more pleasing and stimulating, though it is interesting to note that the light/dark blue page was rated substantially higher than the teal on black page. One might think that the teal on black page would be perceived as more dramatic. It may be that the somewhat "jarring" nature of the teal on black page discouraged users from rating this as highly as the more subdued light/dark blue page, even on "stimulating to the eye". This implies that subtlety is an important factor in aesthetic design.

With respect to "professional looking", users clearly felt that black on white was more professional than other combinations across both content areas. Interestingly, the factor analysis, discussed below, indicates that this may be partly due to the fact that users did not tend to view "professional looking" as a characteristics of "aesthetics" in the sense that "pleasing" and "stimulating" represent this construct. The only case where the experimental groups did not significantly differ with respect to subjective ratings was for the ratings of behavioral intention (only conducted with the commercial group). Based on these results, it does not appear that color combinations have a strong influence on consumers' intended purchasing behavior.

The results of the two regression equations, which examined subjective ratings as predictors of retention (question 5), were somewhat inconsistent with research presented in the introduction (Hill and Scharff 1997). For both content areas, subjective readability ratings significantly predicted quiz scores. The "easy to read" and "easy to study" items significantly predicted quiz scores for the neuron and Hallaview equations respectively. Therefore, it appears that in some cases what users' say is consistent with what they do, assuming that the readability of a page would be related to the amount of information retained. This result must, however, be interpreted with caution, since it's important to note that none of the other subjective-rating variables included met the criteria for entry. Therefore, other ratings of readability, ratings of aesthetics, and behavioral intention were not strongly related to quiz scores once the variance was removed for the primary predictor. Although a significant relationship was found in both equations, the relationship between subjective ratings and retention was certainly less than overwhelming, when taking all of the subjective rating variables into account.

The relationship among the subjective rating variables turned out as anticipated in the factor analyses results for the most part (question 6). The readability, aesthetic, and behavioral intention items loaded strongly on factors they were thought to represent. The one noted exception was the "professional looking" item, which was considered to be a component of the "aesthetic" factor. As it turned out, "professional looking" was much more strongly related to readability items than to the other two "aesthetic" items ("pleasing to look at" and "stimulating to the eye"). This implies that users tend to lump together their perceptions of readability and professionalism of a page. This finding has important implications for those who are trying to present a professional image with web sites. It appears that designers who work to improve the readability of their pages are at the same time improving the professional image of the page as well.

Taken together, these results paint a complex picture of the relationship between the color of text and backgrounds on web pages and their impacts on the user. Moreover, this relationship is clearly more complex than one would gather from reading the opinions, guidelines, and principles posed by web "experts" and "gurus". One important implication is that these color combinations appear to make very little difference on some outcomes (i.e., retention). The impact is principally manifested in terms of subjective user perceptions. Users rate more traditional, and subtler color combinations as the most readable, and rate these more subtle colors as more pleasing and stimulating. They also clearly find the more traditional black on white background to be the most professional color combination. Contrary to previous research, a significant relationship between a some subjective ratings (i.e., readability) and objective outcomes was found. However, this relationship was not reflected in the majority of the subjective perception measures. Finally, these results provide us with some insight into the important but illusive construct, "aesthetics". If aesthetics includes characteristics like the "pleasingness" and emotional "stimulation" of the page, users clearly view the perceived qualities associated with "professionalism" as independent of such a conception of aesthetics.

References

Bouma, H. "Visual Reading Processes and the Quality of Text Displays," in *Ergonomic Aspects of Visual Display Terminals*, E. Grandjean and E. Vigliani (eds.), London: Taylor & Francis, 1980, pp. 673-702.

Bruce, M. and Foster, J.J. "The Visibility of Colored Characters on Colored Backgrounds in Viewdata Displays," *Visible Language*, (16:4), 1982, pp. 382-390.

Hill, A.L. and Scharff, L.V. "Readability of Screen Displays with Various Foreground/Background Color Combinations, Font Styles, and Font Types," *Proceedings of National Conference on Undergraduate Research*, Washington DC, 1999.

Hill, A.L. and Scharff, L.V. "Legibility of Computer Displays as a Function of Colour, Saturation, and Text Backgrounds," in *Engineering Psychology and Cognitive Ergonomics* (Vol. 4), D. Harris (ed.), Sydney: Ashgate, 1999, pp. 123-130.

Mills, C.B. and Weldon, L.J. "Reading Text from Computer Screens," ACM Computing Surveys, (19:4), 1987, pp. 329-358.

Nielsen, J. *Designing Web Usability: The Practice of Simplicity*, Indianapolis, IN: New Riders Publishing, 2000.

Norman, D.A. "Emotions and Design: Attractive Things Work Better," *Interactions Magazine*, (IX:4), 2002, pp. 36-42.

Pace, B.J. "Color Combinations and Contrast Reversals on Visual Display Units," Proceedings of the Human Factors Society, Santa Monica, CA, 1984.

Radl, G.W. "Experimental Investigations for Optimal Presentation-Mode and Colours of Symbols on the CRT-Screen," in *Ergonomic Aspects of Visual Display Terminals*, E. Grandjean and E. Vigliani (eds.), London: Taylor & Francis, 1980, pp. 127-136.