BEHAVIOUR & INFORMATION TECHNOLOGY, MAY–JUNE 2004, VOL. 23, NO. 3, 183–195

Taylor & Francis
Taylor & Francis Group

The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention

RICHARD H. HALL† and PATRICK HANNA‡

†University of Missouri – Rolla, Missouri, USA; e-mail: rhall@umr.edu

‡Matrikon Corporation, USA

Abstract. The purpose of this experiment was to examine the effect of web page text/background colour combination on readability, retention, aesthetics, and behavioural intention. One hundred and thirty-six participants studied two Web pages, one with educational content and one with commercial content, in one of four colour-combination conditions. Major findings were: (a) Colours with greater contrast ratio generally lead to greater readability; (b) colour combination did not significantly affect retention; (c) preferred colours (i.e., blues and chromatic colours) led to higher ratings of aesthetic quality and intention to purchase; and (d) ratings of aesthetic quality were significantly related to intention to purchase.

1. Introduction

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 colours, 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 colour combinations still proliferate.

Web design guidelines often include recommendations for appropriate colour 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 (Nielsen 2000):

> Use colours 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 colour scheme throws people off a little and slows their reading slightly. Legibility suffers much more for colour schemes that make the text any lighter than pure black, especially if the background is made any darker than pure white.

Unfortunately, Nielson does not offer any references for this statement. In fact, an examination of the research that exists on this topic indicates that the relationship between text-background colour combinations and readability is not as clear as it might seem, though it is generally true that a strong contrast leads to more readable text. In addition, colours are used on web pages for purposes other than maximizing readability. These colours enhance the aesthetics of the page, which can potentially impact the user. This will also be addressed below. We will begin with a discussion of the effect of page colour on readability.

1.1. *Readability*

A great deal of 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 colour combinations (Bouma 1980, Mills and Weldon 1987). However, this research

Behaviour & Information Technology
ISSN 0144-929X print/ISSN 1362-3001 online © 2004 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/01449290410001669932

provides a useful background, and results are largely consistent with more recent studies.

Much of the early work on text-background combinations failed to identify specific colour combinations that were the most readable (Radl 1980). For example, one study failed to find any significant difference among 24 different colour combinations on performance with a text search task (Pace 1984). On the other hand, regardless of the specific colour combination, higher levels of contrast generally lead to greater readability (Radl 1980, Bruce and Foster 1982).

More recent research supports the contention that contrast is an important predictor of readability. For example, Shieh and Lin (2000) compared the impact of 12 different colour combinations on participants' ability to perform a basic visual identification task. In addition to colour combination they considered screen type (LCD vs. CRT) and ambient illumination. First of all, colour combination had a greater impact on performance than the other factors, indicating the importance of colour combinations. Blue and yellow combinations lead to the best performance and purple and red the worst. Consistent with previous research, blue and yellow also had the greatest luminance contrast and red and purple the least. In general, the trend across all colour combinations was the higher the luminance contrast, the better the performance. The Shieh and Lin (2000) study also included a measure of subjective preference, and the results with respect to colour combinations, paralleled the readability results to a surprising degree. This is discussed in more detail in the preference and aesthetics section below.

Recent research also indicates that inconsistency in studies of readability as a function of font/background colour combinations may be due to the confounding of contrast of hue with luminance contrast. Hue is the dimension that we normally think of as colour, which is defined by wavelength, while luminance is the 'brightness' of a colour as defined by wave height. Colours not only differ from one another in hue, but they also differ to some degree in luminance. Lin (2003) conducted a series of three experiments where chromatic (i.e., not black/white) colours were placed on a grey and luminance of colours was systematically varied. Readability performance in most cases could be accounted for by luminance contrast, not hue (colour). The one exception was at very low levels of luminance contrast. In this case, purple and cyan resulted in better performance than yellow, despite equivalent luminance contrast of the colour and the background.

There are few empirical studies on readability and text/colour combinations specifically aimed at web pages (Hill and Scharff 1997). Studies specifically aimed at the web are important since the web has come to play such an important role in information distribution and communication. Displays on the web are unique in that a designer cannot be very certain about the browser, system, resolution, or other factors that may affect a given display. Further, many different types of multimedia devices can and are used via the web, which allows for factors such as text dynamics to play a role in impacting colour perception. One interesting study, which relates to the latter, was recently conducted by Wang and colleagues (Wang *et al*. 2003), where scrolling text was examined. They varied a number of factors associated with the scrolling text. Among these factors was text-background colour combination. They found that combinations with positive polarity resulted in better performance (that is dark text on light background), and, as with studies mentioned previously, the greater the contrast between colour combinations the better the performance. It should be noted that a similar positive polarity effect on readability performance was found in the Shieh study discussed above (Shieh and Lin 2000).

A series of two experiments conducted by Hill and Scharff (1997, 1999) focused specifically on web pages, consisting of text presented via a web browser. In the most recent study Hill and Scharff (1999) varied the background texture, colour, 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 colours (blue, grey, and yellow). They found a significant main effect for colour with better performance for the grey and yellow backgrounds than with the blue, again consistent with better performance for higher contrast.

In an earlier study (Hill and Scharff 1997) six colour 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 colour was found with the best performance for green text on a yellow background and the worst for red on green. This poor performance for red and green was likely due to more than just lower contrast ratio, in that opponent colours such as these often appear to 'vibrate' when placed side by side (Clarke 2002). Though this finding appears to be consistent with the high contrast effect, it should be noted that black on white was one of the six 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 colour combination is inconsistent with the contrast effect and clearly inconsistent with Nielson's recommendation in the quote above. This inconsistency with the contrast effect may be due to the fact that

luminescence was not controlled, which is representative of the fact that colours on the web cannot be well controlled, since they vary with the users browser and computer system. In addition, the study found that the colour effect was often mediated by other factors, such as font type. 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 colour combination when Arial font was used.

The 1997 study (Hill and Scharff 1997) also included a comparison of grey and white backgrounds, which was motivated by the fact that most web browsers at the time had grey 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 colours (light grey, dark grey, and white) were used. Surprisingly, they found better performance with the grey 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.)

In April of 2000 the World Wide Web consortium (w3c) published a working draft of a document for 'Techniques for Accessibility Evaluation and Repair Tools' (http://www.w3.org/TR/AERT). This included an algorithm for determining the brightness (luminescence) contrast and colour (hue) contrast between two colours based on the standard method of assigning RGB (red, green and blue) values to colours (http://www.w3.org/TR/AERT#colour-contrast). The author of this technical document and colleagues also carried out an initial evaluation study of the algorithms (http://www.aprompt.ca/WebPageColours.html). In this study, 42 different web pages were created that represented different levels of contrast based on a combined score from the two w3c recommended algorithms. These pages included short text passages. In a within subject design, 50 participants were asked to rate each of the pages using a sliding scale that ranged from 'impossible to read' to 'effortless to read'. Although the relationship between contrast and readability ratings was not perfect, and outliers were noted, a strong and significant relationship was found, adding further support to the importance of contrast as effecting readability, and also supporting the validity of the algorithm.

1.2. *Affect, aesthetics, and preference*

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 endeavours is a balance between the functional and aesthetic. Factors such as aesthetically pleasing colour 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 field, 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 behavioural intention, which could presumably lead to behaviours that would be especially important for commercial sites, in particular purchasing.

There is a long history of research on the impact of colours on emotions independent of computer displays. One consistent finding is that people in general tend to find short wavelength colours (blues and greens) as more pleasant than long wavelength colours (reds and yellows). For example, Guilford and Smith (Guilford 1959) asked participants to rate colours based on preference, which resulted in the following rank ordering from most to least preferred: blue, green, purple, violet, red, orange and yellow. A similar result emerged from a very different study (Osgood *et al.* 1957), in which participants across a number of cultures were asked to rate colour words (e.g., 'red', 'green') using a semantic differential scale. In this study participants associated blue and green with 'good'. However, there was some indication that the relationship between wavelength and preference was not the only relevant dimension. Though yellow was associated with 'bad' and 'weak', red was rated as 'strong' and 'active', which can not be conceived as the opposite end of the preference dimension. Thus there appears to be another, somewhat orthogonal dimension to preference, which is arousal. In fact, studies of the autonomic nervous system response to colours have also found that longer wavelength colours elicit higher levels of autonomic arousal than short wavelength colours (Wilson 1966, Jacobs and Hustmyer 1974). This arousal can, however, be negative or positive depending on context. For example, in contrast to the relatively positive 'strong' and 'active' associated with red mentioned above, another study found that long wavelength colours can also elicit higher levels of state anxiety (Jacobs and Suess 1975).

In a more recent examination of colours on emotions, Valdez and Mehrabian (1995) systematically controlled hue, saturation, and brightness and utilized a pleasure-arousal-dominance emotion model for conceptualizing user responses. Users rated colours using a semantic differential scale. In one experiment participants rated

various colours within a given hue and in a second experiment participants rated different hues. Overall, the expected relationship between pleasure and wavelength was found – short wavelength colours were preferred. However, the effects for arousal were not consistent with previous research, in that the most arousing colours included green and even blue (green-yellow, blue-green, and green), which are short wavelength colours. The authors point out that they also found a strong positive relationship between saturation (i.e., a colour's 'vividness') and arousal, while they controlled carefully for saturation in comparing colours (hues). Thus, the highly arousing effect of red found in previous studies may have been the result of the fact that samples of red tend to be highly saturated, so the high levels of arousal attributed to red may have been due to the confounding of hue with saturation in these previous studies.

There has been an increased interest in emotion as it relates to computers in the form of 'affective computing', which is an area that has become popular in the last decade (Picard 1997). Emotional responses have been identified and are related to characteristics of the interface and computer system. For example, Riseberg and colleagues (Riseberg *et al*. 1998) purposely created frustration in users by offering them a cash reward for performance on a video game, and then purposely creating a 'stuck mouse' effect during the game. Physiological measures of autonomic arousal differentiated between frustrated and non-frustrated states in users. Similarly, in a recent study, increased autonomic arousal was found in response to video and audio that was not properly synchronized (Ali and Marsden 2003). However, none of the studies that have emerged within the affective computing research area have examined the impact of colour on perception of computer displays in general, or web pages in particular. This topic is particularly important since colour and aesthetics can be a very important part of web design as mentioned above.

In general there are few studies that have examined the impact of computer display colour combinations on user emotions. One exception was a study conducted by Pastoor (1990), which included two experiments. In experiment 1, participants viewed a set of nouns on coloured backgrounds in 792 different colour combinations. The participants used a six step scale to rate the words. They were instructed (Pastoor 1990) to 'read some of the displayed words and to emphasize the aesthetic appearance of the screen pages in forming their ratings'. In experiment 2 a greatly reduced set of 18 colour combinations was used and the outcome measures included a reading task, and search task, and subjective ratings of aesthetics, power, legibility, and strain. Summarizing the results of both of these studies the author points out that, although there were a number of colour effects, there was no consistent effect for hue on ratings or performance. The only exception was that short wavelength colours are preferred for combinations with negative polarity (light on dark). Thus, the only clear finding was consistent with the research on colours cited above, that blues and greens are preferred, but, in general, there was minimal effect for colour combinations on subjective measures of preference/aesthetics/affect.

Lastly, we mentioned above that we would revisit the Shieh and Lin study (2000) study in which subjective preference was examined. In this study, the measure of preference partially included affect. Participants were asked to rate the different colour combinations on a 10 point scale with 1 representing 'very poor' and 10 representing 'excellent. In their ratings, users were asked to emphasize 'clearness', 'aesthetic appearance', and 'visual comfort' in making an overall preference rating. Thus this question combined subjective rating of readability with affect/aesthetics. In fact, as mentioned above, the preference ratings strongly paralleled the readability performance. Blue and yellow combinations were rated the highest on preference, while purple and red were rated the lowest.

1.3. *Extension of previous research*

The current experiment extends the research discussed above in two basic ways. First this experiment will examine the affective impact of text-colour combinations as they are presented on web pages, and the associated impact on behavioural intention. As mentioned, an emphasis has been placed on the role of affect and aesthetics in web design recently. Market researchers have recognized for some time the importance that aesthetic factors can play on consumer behaviour, and this almost surely should impact on web design. Among the factors that have been emphasized in this context are more aesthetic visual displays, in which colour will certainly come to play an important role (Jennings 2000). The second basic way in which this research will extend the research reviewed is that a measure of retention will be included as an outcome. 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 colour combinations.

2. Research model and hypotheses

Figure 1 is a graphical depiction of the framework that guided the current research, and represents the relationship between font colour, outcomes, and content.

The model is based on the contention that contrast factors will impact readability and retention, and preference will impact aesthetics and intention in a fairly straightforward manner. These are represented in the first four hypotheses presented below. Similarly, we propose that these consequent measures, readability with retention and aesthetics with intention, will be related to one another, though in a more indirect fashion. Finally, it's important to point out that this is a preliminary and exploratory framework for describing these relations and is principally provided here as an organizational guide for a series of hypotheses and analyses that will follow, not as a representation of a structural statistical model to be tested as a whole.

The hypotheses derived from this model and explanations follow. Note that, in the following hypotheses, when we use the term contrast, we are referring to contrast of brightness and hue combined.

Hypothesis 1: Colour combinations with higher levels of contrast will receive higher ratings in readability.

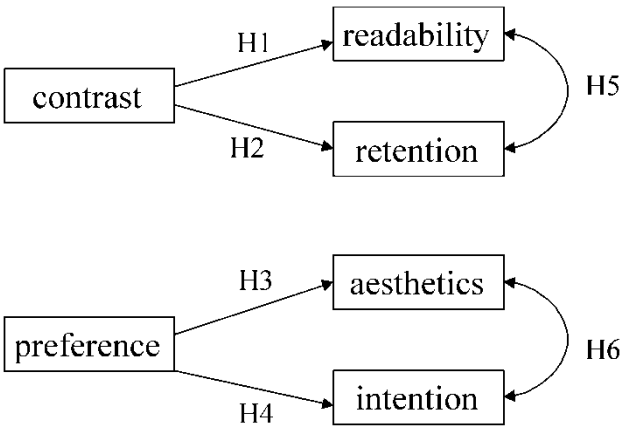

Figure 1. Research model.

The colour combinations used in this study varied along two dimensions: contrast and preference. The latter is discussed below. The four colour combinations (font/background) used were black/white, white/black, light blue/dark blue, and cyan/black. Both black on white and white on black colour combination represented maximal contrast. We also used a combination of light and dark blue, and cyan (blue-green) on black. The former represented a greater degree of both brightness and colour contrast. The contrast ratios for all colours, based on the w3c recommended algorithm discussed above is presented in table 1.

There is a large body of research, reviewed above, that indicates that high levels of contrast leads to better readability. Though most of this was not specifically aimed at web pages, we expect this effect will extend to the web. Specifically, the black/white combinations should result in the highest levels of readability, followed by the dark/light blue combination, followed by the cyan/black combination.

Hypothesis 2: Colour combinations with higher levels of contrast will lead to greater retention than colour combinations with lower levels of contrast.

There is a logical connection between the readability of text materials and the retention of the material, since the latter is not possible without the former. It follows that contrast should also positively impact retention. As with readability, we predict that the black/white combinations should result in the highest levels of retention, followed by the dark/light blue combination, followed by the cyan/black combination.

Hypothesis 3: Preferred colours will lead to higher ratings of aesthetics.

The second dimension that the colours represent is preference. With respect to the colours we selected we conceive the dark and light blue combination as ranking highest on this dimension, since blues are

Table 1. Colour combinations and contrast.

Font/background colour	Contrast	
	Brightness*	Colour**
Black/white	255	765
White/black	255	765
Light blue/dark blue	210	588
Cyan/black	178	510

*Range from 0–255, w3c recommended minimum = 125.
**Range from 0–765, w3c recommended minimum = 500.

consistently preferred across the colour studies reviewed. The cyan and black combination is second on this dimension, since the cyan is a combination of green and blue, which are low-wavelength colours. This is balanced out by the presence of a black background. Although most of the studies reviewed did not examine achromatic colours (black and white), those that did indicate that a chromatic colours are less preferred. For example, in the Osgood cross cultural study on colour names (Osgood *et al*. 1957), black and grey were associated with 'bad', and, though white was associated with 'good' it was also associated with 'weak'. In the Pastoor (1990) study discussed above, in experiment 2 achromatic colour combinations were included. Participants rated the colour combinations that included blue and cyan higher than the achromatic combinations in 15 of 16 combination comparisons on subjective ratings of aesthetics and power; though achromatic colours were preferred in readability and eye strain (achromatic colours were rated as causing less eye strain). We propose that these findings from previous research will extend to the web, such that the preferred colours will be rated as the most aesthetically pleasing. More specifically, we predict that the dark and light blue combination will lead to the highest ratings in aesthetics and behavioural intention, followed by cyan and black, and this will then be followed by the achromatic (black and white) colour combinations.

Hypothesis 4: Preferred colours will lead to higher ratings of behavioural intention.

It is our contention that colours that are preferred will generate positive affect, which will, in turn, lead to a greater intention to purchase a given product. Therefore, we also predict that preference will impact behavioural intention, such that these same preferred colours will have a significant impact on behavioural intention, in the same order as presented above with aesthetic ratings.

Hypothesis 5: Ratings of readability will be significantly related to retention.

Unlike most of the experiments reviewed above, readability in this experiment was rated via participants' subjective ratings. In the studies that used subjective ratings, such as the Ridpath *et al*. study (http://www.aprompt.ca/WebPageColours.html), results were similar to those that used objective measures of readability, such as search tasks, in that contrast was predictive of readability. Retention, on the other hand, was an objective measure in our study consisting of a quiz over participants' retention of information contained on the web pages they viewed and the other three measures were subjective self-report measures, in which they were asked to rate statements, which referred to the pages. We assume that readability will be a basic prerequisite to accurate retention, since information cannot be retained if it is not acquired. As a consequence, a significant relationship between readability and retention is predicted.

Hypothesis 6: Ratings of aesthetics will be significantly related to behavioural intention.

Advertisers in print and television media have long known that the aesthetics of the media can impact buying behaviour (Jennings 2000). Though the web is a different medium, where interactivity plays a much more important role, the impact of aesthetics should still have an important impact on behaviour. E-commerce researchers have suggested that we need to think of users as actors in a play as opposed to observers, as would be the case with traditional media (Laurel 1993). Jennings (2000) argues that principles of aesthetics in design focus principally on visual perception, and that 'pleasing visuals are important because they create first impressions which result in a desire to explore further'. He also notes (Jennings 2000) that many web sites do not take this into account and for such sites 'visual improvements should be made before considering more subtle issues'. Therefore, a significant relationship between ratings of aesthetics and behavioural intention is predicted.

2.1. *Content*

As noted in the model above, we used two different types of content: educational and commercial. We do not propose any specific hypotheses associated with the different content, since we anticipated that the same relationships among colour combinations and outcome measures will be found across content areas. We used these two different content areas for a number of reasons. First, we wanted to examine the generalizability of the results. Second, many web design texts make a distinction among basic types of web sites, and these two types of sites represent two of the basic categories (Lazar 2001, Farkas and Farkas 2002). Third, we propose that the focus of these two types of sites represent well the different types of outcomes proposed in our model. With education the focus is more on retention, while, with commercial sites, the focus is more on behavioural intention. Of course, aesthetic factors are important in education and retention plays an important role in

commerce. However, the primary goal of education oriented sites is to provide the user with information and this often involves encouraging the user to retain the information after they leave the sites. On the other hand, the bottom line for most commercial sites is to increase sales by directly or indirectly encouraging the user to purchase something, and this is often done by focusing on the users' affective states, encouraging them to become excited about a product or service.

3. Research methodology

3.1. *Participants*

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

3.2. *Materials*

3.2.1. *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 colour 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 cyan text on black background (CB). The hexagonal codes for these colours were: black (000000); white (FFFFFF); light blue (DED9FB); dark blue (000066); cyan (00FFFF). The materials used in this experiment can be viewed on the web at http://campus.umr.edu/lite/font_color.

3.2.2. *Outcome measures*: A 10 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 labelled 'strongly disagree' and 10 labelled 'strongly agree'. Both surveys included the following five items:

(1) The colour combination made the text easy to read;
(2) The colour combination made the text easy to study;
(3) I found the colour combination pleasing to look at;
(4) I found the colour combination stimulating to the eye;
(5) I found the colour 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 colour combination made me want to buy this product.

This questionnaire was designed for this experiment. We did not use the same preference measures as the experiments reviewed in the introduction because in some cases they confounded readability and aesthetics, and/or they asked a single question (Shieh and Lin 2000), which would negatively impact reliability. Further, we developed questions based on the model we posed. Within our questionnaire, items 1 and 2 were intended as measures of readability; items 3 – 5 were intended to measure aesthetics; and items 6 and 7 were measures of behavioural intention. We conducted a factor analysis to assure the proper classification of the measures, as well as coefficient alpha analyses in order to assure adequate reliability (see Results section).

3.3. *Procedure*

This experiment took place in 10 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-colour conditions: BW, WB, B, or CB (see section on web pages above for description of colours). 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 10 min, after which they were required to go to the quiz/questionnaire page for 10 min, 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 2.

4. Results

4.1. *Classification of measures*

Two factor analyses were conducted, one for the neuron outcomes and one for the Hallaview outcomes. In both cases a principal components with a Varimax rotation was used. In the first analysis a two-factor solution was forced to represent readability and aesthetics (there were no behavioural intention items in the first post-questionnaire). The items loaded consistent with expectations, with the exception of the professional looking item which loaded on the readability factor. These loadings are displayed in table 3. The rotated solution accounted for 86% of the variance and the aesthetics and readability factors accounted for 45% (Eigenvalue = 2.25) and 41% (Eiganvalue = 2.05) of the variance respectively.

In the second, Hallaview, analysis a three-factor solution was selected to represent readability, aesthetics, and behavioural intention. Again, the items loaded logically as anticipated with the exception that the 'professional looking' item again loaded on the readability factor. The items and loadings are displayed in table 4. The rotated solution accounted for 78% of the variance and the aesthetics, readability, and behavioural intention factors accounted for 30% (Eigenvalue = 2.07), 28% (Eigenvalue = 1.97), and 21% (Eigenvalue = 1.45) of the variance accordingly.

Five factor scores were created for further analyses, consisting of aesthetics and readability scales for both the neuron and Hallaview questionnaires, and a behavioural intention scale for the Hallaview questionnaire. These measures were constructed by averaging the items that primarily loaded on a given factor (the bold items in tables 1 and 2 for each factor). To assess the reliability of these newly created scales, coefficient alphas were computed at the item level and these were $\alpha = 0.85$, $\alpha = 0.89$, $\alpha = 0.80$, $\alpha = 0.85$, and $\alpha = 0.55$ for the neuron-aesthetics, neuron-readability, Hallaview-aesthetics, Hallaview-readability, and Hallaview-behavioural intention scales respectively. Despite the low alpha level for the behavioural intention scale we made the decision to use the scale in subsequent analysis. The decision was based on the identification of the scale in the factor analysis, and our reluctance to use a single item measure, by dividing the scale. Further, the low alpha score is most likely partly attributable to the small number of items in the scale (2), since alpha value is known to decreases with the number of items (Nunnaly 1978).

Table 2. Experimental session schedule.

Time	Activity
0–:10	Introduction, consent
:10–:20	Study content 1
:20–:30	Quiz and questionnaire 1
:30–:40	Study content 2
:40–:50	Quiz and questionnaire 2

Table 3. Factor loadings for neuron outcomes (rotated solution).

	Factor	
Items	Aesthetics	Readability
Easy to read	(0.52)	0.76
Easy to study	(0.55)	0.72
Pleasing to look at	0.91	(0.27)
Stimulating to the eye	0.93	(0.12)
Professional looking	(0.01)	0.92

Table 4. Factor loadings for Hallaview outcomes (rotated solution).

	Factor		
Items	Aesthetics	Readability	Intention
Easy to read	(0.49)	**0.78**	(− 0.02)
Easy to study	(0.52)	**0.73**	(0.06)
Pleasing to look at	**0.88**	(0.22)	(0.16)
Stimulating to the eye	**0.86**	(0.15)	(0.22)
Professional looking	(− 0.02)	**0.84**	(0.23)
Like to buy	(0.16)	(− 0.03)	**0.85**
Colours made me want to buy	(0.13)	(0.26)	**0.78**

4.2. *Hypotheses 1 and 2: Impact of colour-combinations on readability and retention*

In order to address the first two hypotheses that colours with higher contrast would have a greater impact on readability and retention, a one-way between-subjects multivariate analysis of variance (MANOVA) was computed with experimental group (BW vs WB vs B vs CB) as the independent variable and neuron readability, Hallaview readability, neuron quiz score, and Hallaview quiz score as the dependent variables. The number of participants per group were: 29, 31, 39, and 35 for the BW, WB, B, and CB groups respectively. The MANOVA was significant $\Lambda(12,336) = 0.771$, $p < 0.001$. Due to the significant MANOVA, a series of four univariate ANOVAs were

conducted, one for each of the four dependent variables. The two readability ANOVAs were statistically significant, while the two retention ANOVAs were not. Tukey's *post hoc* tests were then computed for both of the readability ANOVAs. For both ANOVAS the CB group scored significantly lower than all other groups. In addition, for the neuron ANOVA, the BW group was marginally significantly higher ($p = 0.062$) than the WB group. For the Hallaview ANOVA, the BW group was also significantly higher than the B group and marginally higher ($p = 0.062$) than the WB group. No other mean comparisons were significant. The readability and retention descriptive statistics are displayed in table 5.

4.3. *Hypotheses 3 and 4: Impact of colour-combinations on aesthetics and behavioural intention*

In order to address the third and fourth hypotheses, a one-way between-subjects multivariate analysis of variance (MANOVA) was computed with experimental group (BW vs. WB vs. B vs. CB) as the independent variable and neuron aesthetics, Hallaview aesthetics, and Hallaview behavioural intention as the dependent variables. The number of participants per group were: 30, 32, 39, and 35 for the BW, WB, B, and CB groups respectively. The MANOVA was marginally significant $\Lambda(9,316) = 0.889$, $p = 0.08$. Due to the marginally significant MANOVA a series of three univariate ANOVAs were performed on neuron aesthetic ratings, Hallaview aesthetics, and Hallaview behavioural intention. The neuron aesthetics ANOVA was statistically significant but neither of the Hallaview ANOVAs were significant. Tukey's *post hoc* tests were conducted to compare the means for the neuron aesthetics ANOVA and the mean difference between the blue and black/white group means was marginally significant ($p = 0.058$). The descriptive statistics associated with these ANOVAs are presented in table 6.

4.4. *Hypotheses 5 and 6: Readability-retention relationship and aesthetic-intention relationship*

In order to address hypotheses 5 and 6, Pearson correlations between readability and retention were computed for both the neuron and Hallaview sites. The readability and retention (quiz) scores were significantly related for the neuron page but not for the Hallaview page. To address hypothesis 5, a correlation between aesthetics and behavioural intention was computed for the Hallaview page (there was not a behavioural intention factor for the Neuron page). This correlation was statistically significant. The correlations and significance/probability levels for these analyses are displayed in table 7.

5. Discussion

5.1. *Hypotheses 1 and 2: Impact of colour-combinations on readability and retention*

According to hypothesis 1, colours with higher levels of contrast were expected to lead to higher

Table 6. Aesthetics and behavioural intention scores for the neuron and Hallaview page as a function of colour. Mean (standard deviation).

Font/background colour	Neuron	Hallaview	
	Aesthetics	Aesthetics	Behaviour
Black/white	5.53 (2.54)	5.47 (2.23)	4.43 (2.10)
White/black	5.70 (2.58)	6.08 (2.44)	3.98 (2.16)
Light blue/dark blue	6.97 (1.86)	6.60 (2.08)	4.94 (2.30)
Cyan/black	6.06 (2.39)	6.13 (2.17)	4.87 (2.33)
F (degrees of freedom)	2.72 (3,132)*	1.48 (3,132)^{ns}	1.33 (3,132)^{ns}

*$p < 0.05$; ^{ns}not significant.

Table 5. Readability and retention scores for the neuron and Hallaview page as a function of colour. Mean (standard deviation).

Font/background colour	Neuron		Hallaview	
	Readability	Retention	Readability	Retention
Black/white	7.63(2.20)	8.93(1.51)	7.66(2.02)	8.45(1.76)
White/black	6.25(2.19)	8.29(1.44)	6.43(1.93)	8.06(1.61)
Light blue/dark blue	6.47(1.99)	9.00(1.36)	6.25(1.84)	8.08(1.53)
Cyan/black	5.05(1.96)	8.49(1.63)	5.03(1.88)	8.06(1.37)
F (degrees of freedom)	8.52(3,132)**	1.975(3,131)^{ns}	10.36(3,131)**	0.497(3,132)^{ns}

**$p < 0.001$; ^{ns}not significant.

Table 7. Readability/retention and aesthetics/behaviour correlations for neuron and Hallaview.

	Neuron	Hallaview	
Measures	Readability/ retention	Readability/ retention	Aesthetics/ behaviour
r (degrees of freedom)	0.211*	0.134^{ns}	0.340^{***}

*$p < 0.05$ (2-tailed); ***$p < 0.001$; ^{ns}not significant.

readability ratings and retention (quiz) scores. This hypothesis was largely supported with respect to participants' perceived readability. For both types of material, the means were significantly different, and were in the correct order, with the exception that the mean for the light blue on dark blue rating was higher than the white on black rating with the educational page. The traditional black on white page was clearly the most readable based on participant ratings. Tukey's *post hoc* tests indicated that the black on white page was significantly or marginally significantly higher than all other colours. Surprisingly, the white on black and light blue on dark blue pages were largely equivalent on readability ratings, despite the fact that the white on black page represents maximum contrast. Two potential factors could be responsible for this unexpected result. First, users are more familiar with black on white, which may in turn have a positive impact on readability. This would be partially consistent with the Nielsen quote that begins this paper (Nielsen 2000), though white on black was not found to be 'almost as good' as black on white, as stated in the quote. Another factor that may have influenced the high rating of the blue page is that previous research has found a significant relationship between readability and subjective preference (Shieh and Lin 2000), and the blue page was the most preferred page as predicted. Although, it's important to note that we cannot say if the readability lead to the preference or vice versa.

The second hypothesis was not supported. Retention scores did not differ significantly as a function of colour for either type of content. Further, the order of the means was not even as anticipated. Though those in the black on white group scored higher than other groups with the commercial content, a lower contrast colour combination (light blue/dark blue) resulted in a slightly higher score than the black on white with the educational content. It may simply be that colours do not affect retention the way they impact readability. The relationship between these two factors, though significant with one passage, was moderate at best, as indicated by the correlational analysis. It is also possible that the difference in contrast ratio for the different colour combinations was not great enough to have an impact. Note that all of the colour combinations that were used in this experiment were above the minimum based on w3c recommendations (see table 1). There is some evidence that contrast ratio only has an impact on readability performance when the contrast ratio for some colours is below a minimum baseline (Lin 2003). Though this minimum contrast finding refers to readability, and we did find a significant contrast effect on readability in this study, it is possible that this minimum baseline effect is even stronger for higher level processes such as retention.

5.2. *Hypotheses 3 and 4: Impact of colour-combinations on aesthetics and behavioural intention*

The third and fourth hypotheses were partially supported in that, overall, differences among colour groups were marginally significant with respect to measures of aesthetics and behavioural intention. Further, for the education passage the mean aesthetic ratings differed significantly. Moreover the order of the means was consistent with expectations in that the blue group was highest on aesthetics and behavioural intention scores followed by the cyan on black group (table 3). These results also substantially contrast with the readability and retention outcomes, since learners consistently viewed the combinations that included chromatic colours as more pleasing, stimulating, and more likely to lead them to buy the product in the case of the commercial site.

It is somewhat surprising that the white on black colour (negative polarity) was rated higher than the black on white (positive polarity). As noted above, black often has negative associations (Osgood *et al.* 1957) and, when a difference is found, users generally prefer positive polarity (dark on light) (Shieh and Lin 2000, Wang *et al.* 2003). Though this is a difficult finding to explain, one possible explanation is that the novelty of the white/black combinations somehow affects aesthetic ratings in comparison to the traditional black/white. Two disclaimers worth noting about this unexpected effect are that these two colour combinations did not significantly differ, and the white/black combination was rated lowest in the degree to which participants were encouraged to buy the product (behavioural intention) based on colour (perhaps reflecting the negative connotations of the black colour).

5.3. *Hypotheses 5 and 6: Readability-retention relationship and aesthetic-intention relationship*

The fifth hypothesis that readability would be significantly related to retention was supported for the commercial site, but not for the educational site. The correlation was also relatively low (0.21) even for the commercial site. It may simply be that low level processes of readability are not as strongly related to retention as was anticipated. It may also be due to the fact that the measure of readability was a subjective rating, while the measure of retention was objective recall.

The aesthetic factor score proved to be significantly related to behavioural intention, which is consistent with the sixth hypothesis. It appears, then, the degree to which the participants saw the pages as pleasing and stimulating was linked with the degree to which they intended to purchase a given product. This effect is not surprising given the fact that aesthetics had been identified with other media as being an important factor in influencing consumer behaviour. However, this relationship is relatively unexplored with respect to web pages. This supports the view expressed by Jennings (2000) that visual aesthetics are a fundamental component in determining the effectiveness of e-commerce sites.

5.4. *Classification of measures*

When the questionnaire was designed it was anticipated that outcome scores would fall into two factors for the neuron questionnaire (readability and aesthetics) and three factors for the commercial page (readability, aesthetics, and behavioural intention). For the most part measures loaded as anticipated with the exception that the item, which asked participants to rate the degree to which the page was professional looking, loading most strongly on the readability, rather than the aesthetic factor. This finding is interesting, though not too surprising, that readers view the professional nature of a site to be more tied to its function than its appearance.

It's also interesting that the 'easy to study' and 'easy to read' items had relatively large loadings on the aesthetics factor, indicating that aesthetics and readability were not completely independent. In fact, this result is consistent with the Shieh and Lin (2000) study reviewed in the introduction, where preference for colours paralleled users' performance on readability measures in both studies. Thus, while this factor analysis indicates that it is reasonable to conceive aesthetics and readability as different outcomes, they are certainly related.

5.5. *Implications for designers*

As stated in the introduction, one of the primary purposes of this experiment was to provide a systematic and empirical investigation of the impact of colour combinations on outcomes, in order to provide designers with practical evidence-based guidelines. It is important to keep in mind that this is a, controlled, single experiment, conducted with college students, therefore results should be interpreted accordingly. Despite these constraints, we do feel confident that there are a number of guidelines that can be derived from these results that can aid the designer in selecting background/text colour combinations.

- For educational sites, where retention and readability, especially readability, are a major concern; black on white or a closely related combination of text should be used. This advantage appears to be the result of both the contrast ratio of black and white and the convention or familiarity, since white on black text (equivalent contrast, but much less common) was rated much lower on readability. Therefore, if other colour combinations are the convention for a given context, then the convention should weigh as heavily in the decision as contrast.
- A site that is viewed as readable is also viewed as professional, so these same readability guidelines should be applied if 'professional' is an important part of the image to be projected.
- For commercial sites, where aesthetic and purchasing behaviour factors are a major concern, chromatic (coloured) text/background combinations should be used. Chromatic colours are more likely to lead the viewer to see a site as more visually pleasing and stimulating. Most importantly, these colours are more likely to lead a viewer to the intention to purchase products advertised on the site. Combinations involving the colour blue, and including two chromatic colour (e.g., light blue on dark blue) appear to be preferable to a combination with less contrast and including a chromatic colour (e.g., cyan on black) for promoting positive affect and behavioural intention.

5.6. *Limitations*

Though systematic and controlled, it's important to keep in mind that this was an initial exploratory experiment on the impact of web page colour combinations on a number of outcomes. It's important to note

limitations to better provide a context for interpretation. First, we used a relatively small set of colour combinations as a starting point, and purposely selected colours that varied on a number of dimensions in an effort to gather as much initial information as possible. As a consequence, this does not allow for the specific isolation of the impact of individual factors. Second, colour preference can certainly be influenced by experience and culture (Morton 1997), and the sample of participants consisted of college students at a technology oriented school in the USA Midwest, which is a relatively restricted sample. Third, due to time constraints we did not include any pre-tests for determining participants pre-knowledge and skills. We could have used this information to remove variance associated with these individual difference and/or examined the impact of these factors in mediating outcomes. Fourth, we did not include behavioural intention measures for the educational site outcomes, since it did not explicitly involve the possibility of product purchase. However, we could have included intentions in the form of intentions or motivation to use or study the educational information, which would have provided additional information on the relationship between cognitive outcomes and behavioural intention. Fifth, due to the limited and focused nature of an experiment such as this, our test stimuli could only consist of small amounts of material on single web pages, whereas most web users form impression based on experience with sites consisting of a number of linked pages. Despite these limitations there are a number of interesting findings that emerged, raising a number of important issues to be addressed in future research.

5.7. *Future research*

This research could be extended in a number of directions. First, a more controlled systematic study of colour combinations could be conducted. Hues could be selected to better represent wavelengths across the spectrum – in particular including long wavelength colours. Further, these different colour combinations could be presented more systematically using a fully crossed factorial design. Second, a number of alternative outcomes could be explored. Objective measures of readability could be utilized, such as most previous studies and retention measures could be expanded to include even more complex learning measures such as problem solving and structural knowledge. Physiological measures of affect, which are popular within the area of affective computing could be used. Third, a more applied direction could be pursued. More realistic and detailed e-learning or e-commerce prototypes could be created and examined, or existing sites could be evaluated in an applied context. Finally, a more general examination of the impact of colours in other web-based contexts would be interesting, and more complex measures of aesthetic and affective qualities such as flow could be considered.

Acknowledgements

This research was supported in part by the Instructional Software Development Center at the University of Missouri – Rolla.

References

Ali, A. N. and Marsden, P. H. 2003, Affective muli-modal interfaces: The case of mcgurk effect. Proceedings of the Intelligent User Interfaces Conference, pp. 224 – 226.

Bouma, H. 1980, Visual reading processes and the quality of text displays. In E. Grandjean and E. Vigliani (eds) *Ergonomic Aspects of Visual Display Terminals* (London: Taylor & Francis), pp. 101 – 114.

Bruce, M. and Foster, J. J. 1982, The visibility of colored characters on colored backgrounds in viewdata displays. *Visible Language*, **16,** 382 – 390.

Clarke, J. 2002, *Building accessible web sites* (Boston, MA: New Riders).

Farkas, D. K. and Farkas, J. B. 2002, *Principles of web design* (New York: Longman).

Guilford, J. P. 1959, A system of color preferences. *American Journal of Psychology*, **72,** 487 – 502.

Hill, A. L. and Scharff, L. V. 1997, Readability of screen displays with various foreground/background color combinations, font styles, and font types. Proceedings of the Eleventh National Conference on Undergraduate Research, pp. 742 – 746.

Hill, A. L. and Scharff, L. V. 1999, Legibility of computer displays as a function of colour, saturation, and texure backgrounds. In D. Harris (ed) *Engineering Psychology and Cognitive Ergonomics* (Sydney: Ashgate), pp. 123 – 130.

Jacobs, K. W. and Hustmyer, F. E. 1974, Effects of four psychological primary colors on gsr, heart rate, and respiration rate. *Perceptual and Motor Skills*, **38,** 763 – 766.

Jacobs, K. W. and Suess, J. F. 1975, Effects of four psychological primary colors on anxiety state. *Perceptual and Motor Skills*, **41,** 207 – 210.

Jennings, M. 2000, Theory and models for creating engaging and immersive e-commerce websites. Proceedings of the ACM Computer Personnel Conference, pp. 77 – 85.

Laurel, B. 1993, *Computers as Theater* (Reading, MA: Addison-Wesley).

Lazar, J. 2001, *User-centered Web Development* (Sudbury, MA: Jones and Bartlett).

Lin, C. 2003, Effects of contrast ratio and text color on visual performance with tft-lcd. *International Journal of Industrial Ergonomics*, **31,** 65 – 72.

Mills, C. B. and Weldon, L. J. 1987, Reading text from computer screens. *ACM Computing Surveys*, **19,** 329 – 358.

Morton, J. 1997, *Guide to color symbolism* (Manoa, HI: Colorcom).

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

Norman, D. A. 2002, Emotions & design: Attractive things work better. *Interactions Magazine*, **ix,** 36–42.

Nunnaly, J. 1978, *Psychometric Theory* (New York: McGraw-Hill).

Osgood, C. E., Suci, G. J. and Tannenbaum, P. H. 1957, *The Measurement of Meaning* (Urbana, IL: University of Illinois Press).

Pace, B. J. 1984, Color combinations and contrast reversals on visual display units. Proceedings of the Human Factors Society 28th Annual Meeting, pp. 326–331.

Pastoor, S. 1990, Legibility and subjective preference for color combinations in text. *Human Factors*, **32,** 157–171.

Picard, R. 1997, *Affective Computing* (Cambridge, MA: M.I.T. Press).

Radl, G. W. 1980, Experimental investigations for optimal presentation-mode and colours of symbols on the crt-screen. In E. Grandjean and E. Vigliani (eds) *Ergonomic Aspects of Visual Display Terminals* (London: Taylor & Francis), pp. 127–136.

Riseberg, J., Klein, J., Fernandez, R. and Picard, R. 1998, Frustrating the user on purpose: Using biosignals in a pilot study to detect the user's emotional state. Proceedings of the ACM Special Interest Group on Computer-Human Interactions, pp. 227–228.

Shieh, K. and Lin, C. 2000, Effects of screen type, ambient illumination, and color combination on vdt visual performance and subjective preference. *International Journal of Industrial Ergonomics*, **26,** 527–536.

Valdez, P. and Mehrabian, A. 1995, Effects of color on emotions. *Journal of Experimental Psychology*, **123,** 394–409.

Wang, A., Fang, J. and Chen, C. 2003, Effects of vdt leading-display design on visual performance of users in handling static and dynamic display information dual-tasks. *International Journal of Industrial Ergonomics*, **32,** 93–104.

Wilson, G. D. 1966, Arousal properties of red versus green. *Perceptual and Motor Skills*, **23,** 942–949.