HUMAN FACTORS, 1990, 32(2), 157–171

Legibility and Subjective Preference for Color Combinations in Text

SIEGMUND PASTOOR,[1] *Heinrich-Hertz-Institut für Nachrichtentechnik, Berlin, West Germany*

This study examined legibility performance and subjective preference for text/background color combinations displayed on a video monitor. Luminance contrast was fixed at two preoptimized levels, either with text brighter than the background (10:1) or vice versa (1:6.5). In Experiment 1, 32 subjects rated about 800 color combinations. No evidence suggested differential effects of luminance polarity or hue, with the only exception that cool background colors (blue and bluish cyan) tended to be preferred for the light-on-dark polarity. Saturation had the most important influence on ratings. Any desaturated color combination appears to be satisfactory for text presentation. In Experiment 2 a reduced set of 18 color combinations was investigated with a new sample of 18 subjects. Reading and search times as well as multidimensional ratings were evaluated. There was no evidence for an influence of luminance polarity or chromaticity on performance. Subjective ratings corresponded well with the results of Experiment 1.

INTRODUCTION

Today's computer display technology offers a wide variety of colors that may be used to improve the presentation of graphic and textual information. The applications of multicolor displays range from professional use by highly trained operators (e.g., in air traffic control and industrial process control) to typically nonprofessional, occasional use by owners of home computers or videotex terminals. The European videotex system, for example, allows for a choice from among 4096 different colors for composing a screen page. With large varieties of available colors, there is a growing need to understand how best to employ them.

Based on the experimental literature, there is no question about the potential advantages afforded by color versus achromatic displays (see, e.g., Silverstein, 1987, for an overview). Color can be an effective means to improve the information acquisition and generation process—that is, to make human-computer communication faster, safer, and easier. Besides this objective advantage, color is commonly expected to provide an additional, subjective benefit by making display work more pleasant.

The objective improvements are attributed to the coding properties of color and to a potential enhancement of display legibility by the chromatic contrast between characters and background.

[1] Requests for reprints should be sent to Siegmund Pastoor, Heinrich-Hertz-Institut für Nachrichtentechnik Berlin GmbH, Einsteinufer 37, D-1000 Berlin 10, West Germany.

Color Coding

Overall, color coding has proven to be the most effective form of coding on search tasks (Christ, 1975) and is generally superior or at least equal to achromatic codes (such as size, brightness, and shape) on identification tasks (Luder and Barber, 1984). Teichner (1979) has stressed the usefulness of color in demarcating special areas on the screen. His conclusion was supported by Wöpking, Pastoor, and Beldie (1985), who investigated different codes to demarcate the user guidance information on videotex pages. Color was ranked superior to achromatic codes such as separating lines and type font.

Saito and Bender (1985) propagated the use of color coding in word-processing tasks. In their text editor, color serves to indicate the source and history (age of revision, refinements) of edited and inserted text. As a very interesting application of quantitative color coding, the background of revised text diminishes in saturation in proportion to the age of the revision.

Unfavorable results of color coding are to be expected, especially whenever color may act as a distractor in a given application (Teichner, 1979). Usually problems will occur when the meaning of the color code is not known to the user or when the colors are not properly correlated with the displayed information.

Display Legibility

Visual perception of a color stimulus is generally based on the information of three opponent visual channels: one channel signals the luminance information and two channels convey the chromaticity information of the stimulus (blueness/yellowness and redness/greenness channels). The combined action of the three visual channels for different types of visual tasks (e.g., color discrimination, identification of alphanumeric symbols) has been a very active area of research in recent years. With respect to the topic of color and legibility, Lippert, Farley, Post, and Snyder (1983) have compared the usefulness of standardized three-dimensional color coordinate systems as a performance metric of color contrast. A modified CIE-based color space yielded accurate predictions of reading performance from color differences between text and background. This conclusion was confirmed by a more recent study by Lippert (1986), who also presented a performance nomogram relating luminance contrast and chromatic contrast to display legibility. The nomogram predicted a positive effect of chromatic contrast for low-luminance contrast displays (character luminance within a range of 2.8 times higher or lower than background luminance). Matthews and Mertins (1988), however, found that differences in color space were a minor factor in predicting visual performance. They concluded that there appears little need to develop a metric to aid display designers for tasks in which reading information from the screen is the main requirement. Hence in these applications the selection and specification of colors should be based primarily on subjective preferences.

Negative side effects of color on the legibility of a display may result from the chromatic aberration effect. However, the research literature on relations between CRT-generated colors and visual accommodation has not provided convincing evidence of strong chromatic aberration effects. With the exception of primary emissions from the relatively saturated blue phosphor, all display colors tend to fall within the depth of focus of the normal eye and only minimally affect visual accommodation (Donohoo and Snyder, 1985; Murch, 1982). Matthews and Mertins (1987) reported that in a visual search and decision-making task the influence of screen color characteristics on visual performance or subjective fatigue was minimal, even after hours of task performance. No support was obtained for the general recommendation to avoid the use of red and blue stimuli either alone or in combination in CRT displays. In a more recent study (Matthews and Mertins, 1988), again, no evidence was obtained that suggested a substantial effect of color on performance or well-being.

Subjective Benefits

In contrast with the coding and legibility themes, there are very few experimental results on the subjective benefits of color displays. From the data reviewed by Christ (1975), a general preference for multicolored over monochromatic presentations may be derived. Subjective ratings indicate that color tends to render a display less monotonous and to reduce eye strain and fatigue. Experienced operators often believe that color improves their task performance, even under conditions in which the opposite is true.

Besides these general assessments, systematic studies on the more specific question of which chromaticities should be chosen in order to generate an aesthetically appealing display are scarce. Existing recommendations concerned with display terminals (Lalomia and Happ, 1987; Long, 1984; van Nes, 1986) or print (Tinker, 1969) are based on luminance contrast demands rather than on chromaticity effects. The recommended colors are those that produce the maximum luminance contrast within the inherent limits of the display medium. Therefore the aim of the present study was to explore the role of chromaticity (defined by the two color characteristics, hue and saturation, at constant levels of luminance) on subjective quality assessments.

The experiments focused on videotex applications. Pages of a uniformly colored text were displayed on a video screen. Evaluations were primarily based on categorical ratings of nonexpert subjects. Performance measures (reading and search times) were obtained only for some selected color combinations because an effect on legibility was not to be expected under conditions of matched luminance and high luminance contrast.

The contents of the screen pages were discrete, unrelated words. The displayed text served primarily to demonstrate the color combinations under test conditions (Experiment 1) and to allow evaluation of legibility performance effects (Experiment 2). Hence the scope of the study did not include determination of the suitability of colors for specific text contents such as advertising, letters, and so on (Sallio and de Legge, 1982).

EXPERIMENT 1: QUALITY RATING METHOD

Screen pages consisting of randomly chosen nouns were presented on a computer-driven, high-resolution color monitor (51 cm diagonal, 0.31 mm pitch, 13 MHz bandwidth with red-green-blue [RGB] input, European Broadcasting Union [EBU] standard-phosphor set). In order to avoid interline flicker, the monitor was driven at a frequency of 50 frames/s and 313 picture lines per frame without interlacing. Convergence errors were less than 1.5 min of arc, with reference to a viewing distance of six screen heights (1.6 m).

The viewing conditions were arranged in conformance with Comité Consultatif Internationale des Radiocommunications Recommendation 500 (CCIR, 1974). The background behind the monitor was white (Y = 9 cd/m^2, u' = 0.21, v' = 0.50 according to the 1976 CIE uniform chromaticity scale). The ambient room illumination was low (about 50 lux at observers' seats). Either text was brighter than the background or vice versa, and each luminance contrast polarity condition was tested in a separate session. The luminance contrast was fixed at previously optimized levels obtained in a pretest with luminance being adjusted by the subjects. The following combinations of text versus background luminance were selected: 4 cd/m^2 text versus 26 cd/m^2 background and 40 cd/m^2 text versus 4 cd/m^2 background for the dark-on-light and light-on-dark luminance polarities, respectively.

The character set was based on a matrix size of 9 × 13 pixels for capital letters, subtending approximately 25 min of visual arc. Both this character size and the selected luminances are clearly above the limits for reliable color discrimination and comfortable legibility (Post, 1985; Silverstein, 1982).

The following colors were selected for the quality rating experiment (color name abbreviations in parentheses): green (GR), yellow (YE), orange (OR), red (RE), reddish magenta (RM), magenta (MA), bluish magenta (BM), blue (BL), bluish cyan (BC), and cyan (CY). Figure 1 shows a 1976 CIE uniform color space (UCS) diagram with the chromaticity coordinates of the selected colors. Because of inherent limits on the CRT's color gamut, it was not possible to replicate all 22 chromaticities shown in Figure 1 at all three luminance levels. Consequently, of the (22 × 22) 484 combinations of foreground and background colors that would otherwise have been possible within each luminance/polarity condition, 440 were used for the condition in

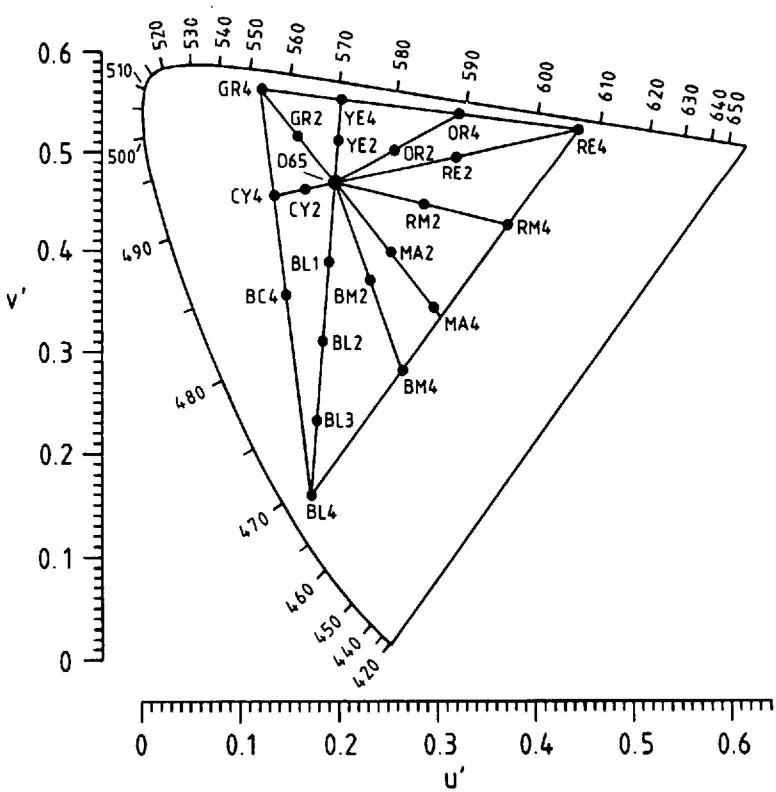

Figure 1. *Selected colors for Experiment 1. u' and v' are the chromaticity coordinates of the 1976 CIE uniform color space. The horseshoe-shaped curve represents the locus of spectrally pure colors (wavelength in nm). This curve encircles the gamut of visible colors. The dominant wavelength of a color sample is given by the intersection of a straight line through the color sample and the achromatic reference (D65) and the spectral colors curve. Excitation purity increases along this line from 0 at the achromatic reference to 1 at the boundary of the color gamut. BC = bluish cyan; BL = blue; BM = bluish magenta; CY = cyan; GR = green; MA = magenta; OR = orange; RE = red; RM = reddish magenta; YE = yellow.*

which the text was darker than the background and 352 were used for the opposite polarity. Numbers 1 to 4 assigned to the color names in Figure 1 label the saturation level: 1 means ¼ of the CRT's maximum color saturation, 2 means 2/4, and so on. Thus a Level 2 saturation (i.e., 50%) represents 50% of the maximum saturation producible by the EBU standard phosphor set along any particular colorimetric axes, but it does not represent an excitation purity or true saturation value of 50%. D65 denotes the achromatic CIE standard illuminant.

Rating Task

Subjects were asked to rate each presented color combination on a six-step scale that was adapted from the CCIR quality scale (CCIR, 1974). The rating categories were *excellent, good, fairly good, rather poor, poor,* and *very poor.* The German translations of these items represent a sufficient approximation to an equal interval scale (Schertz, 1987). Subjects were instructed to read some of the displayed words and to emphasize the aesthetic appearance of the screen pages in forming their ratings.

Subjects

A sample of 32 subjects (15 male, 17 female) participated in the first experiment. They ranged in age from 20 to 56 years with a mean of 30 years and had little or no experience with display workstations. All subjects were screened for normal visual acuity, either with or without corrective lenses, and normal color vision as indicated by the Velhagen color plates (Velhagen, 1974).

Subjects registered their favored colors by rank-ordering eight colored fields (red, green, blue, gray, magenta, cyan, yellow, and orange), which were presented together on one screen page with a light gray background.

Male/female differences in the rankings of the eight colors were evaluated by a series of eight Mann-Whitney U tests, one for each of the ranked colors. None of the comparisons achieved statistical significance ($ps > 0.05$).

The overall ranks of the test colors were determined by calculating their median rank values. The subjects favored (1) blue, cyan, and red (equally ranked), followed by (2) yellow, (3) magenta and green, (4) orange, and (5) gray.

Procedure

The experiment was subdivided into separate sessions for the two luminance polarity conditions (balanced between subjects). During each session two subjects were tested. The color combinations were presented in random order with unlimited viewing time. After both subjects had entered their ratings on small hand-held keypads, the screen was gray for about 5 s in order to avoid successive contrast effects. The sessions typically lasted for 3 h, including regular intermissions to reduce fatigue. The trials were spread over at least two days.

Statistical Evaluation

The quality categories were mapped into the numbers 6 (for *excellent*) to 1 (for *very poor*). First, a three-way (Subject × Text Color × Background Color), fixed-effects, within-subjects ANOVA was used to examine the results from each luminance polarity. Second, a combined data set containing the color combinations common to the two polarities was created and used to perform a four-way (Subject × Text Color × Background Color × Polarity) ANOVA. Third, reduced data sets containing only those text and background colors common to two of the saturation levels—Levels 2 (desaturated colors) and 4 (saturated colors)—were created for both luminance polarities. The ratings for these data sets were examined by four-way (Subject × Text Color × Background Color × Saturation) ANOVAs. In these ANOVAs the saturation factor covered four levels: (1) desaturated colors for both text and background, (2) saturated colors for both text and background, (3) saturated text on desaturated background, and (4) desaturated text on saturated background.

Because two of the independent variables (text color and background color) had a large number of levels, the ANOVAs were supplemented by the conservative Geisser-Greenhouse F test (see Keppel, 1973). In the following section the conservative test will be addressed only in cases of divergent results.

RESULTS OF EXPERIMENT 1

The three-way ANOVAs revealed significant main effects of text color and background color as well as significant interactions (i.e., color combinations) for both luminance polarities.

Light-on-Dark Polarity

The respective statistics are $F(15,465) = 43.62$, $p < 0.0001$ for text color, $F(21,651) = 17.56$, $p < 0.0001$ for background color, and $F(315,9765) = 1.55$, $p < 0.0001$ for Text Color × Background Color interaction. The Geisser-Greenhouse test indicates that this interaction is not significant.

Mean values of the ratings are not presented here in detail but are available from the author. The following paragraphs provide a survey of overall trends along with some selected results.

The following rank order of text colors results when the text colors are arranged according to their mean ratings (i.e., ratings of color combinations averaged across subjects and background colors): D65 (mean = 3.9), YE2, CY2, OR2, BL1, GR2, CY4, YE4, RM2, BM2, MA2, RE2, OR4, BC4, GR4, MA4 (mean = 2.1). The corresponding rank order for background colors is BL1 (mean = 3.9), BC4, BL2, D65, CY2, BL3, CY4, GR2, YE2, RM2, BM2, OR2, MA2, RE2, BL4, YE4, BM4, GR4, RM4, MA4, OR4, RE4 (mean = 2.3).

The mean ratings of individual color combinations range from 4.8 (for achromatic text on a desaturated blue [BL1] background) to 1.7 (for saturated magenta text on a saturated red background). Because of the large number of experimental conditions, post hoc paired comparisons among color combination means yielded overly large critical differences. For example, the critical difference as indicated by the Scheffé test is 4.24 scale units.

Hence for identifying differences that are definitely *not* statistically reliable, Fisher's LSD was used (Keppel, 1973). This method yields a least-significant difference of 0.44. By this criterion the following text/background color combinations do *not* statistically differ from the best rating: D65/BL1, YE2/BL1, CY2/BL1, OR2/BL1, D65/BC4, CY2/BC4, D65/BL2, and YE2/BL2.

Dark-on-Light Polarity

The statistics are $F(21,651) = 6.52$, $p < 0.0001$ for text color, $F(19,589) = 49.98$, $p < 0.0001$ for background color, and $F(399,12369) = 1.75$, $p < 0.0001$ for Text Color × Background Color interaction. The interaction is not significant in the Geisser-Greenhouse test.

The rank order of text colors is BC4 (mean = 3.4), D65, BL1, CY4, BM2, BL2, CY2, MA2, YE2, RM2, GR2, OR2, GR4, YE4, BL3, BM4, RE2, OR4, MA4, RM4, BL4, RE4 (mean = 3.0). The background colors rank as follows: D65 (mean = 4.5), CY2, YE2, BL1, GR2, OR2, CY4, MA2, RE2, BC4, BM2, RM2, BL2, YE4, OR4, BM4, GR4, MA4, RM4, RE4 (mean = 1.7).

Obviously the variation of means for text color is very small. The estimates of variance components support the conclusion that this factor is not important for dark-on-light polarity displays (0.01 for text color versus 0.67 for background color).

The mean ratings of individual color combinations range from 4.9 (for bluish cyan text [BC4] on an achromatic background) to 1.6 (for saturated red text on a saturated red background).

The following text/background color combinations are within Fisher's LSD range (0.39 scale units) for the best-rated conditions: BC4/D65, CY4/D65, BM2/D65, BL2/D65, MA2/D65, RM2/D65, BL3/D65, BM4/D65, RE2/D65, MA4/D65, RM4/D65, BC4/CY2, D65/CY2, BL1/CY2, BL2/CY2, CY2/CY2, MA2/CY2, OR2/CY2, RM4/CY2, BC4/YE2, and BL1/YE2.

The four-way ANOVA on color combinations common to the two polarities revealed that polarity has no effect on the subjective color ratings, $F(1,31) = 2.47$, $p > 0.05$. The effect of text and background color, as well as all interactions, was significant in the combined data set.

A plot of the Polarity × Background Color interaction (not presented here) showed that this interaction results from the fact that combinations with dark background colors are rated more moderately than are combinations with light background colors, $F(19,589) = 21.18$, $p < 0.0001$. The same result holds for dark versus light text colors: Polarity × Text Color interaction, $F(15,465) = 36.01$, $p < 0.0001$—that is, with dark text there is only a little variation of the ratings when text color varies, compared with conditions with light text.

The third evaluation step revealed a predominant influence of saturation on the subjective ratings at both polarities. The respective variance components are 0.30 (saturation), 0.13 (text color), and 0.06 (background color) for the light-on-dark polarity and 0.45 (saturation), 0.01 (text color), and 0.28 (background color) for the dark-on-light polarity.

For the light-on-dark polarity the effect of saturation turns out with $F(3,93) = 70.84$, $p < 0.0001$. The rank order of preferred text/background saturation pairings is as follows (mean values in parentheses): first, desaturated text on desaturated background (3.7); second, desaturated text on saturated background (3.2); third, saturated text on desaturated background (2.8); and last, both text and background saturated (2.4). All differences between means are above the critical difference for post hoc paired comparisons (0.26 Scheffé test).

The respective F value for the dark-on-light polarity is $F(3,93) = 128.66$, $p < 0.0001$. The mean ratings of the four pairings of text and background saturation are as follows: 3.8 for desaturated text on desaturated background, 3.6 for saturated text on desaturated background, 2.6 for desaturated text on saturated background, and 2.5 for saturated text on saturated background. The critical difference is 0.24. This result suggests that for dark text, text color saturation has no significant effect on the ratings either in combination with saturated or with desaturated backgrounds. Combinations involving desaturated background colors yield significantly higher ratings than do combinations with saturated backgrounds.

The ANOVAs also established significant main effects of text color and background color as well as significant Saturation × Text Color and Saturation × Background Color interactions. The statistical values are not reported here. An inspection of the data revealed that these interactions are an experimental artifact because they result from the fact that the differences in color excitation purity between saturation Levels 2 and 4 are not equal for the different hues selected for this experiment (see Figure 1).

DISCUSSION OF EXPERIMENT 1

The results of the first experiment may be interpreted as follows. First, on the condition that luminance contrast is well adapted to a given viewing situation (see method section), luminance polarity does not affect the subjective quality of a colored text presentation. Second, comparing the magnitude of the various factors' effects on the ratings, saturation is the most important factor, with one exception: for the dark-on-light polarity the effect of text color saturation as well as the effect of text color by itself is only marginal. This result may be explained by the findings of Connors (1964), which show that hue discrimination is depressed by background luminances that exceed stimulus luminance by a factor of three or more.

Hence for dark-on-light polarity displays with high luminance contrast, any text color will do if the background is chosen properly. Saturated backgrounds yield unsatisfactory ratings, whereas the less saturated backgrounds (i.e., those nearest D65 in Figure 1) generally receive high ratings with any text color.

For the light-on-dark polarity, combinations involving saturated colors tend to be unsatisfactory and less saturated combinations are generally preferred. Figure 2 gives a visual impression of this trend. In order to illustrate the rating tendency three categories were defined, with boundaries fixed at mean rating scores of 4.5 and 3.5. The boundaries shown in Figure 2 were obtained by interpolating and graphically smoothing the rating data of adjacent test chromaticities. Text color saturation is increasing from white to saturated yellow. The diagrams (a) to (c) demonstrate that the range of background colors yielding high ratings shrinks as text color saturation increases.

The general tendency for downgrading quality as saturation increases is consistent with the findings of Fukuzumi, Yamamoto, and Hayashi (1987). These authors have suggested a psycho-optical explanation of this effect. They argue that the visual accommodation process is affected by the spectral energy distribution of a colored target: if the spectral energy distribution is wide, easy accommodation and a stable visual percept are to be expected. However, a small spectral density distribution tends to destabilize the percept. Because the spectral energy distribution determines saturation (desaturation is related to a broad spectral distribution), the stability of the visual percept is linked to saturation in such a way that desaturation results in a stabilized percept. From this explanation it would follow that the black-and-white condition facilitates the most stable percept and consequently is the optimum condition for reading text.

Overall preferences for specific colors are not firmly established by the ratings. However, an inspection of the rank orders of text and background colors indicates a tendency to prefer achromatic and cool background colors (i.e., short-wavelength hues such as blue, bluish cyan, and cyan) in the light-on-dark condition (see also Figure 2).

This result is partly in accord with the subjects' favorite colors (see subjects section): blue and cyan rank in the highest category. However, it should be noted that red also ranks in the highest category, whereas the achromatic color gray ranks at the bottom. It seems that color preference is not an altogether stable concept but reasonably sensitive to the type of application purpose.

It should be emphasized that the findings of this experiment probably apply only in cases in which legibility is satisfactory and luminance contrast is comparable to the experimental settings. The latter point is especially relevant for the dark-on-light polarity because background colors whose luminance exceeds text luminance by a factor of 6.5 (as in this experiment) most certainly affect text color perception. Presumably the ratings

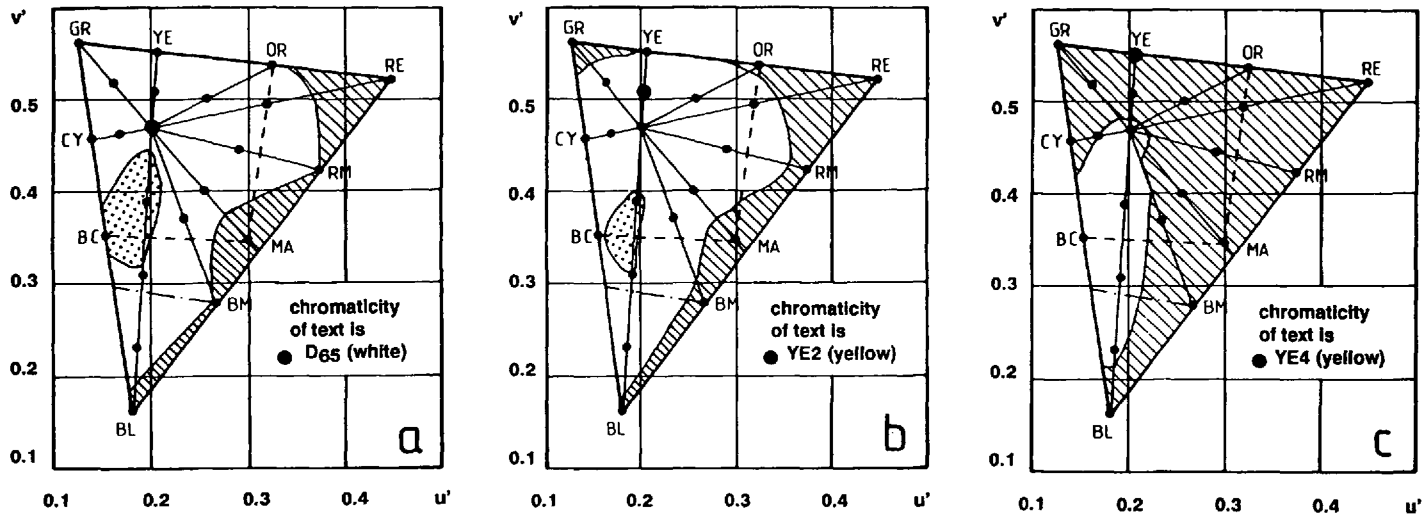

Figure 2. *Rating boundaries for background colors. Dotted areas indicate ratings better than 4.5; striped areas denote ratings worse than 3.5. In this example luminance contrast is 10:1 (text vs. background) and chromaticity of text is (a) white, (b) desaturated yellow, and (c) saturated yellow. BC = bluish cyan; BL = blue; BM = bluish magenta; CY = cyan; GR = green; MA = magenta; OR = orange; RE = red; RM = reddish magenta; YE = yellow.*

would have been different had the luminance contrast been substantially lower (or higher) because the color contrast effects would not have been the same.

EXPERIMENT 2: PERFORMANCE AND MULTIDIMENSIONAL RATING

The main purpose of the follow-up experiment was to test whether or not the spontaneous ratings from Experiment 1 would be replicated under conditions that allowed subjects a more thorough examination of a reduced sample of color combinations. For this reason a reading task and a search task were introduced prior to the assessment of several features (given by semantic differentials) of the color combinations being tested. Legibility-related performance measures were evaluated even though significant legibility effects were not expected, given that luminance contrast was sufficient in all conditions.

METHOD

Nine color combinations for each of the two luminance polarities were selected to represent three levels of coloring a display: (1) no colors (black and white), (2) color exclusively for text or background, and (3) full-color presentation. Because humans cannot simultaneously perceive opponent colors such as red and green or blue and yellow (Pokorny and Smith, 1986), opponent as well as similar colors for text and background were included in order to evaluate the potential influence on performance and subjective preference (see Table 1). Between the two luminance polarities color combinations were matched not for identical chromaticities but, rather, according to their ratings in Experiment 1. High- and low-ranking combinations were included in order to support a potential effect of preferred color combinations on performance and to determine whether high-ranking color combinations provide any subjective benefits compared with black and white. The viewing conditions and basic test apparatus were the same as in Experiment 1.

Tasks and Measures

Reading task. Three screen pages of text (nouns arranged in random order, same page layout as in Experiment 1) had to be read aloud as quickly as possible. A small hand-held keypad was used for forward paging. The total reading time was measured.

Search task. A target word had to be found as quickly as possible out of a column of 10 similar-looking items (they had equal number of characters and same first syllables). Numbers placed in front of the words served

TABLE 1

Text/Background Color Combinations Selected for Experiment 2

Condition	*Light on Dark*	*Dark on Light*
1 Black and white	D65/D65	D65/D65
2 Chromatic/achromatic (+)	CY2/D65	BC4/D65
3 Achromatic/chromatic (+)	D65/BL1	D65/CY2
4 Chromatic/achromatic (−)	MA4/D65	RE4/D65
5 Achromatic/chromatic(−)	D65/RE4	D65/RE4
6 Opponent colors (+)	YE2/BL1	BL2/YE2
7 Similar colors (+)	CY2/BL1	BL2/CY2
8 Opponent colors (−)	MA4/GR4	BL4/RE4
9 Similar colors (−)	MA4/RE4	MA4/MA4

(+) = high-ranking color combination according to Experiment 1; (−) = low-ranking color combination according to Experiment 1.

for identification. Subjects were asked to enter the number of the searched item. The task was repeated 10 times for each condition, with the positions of the targets varying at random. Mean search times were used in the data analyses.

Rating task. After completion of the performance tasks, 18 rating scales were sequentially inserted into the bottom line of the screen pages. The scales were made up of pairs of opposites—for example, pretty/ugly, calm/vivid, easily legible/poorly legible. In a pretest with a pool of 43 scales, the selected scales had shown the highest factor loadings. Ratings were entered using a keypad providing six steps to differentiate the judgments.

Subjects

All 18 subjects (11 females and 7 males ranging in age from 20 to 39 years) had normal color vision and acuity. None of the subjects had participated in Experiment 1.

Procedure

For pretraining of the experimental tasks, color combinations that had received near-indifferent ratings (3.5) in Experiment 1 were used. The experimental conditions were grouped for the two luminance polarities and performance tasks. All conditions were sequentially balanced between subjects (Latin-square design). The rating task was completed after each performance task. Because the experiment typically lasted about five hours, it was split into two sessions on different days.

Statistical Evaluation

Performance data. The distributions of both reading and search times showed a slight positive skew, which was reduced by a square-root transformation prior to analysis. Three-way (Subject × Polarity × Color Combination) fixed-effects, within-subjects ANOVAs were used to examine the results for each task.

Rating data. First, a principal components analysis was performed on the multidimensional scales. A four-factorial solution was chosen, which explained 84.6% of the total variance (after Varimax rotation). On the basis of the characteristic scales, the four factors were designated *aesthetics* (pretty, appealing, pleasant, harmonious), *power* (expressive, vivid, loud), *legibility* (legible, recognizable), and *strain* (sharp, radiant, demanding).

Next, four composite scales were defined by assigning each of the rating scales to one of the four factors according to its factor loading. After the polarity of the scales had been unified, all of the ratings belonging to the same factor were combined by averaging, thus yielding composite scales with the same range as the one-dimensional scale used in Experiment 1.

Four-way (Subject × Luminance Polarity × Task × Color Combination) fixed-effects, within-subjects ANOVAs were then performed on each of the four composite scales.

Color combinations common to Experiment 1 and 2 were also evaluated through regression analyses. The analyses were performed on the mean values of the one-dimensional ratings (Experiment 1) and the mean values of each of the four composite scale ratings (Experiment 2).

RESULTS OF EXPERIMENT 2

Performance

According to the ANOVAs, neither polarity nor color combination had any significant effect on either reading or search times (all $ps > 0.05$).

Ratings

The regression analyses performed on the one-dimensional scale of Experiment 1 and each of the four composite scales of Experiment 2 yielded significant coefficients of determination ranging from $r^2(7) = 0.69$ to $r^2(7) = 0.96$, $p < 0.05$.

The ANOVAs established significant main effects of color combination and significant Color Combination × Polarity interactions on each of the four composite scales. The respective F ratios of the main effects are $F(8,136) = 32.55$ (aesthetics scale), $F(8,136) = 72.27$ (power scale), $F(8,136) = 28.58$ (legibility scale), and $F(8,136) = 81.21$ (strain scale). The interactions have F ratios as follows: $F(8,136) = 16.20$ (aesthetics), $F(8,136) = 9.12$ (power), $F(8,136) = 15.99$ (legibility), and $F(8,136) = 18.35$ (strain). All reported F ratios are significant with $p < 0.0001$.

Mean values are given in Table 2. Critical differences for multiple comparisons as determined by the Scheffé test are 1.35 (aesthetics), 0.92 (power), 1.15 (legibility), and 0.97 (strain).

Neither polarity nor task had a significant main effect on the ratings.

DISCUSSION OF EXPERIMENT 2

Performance versus Rating Scores

Performance scores do not coincide with subjective ratings on the suitability of the color combinations used in this experiment. On one hand, subjects' ratings reflect a perceived reduction of legibility for certain color combinations. This result was to be expected considering that color combinations with very different dominant wavelengths were used (e.g., Condition 8 in Table 1), and these combinations may be difficult to accomodate because of chromatic aberration. On the other hand, the two performance tasks gave no hint for any deterioration caused by these color combinations.

There may be two explanations for this result. First, the performance tasks may not have been sensitive enough to test legibility. Second, the subjects may have been able to balance out the reduced legibility of a difficult color combination by enhanced effort. The second reason seems more plausible because (1) the tasks did show differential effects in a similar context (effect of character size and spacing; see Pastoor, Schwarz, and Beldie, 1983) and (2) legibility was supported by the use of relatively large-sized characters and a high luminance contrast. Hence it may be concluded that no color combination should be generally excluded from text presentation, given that character size and luminance contrast are sufficiently high. This con-

TABLE 2

Mean Values of Composite Rating Factors for the Text/Background Color Combinations Shown in Table 1

	Aesthetics		*Power*		*Legibility*		*Strain*	
1*	4.00	4.11	2.42	2.74	5.09	5.17	2.40	2.33
2	3.83	4.91	2.49	3.10	4.92	5.21	2.39	2.27
3	5.04	4.51	3.30	3.22	5.30	5.03	2.49	2.67
4	2.44	4.35	4.72	3.67	3.25	5.20	5.04	3.25
5	3.87	2.44	4.51	4.99	4.25	3.54	4.17	5.10
6	4.54	4.37	3.73	3.56	4.98	4.92	3.03	2.90
7	4.95	4.71	3.51	3.72	5.27	4.83	2.61	2.89
8	2.20	1.92	5.05	5.31	3.03	2.39	5.24	5.56
9	2.30	2.76	4.88	4.91	2.95	3.43	5.15	4.87

Left columns under each factor are for light on dark and right columns are for dark on light luminance polarity. Maximum score = 6.
* See Table 1 for explanation of condition numbers.

clusion is supported by the aforementioned study by Matthews and Mertins (1987) on long-term effects of color combinations in CRT displays.

Comparison of Results from Experiments 1 and 2

The rating data of Experiments 1 and 2 are highly correlated, though subjects and experimental conditions were not the same. This result confirms the reliability of the data obtained in the first experiment. It suggests that rather spontaneous assessments of the suitability of display colors do not change substantially in the course of a more thorough examination.

Multiple Comparisons

In the discussion to follow, the conditions of Experiment 2 will be compared on the basis of the mean ratings listed in Table 2 and the related critical differences. Note that these comparisons are based on a rather small pool of color combinations; therefore conclusions derived from these comparisons should be regarded as preliminary, pending confirmation by more detailed research.

Black-and-white versus high-ranking color-combinations. These levels are represented by Conditions 1 versus 2, 3, 6, and 7. Only the power scale showed any significant differences: the full-color Conditions 6 (only for light on dark) and 7 (for both polarities) were reliably rated as more powerful than black and white. This result provides no evidence that color can significantly improve aesthetics or legibility or reduce strain, compared with black and white. Conversely, the effects on the power scale suggest that desaturated colors (see Conditions 6 and 7) are suitable for rendering a display more expressive without any negative side effects on aesthetics, legibility, or strain.

Chromatic text versus chromatic background. These conditions do not differ for high-ranking combinations (Conditions 2 vs. 3). Low-ranking combinations (Conditions 4 vs. 5) show different effects for each luminance polarity.

For the light-on-dark polarity, a significant effect was found on the aesthetics scale: chromatic text on an achromatic background was rated as less aesthetic than achromatic text on a chromatic background. Regarding legibility, power, and strain, it makes no difference if color is used for text or background.

For the dark-on-light polarity there are significant differences on each of the four scales. Chromatic text on an achromatic background is rated as less powerful and less demanding but more aesthetic and legible than achromatic text on a chromatic background. Hence this result suggests that for low-ranking combinations with dark-on-light polarity and one chromatic component, color should be used for text rather than for background.

Similar versus opponent color combinations. Comparing high- and low-ranking combinations of similar versus opponent colors separately (i.e., Conditions 6 vs. 7 and 8 vs. 9, respectively) revealed no significant differences on any scale. Hence there is no indication of a similar-versus-opponent effect on the subjective rating of color combinations.

CONCLUSIONS

In a number of studies color proved to be a very effective means of coding information on a display—for instance, as an aid to visualizing relationships among items on the screen and to displaying status information. In a different approach, the present paper focuses on applications in which color definitely bears no information per se but is to be used in order to improve the aesthetic appearance of a display. There are two questions of primary concern: first, are there any combinations of text and background colors prominently preferred by the users? Second, are there any overall effects of characteristic features, such as hue, saturation, or luminance polarity, on performance and ratings? The results of the present study are summarized as follows.

Despite highly significant overall effects of display colors on subjective ratings, it is not appropriate to answer the first question with statements like "use blue for text and yellow for background" because no evidence suggested a differential effect of hue on either ratings or performance. The only exception is a slight indication that cool or short-wavelength background colors (i.e., blue and bluish cyan) may be favorable for displays with light text on dark backgrounds.

Saturation exerts the most important influence on ratings. Generally, desaturated color combinations tend to yield the best results. For a dark-on-light display polarity, however, saturation of the text or foreground color affects ratings only marginally. Thus it may be concluded that any dark text color is satisfactory given that the luminance contrast is high (e.g., 1:6.5, as in this study).

No evidence was found to suggest that the use of color improves a display significantly on such psychological dimensions as aesthetics, legibility, and strain compared with black and white displays. Regarding the power dimension, some desaturated color combinations were rated higher (i.e., more expressive and vivid) than black and white, with no negative side effects on the other psychological dimensions. This means that expressive presentations can be realized without saturated colors.

Provided that the luminances of text and background have been adapted to a given viewing situation, no evidence was obtained that suggested an influence of luminance polarity on either subjective ratings or performance.

There is no evidence for a differential effect of chromaticity on performance, assuming that character size and luminance contrast are properly adjusted for adequate legibility.

Although videotex applications were of focal interest in this study, the findings seem to be applicable to more than merely television screens (e.g., viewgraphs) because the experimental tasks have some validity beyond videotex. There is good reason to believe that the findings reported in this paper may apply to any text presentation characterized by such factors as adequate character size and high luminance contrast between text and background.

ACKNOWLEDGMENTS

This study was conducted at the Human Factors Laboratory, Heinrich-Hertz-Institut, West Berlin, in 1983. It was supported by German Ministry of Research and Technology Grant TK 135. The responsibility for the contents is exclusively taken by the author. I would like to thank the anonymous reviewers for their helpful comments.

REFERENCES

Christ, R. E. (1975). Review and analysis of color coding research for visual displays. *Human Factors, 17*, 542–570.

Comité Consultatif International des Radiocommunications (International Radio Consultative Committee). (1974). *Subjective assessment of the quality of television pictures* (Report 405-2). Geneva: International Telecommunication Union.

Connors, M. M. (1964). Effect of surround and stimulus luminance on the discrimination of hue. *Journal of the Optical Society of America, 54*, 693–695.

Donohoo, D. T., and Snyder, H. L. (1985). Accommodation during color contrast. In *Society for Information Display Digest of Technical Papers* (pp. 200–203). New York: Palisades Institute for Research Services.

Fukuzumi, S., Yamamoto, I., and Hayashi, Y. (1987). Study on VDT display color and readability. In G. Salvendy, S. L. Sauter, and J. J. Hurrell (Eds.), *Social, ergonomic and stress aspects of work with computers* (pp. 189–196). Amsterdam: Elsevier.

Keppel, G. (1973). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice-Hall.

Lalomia, M. J., and Happ, A. J. (1987). The effective use of color for text on the IBM 5133 color display. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 1091–1095). Santa Monica, CA: Human Factors Society.

Lippert, T. M. (1986). Color-difference prediction of legibility performance for CRT raster imagery. In *Society for Information Display Digest of Technical Papers* (pp. 86–89). New York: Palisades Institute for Research Services.

Lippert, T. M., Farley, W. W., Post, D. L., and Snyder, H. L. (1983). Color contrast effects on visual performance. In

Society for Information Display Digest of Technical Papers (pp. 170–171). Coral Gables, FL: Lewis Winner.

Long, T. (1984) Human factors principles for the design of computer colour graphics displays. *British Telecom Technology Journal, 2*(3), 5–14.

Luder, C. B., and Barber, P. J. (1984). Redundant color coding on airborne CRT displays. *Human Factors, 26,* 19–32.

Matthews, M. L., and Mertins, K. (1987). The influence of color on visual search and subjective discomfort using CRT displays. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 1271–1275). Santa Monica, CA: Human Factors Society.

Matthews, M. L., and Mertins, K. (1988). Working with color CRTs: Pink eye, yellow fever, and feeling blue. In E. D. Megaw (Ed.), *Proceedings of the Ergonomics Society's 1988 Annual Conference* (pp. 228–233). London: Taylor & Francis.

Murch, G. M. (1982). Visual accomodation and convergence to multichromatic information displays. In *Society for Information Display Digest of Technical Papers* (pp. 192–193). Coral Gables, FL: Lewis Winner.

Pastoor, S., Schwarz, E., and Beldie, I. P. (1983). The relative suitability of four dot-matrix sizes for text presentation on color television screens. *Human Factors, 25,* 265–272.

Pokorny, J., and Smith V. C. (1986). Colorimetry and color discrimination. In K. R. Boff, L. Kaufman, and J. P. Thomas (Eds.), *Handbook of perception and human performance* (vol. 1, pp. 8-1–8-51). New York: Wiley.

Post, D. L. (1985). Effects of color on CRT symbol legibility. In *Society for Information Display Digest of Technical Papers* (pp. 196–199). New York: Palisades Institute for Research Services.

Saito, M., and Bender, W. (1985). A color-based text editor. In *Society for Information Display Digest of Technical Papers* (pp. 24–27). New York: Palisades Institute for Research Services.

Sallio, P., and de Legge, J. (1982). *Quelques aspects opérationnels concernant l'utilisation des contrastes colorés dans le cas d'images purement textuelles sur écran videotex* [Some operational aspects concerning the use of color contrasts for purely textual images on videotex screens] (TMO Report No. 8198). Cesson-Sévigné, France: Centre Commun d'Études de Télévision et Télécommunications.

Schertz, A. (1987). *Grafische Skalierung von Begriffen zur Bewertung der Ton- und Bildqualität* [Graphical scaling of items for the rating of sound and picture quality] (Tech. Report B88/87). Munich, West Germany: Institut für Rundfunktechnik.

Silverstein, L. D. (1982). Human factors for color CRT displays. In *Society for Information Display: Seminar Lecture Notes* (vol. 2, pp. 1–41), Los Angeles: Society for Information Display.

Silverstein, L. D. (1987). Human factors for color display systems: Concepts, methods, and research. In H. J. Durrett (Ed.), *Color and the computer* (pp. 27–61). Orlando, FL: Academic.

Teichner, W. H. (1979). Color and visual information coding. *Proceedings of the Society for Information Display, 20*(1), 3–9.

Tinker, M. A. (1969). *Legibility of print.* Ames: Iowa State University Press.

van Nes, F. L. (1986). Space, colour and typography on visual display terminals. *Behaviour and Information Technology, 5*(2), 99–118.

Velhagen, K. (Ed.). (1974). *Tafeln zur Prüfung des Farbensinnes* [Color-vision test plates]. Stuttgart, West Germany: G. Thieme Verlag.

Wöpking, M., Pastoor, S., and Beldie, I. P. (1985). Design of user guidance in videotex systems. In *Proceedings of the 11th International Symposium on Human Factors in Telecommunications* (pp. 1/8–8/8). Cesson-Sévigné, France: Centre Commun d'Études de Télévision et Télécommunications/Boston: Information Gatekeepers.