This paper summarizes the results of an analysis of literature on color use in screen design. Fundamental problems with the information base are identified as follows: (1) advise becomes obsolete as newer equipment becomes available; (2) the nature of the task used in the research is not sufficiently similar to tasks performed during teaching and learning; (3) generalizations are either apocryphal in origin or based on empirical results from different display media and transferred to video display terminals automatically; and (4) distinctions are not made with respect to intended uses of the screen display. Instructional designers must ascertain the currency of the equipment used before accepting research generalizations as guides; researchers and authors of literature reviews should indicate which generalizations might no longer be valid. A careful task analysis is necessary in order to maintain a high degree of similarity between the research tasks and actual teaching and learning strategies. Differences in display medium and textual presentation should be noted. There is a need for research on color use in the screen design for instruction. Three tables provide a classification of articles according to information type, tasks used in research, and a summary of empirical research studies is provided in three tables. (Contains 65 references.)

(Author/AEF)
The Art and Science of Color in Multimedia Screen Design, Part II: Science, Myth, and Legend

Earl R. Misanchuk
Instructional Design
Extension Division
(MISANCHUK@SASK.USASK.CA)
306-966-5555

and

Richard A. Schwier
Curriculum Studies
College of Education
(SCHIWER@SASK.USASK.CA)
306-966-7641

The University of Saskatchewan
Saskatoon, SK S7N 0W0
Canada

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Earl R. Misanchuk

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The effective use of color in designing computer screen displays is both an art and a science. Considerable empirical research has been done on the use of color on video display terminals (VDTs) and computer screens, and a number of summaries of this research contribute to the "science" of screen design. This article deals primarily with generalizations gleaned from empirical studies, while Part I of this paper (Schwier and Misanchuk, 1995) looks at the "art"—some rules of thumb that have emerged experientially to advise on how to avoid the garish atrocities that sometimes are produced by neophytes in screen design.

This article eschews a considerable body of research on the use of color on VDTs where the particular use seems quite different than would be found in instructional situations (e.g., air traffic monitoring, airline arrival/departure schedules, pilot/driver navigation systems, on-line job aids). This poses some risk, of course, as sometimes instruction and training approximate the situations we have chosen to ignore. Nevertheless, the danger of over-generalization seems greater than that of under-generalization, and we have chosen to present only the most generic conclusions.

Categories of Articles Reviewed

There have been a number of recent major reviews of the literature pertaining to the use of color that can likely be generalized to multimedia screen design. Some deal with color more or less exclusively (e.g., Brockmann, 1991; Christ, 1975; Davidoff, 1987; Holcomb, 1991; Horton, 1991; Murch, 1987; Winn, 1991) while others deal with color on screens in passing, as part of a review of a related topic (e.g., Gillingham, 1988; Hathaway, 1984; Isaacs, 1987; Mills and Weldon, 1987; Sawyer, 1985; Shaw, 1991; Tuillus, 1983). In addition, there have been numerous reviews which are either mostly focused on topics other than the use of color, or less comprehensive in nature (e.g., Chapman, 1993; Milheim and Lavix, 1992; van Nes, 1986). Finally, there are frequently brief reviews of related literature associated with empirical studies in related areas (e.g., Anglin and Towers, 1993; Baek and Layne, 1988; Baker, Belland, and Cannmcr, 1985, 1986; Bruce and Foster, 1982; Clausing and Schmitt, 1989, 1990; D'Angelo, 1991; Hativa and Teper, 1988; Kerr, 1987; Livingston, 1991; McDonald, Molander, and Noel, 1988; Ohlsson, Nils, on, and Rönnberg, 1981; Pace, 1984; Pastoor, 1990; Radl, 1980; Simmers, 1988; Tuillus, 1981; Wright and Lickorish, 1988).

Table 1 classifies articles on the basis of whether they are primarily:

- summaries of empirical research (i.e., they do not include new empirical data).
• empirical studies (i.e., they do include new empirical data) that may include brief but not necessarily comprehensive reviews of related literature, or

• non-empirical in nature (i.e., generalizations gleaned from experiential or theoretical propositions).

Insert Table 1 about here.

Instructional designers naturally turn to recent review articles in order to keep themselves abreast of the most current thinking on how to most effectively use technology. However, the advice they get there may not be the best possible. While we have no desire to impugn the scholarship of the authors of the articles mentioned above, or others, we wish to point out some major problems in most summaries of the literature dealing with aspects of multimedia, using the case of color in screen design as a case in point. Through the normal practice of exemplary scholarship, myths and legends have crept into our knowledge base on screen design. Authors cite previous authors’ works, but in their efforts to be comprehensive sometimes report outdated or only marginally related literature which then tends to become part of mainstream advice and generalization (much as a legend gets handed from one generation to the next). Thus a certain amount of current instructional design practice, as it is applied to screen design, may actually be based on myth.

Careful perusal of the articles listed in Table 1 reveals four fundamental problems with the information base:

• Some of the advice that is promulgated from article to article may be obsolete, in the sense that the generalizations were formulated using equipment that has been superseded technologically:

• The nature of the task used in the research is not sufficiently similar to tasks typically performed during teaching and learning:

• The generalizations being passed on are either apocryphal in origin or else have been based on empirical results from different display media and transferred to VDTs on the assumption that, say, whatever was found to be an effect of color when paper was the medium of display would automatically transfer to VDTs:

• virtually none of the literature makes distinctions with respect to intended uses of the screen display.
Possibly Obsolete Advice

A shortcoming of many of the published summaries of research on the use of color in screen design is related to the rapid advance of technology: Generalizations based on studies employing obsolete equipment tend to be included in summaries of research alongside contemporary ones, although they may no longer be valid as a result of technological advances. For example, we examined the lists of references attached to the articles in Table 1, and counted the number of times certain articles were cited. The two most-often cited articles are in one case more than a decade old (Tullis, 1981) and in the other, two decades old (Christ, 1975). In the latter review, 75% of the articles reviewed were written in or prior to 1971, 50% of them were written in or prior to 1965, and 25% of them were written in or prior to 1960. Thus the widespread citing of Christ’s conclusions promulgates what may be some dubious, aging generalizations. One has to wonder whether the results of investigations conducted on the hardware available in the 1960s or 1970s really has currency in today’s rapidly shifting technological world. The rapid emergence and widespread dissemination of high-resolution, many-bits-deep color monitors throws into question generalizations derived from studies conducted on relatively coarse-grained monitors capable of displaying only six or eight colors. Today, sixteen-bit color is fairly common, and many systems sport twenty-four-bit and thirty-two-bit color. In addition to the greater number of hues these systems make available, they afford much more control over saturation and, therefore, contrast, which has been shown to have considerable importance (Adkins and Pease, 1991; Baker, Belland, and Cambre, 1985, 1986; Faiola and DeBloois. 1988; Mills and Weldon. 1987; Pace. 1984; Radl. 1980; van Nes. 1986). Given that the human eye can distinguish thousands of different colors and that we are capable of providing at least that range on commonly-available VDTs, how useful is empirical evidence about the optimality of a given color of text on a given color of background, unless the colors involved are described in a much more specific fashion (e.g., Munsell color system coordinates, or RGB values) than has been done to date?

As another example, consider the widely-promulgated advice that navigation elements of a screen be consistent in placement and type, a notion that appears to be much more experientially derived than empirically. Consistency may still be good advice, but given that the generalization was derived primarily on the basis of experience with mainframes that were character-display and command-line or text-menu-based, one wonders how rigidly to apply that advice to a graphical-user interface with hypertext capabilities. To take a more extreme example, some of the literature contains advice that is plainly obsolete (e.g., “use character sets with true descenders”).

What this means, then, is that instructional designers must learn to pay close attention to the dates when research was conducted, and attempt to ascertain the currency of the equipment used, before accepting generalizations as guides to their efforts. By the same token, researchers and authors of literature reviews should be sensitive to the issue, and make the reader aware when generalizations might no longer be valid.
The Nature of the Task

Another shortcoming of many of the published summaries of research on the use of color in screen design is that they do not take into account the type of task used in the research being summarized. Hence both instructional and non-instructional uses of color in screen displays are lumped together in recommendations for practice. Results of studies conducted in pursuit of improved air traffic control systems are sometimes mixed in with others to set the stage for an experimental procedure in education, or are offered by reviewers as purportedly relevant to instructional practice. But are they? A careful analysis of task demands seems warranted.

We classified recent empirical studies according to whether the nature of the task employed was similar to instruction. Only about one-third of those employed tasks that were clearly similar to instruction; about one-sixth were classified as "maybe" instructional (meaning that arguments might be made for them, or that it was not possible to tell from the description what the task was). The remaining half of the studies used tasks that were not related to common instructional activities. Clearly, at minimum, great caution must be used when generalizing the results of the third group to instructional situations.

Table 2 classifies recent empirical studies according to task employed.

---

Table 2: Classifying recent empirical studies according to task employed.

---

Research methods for studying color in screen design are subject to the same debate that has flourished since nonsense syllables were first used to investigate processes of learning: How important is it that the tasks employed in research studies approximate real life? Is it better to risk confounding from the content or instruction, or to employ a task that is "content-free"? The current consensus seems to be that a high degree of similarity between a research task and real life is essential. That fact seems to have been glossed over in some of the recent research on color in screen design.

Different Display Media

Another problem centers on studies involving color that were conducted on media other than VDTs. While the results of such studies may, in fact, be valid for VDTs as well, there is reason to be cautious in making the generalization. For example, it has been shown that reader preferences for fonts in printed materials are quite consistent (Misanchuk, 1989a; Tinker, 1963, 1965). However, users prefer quite different fonts on computer screens than they do on paper (Misanchuk, 1989c). There is also some indication that leading (vertical spacing of text) on paper and on a VDT might show similar differences (Misanchuk, 1989b). Might color effects and preferences suffer similar changes in response to changed display media?
Furthermore, CRTs are radiant light sources, operating on the additive color system, while traditional color theories are based primarily on reflected light, using the subtractive system. It is a qualitatively different experience to view text or images on paper and on a VDT, and "artists are discovering that certain aspects of color theories used in traditional art media are not applicable to computer graphics" (Collery, 1985, p. 1).

Closely related to the problem of possibly inappropriate generalizations based on research conducted on different media is another one, involving the intended display medium. Our observation (subjective, to be sure, but powerful nevertheless) has been that what appears useable and attractive on the VDT screen may not be useable or attractive when displayed on a projected liquid-crystal display (LCD), and vice versa. Virtually none of the literature describes uses other than display on the VDT screen, but it would be easy for a novice screen designer to take generalizations derived from VDT screen research (especially when they are cited out of context) and inappropriately generalize them to displays intended for presentation to a group via LCD technology. Hence this is not a problem with the literature, per se, but rather a potential problem in generalizing the literature.

**Intended Use**

In a similar manner, the literature does not distinguish between screens that are intended for different uses. Being unsure about correct terminology (or even whether such exists), we will use terms appropriate to displays on paper to describe the phenomenon: There is a difference between body text (text that presents the message and elaborates it) and titles or headlines intended to serve as organizing elements for the body text. Without getting into a discussion of related concepts such as screen density (the amount of text presented in a single display, and/or its complexity), on which a separate literature exists, we recognize intuitively and experientially that generalizations appropriate for the use of color in headlines and titles might very well differ from those appropriate for use in body text, and the visual presentation accompanying this paper illustrates some such cases. The literature does not make the distinction: it tends to speak in terms of displayed text, without reference to whether there are four words on the screen in large type or forty in much smaller type (although in a few instances it is possible to make inferences). We believe that researchers, and summarizers and interpreters of research, should be more attentive to these differences.

**What Has “Science” Taught Us About Color?**

To determine what we really know—from empirical evidence—about the use of color for screen design of instructional materials, we review briefly below those recent empirical studies from Table 3 that employed only tasks which were clearly or possibly instructional in nature.
Baek and Layne. 1988: Grades 9–12 students (n = 119) were given a CAL lesson in mathematics (calculating average speed) under a 2 x 3 design (color vs. black and white; text, graphics, animation). It is not clear exactly how color was used, but the authors admit that it was likely used in a non-salient way. Color had no significant effect.

Baker, Belland, and Cambre. 1985, 1986: Elementary-school children were shown an Apple II low resolution color graphic on both color and black-and-white monitors. There was evidence that the displays on the black-and-white monitors may have lacked sufficient figure/ground contrast and therefore may have adversely affected picture comprehension. When the graphic was modified to take into account the fact that the original colors would be displayed on a black-and-white screen, recognition improved.

Clausing and Schmitt. 1989, 1990: Clausing and Schmitt’s two studies both varied text/background parameters, employing white text on a black background, black text on a white background, white text on a light blue background, and black text on a light blue background with eighth-grade students on a close reading exercise. The 1990 study also manipulated line length. Neither study yielded significant results.

D’Angelo. 1991: Subjects below and above 40 years of age showed no significant differences in performance attributable to color combinations.

Hativa and Teper. 1988: One hundred and nine ninth-grade students learned geometric concepts when a microcomputer was used as an “electronic chalkboard” in a lecture/discussion/recitation mode under one of three conditions: monochrome, functional color (“used as a cue”), and non-functional color (used indiscriminately). Immediate and delayed (one month) posttests showed significantly better learning under the functional color treatment, with low-ability learners benefiting the most. High-aptitude students showed the most positive attitude toward the treatment, however.

Holcomb. 1991: Ten males and ten females over age 40 were shown screen dumps with different color combinations, each of four different extant software packages (data base, word processor, spreadsheet, integrated package). One color combination was the normal (default) one used by the software. The others were: “gray text on a vivid blue background, green text on a red background, bright white on a vivid blue background, and yellow text on a black background” (p. 4). Subjects were asked to identify the screen color combination they preferred. Significant differences in preference were found in favor of the white on blue combination for three of the four packages.

Kerr. 1987: Kerr compared the use of color as one alternative cueing method to aid users locating specific information within a database (the other methods were headers, icons (graphics), and a combination of headers, color, and graphic cues; a control database had no cues). He found no significant differences in speed, efficiency, or accuracy of locating the information, but nevertheless concluded that color cues seemed less impressive than textual or
graphic cues. The most important factor was the user's ability to represent the structure of the information internally, he concluded.

Simmers, 1988: Partially-sighted junior and senior high-school students read passages on-screen with text/background combinations of white, green, or yellow on black, and black on white, green, or yellow. Dependent variables were oral reading rate, comfort rating, and brightness-contrast adjustment. Significant results were obtained for brightness-contrast only.

Wright and Lickorish, 1988: Wright and Lickorish reported two studies that investigated the use of color as a cue on computer screens [other studies reported in this article dealt with paper-based text or manipulated variables other than those related to screen presentation]. In both studies, the subjects were to locate specific information that they had previously read. In the first, colored text (green, yellow, white, or cyan) was presented (presumably on a black background, although the report does not make that clear). The second study was exactly the same except for the content. No significant differences were detected in reading time in either study. However, differences were detected in both studies when the color-cued versions were compared to a non-cued version: in both cases, the use of color cues "...may have hampered people from attaining as good a knowledge of content location as they would otherwise have done" (p. 18).

Table 3 summarizes the studies described above into four categories, representing those that:

- showed no significant difference;
- indicated user preferences;
- possibly showed a negative effect; and
- may have restricted generalizability because of special characteristics of the subjects or the hardware used.

Holcomb's study used only five color combinations (hence did not cover the complete range of possibilities), tested only for preference (hence efficacy is unknown), and involved subjects with special characteristics (over age 40). Wright and Lickorish used fairly rudimentary equipment by today's standards (a Z-80 machine with an 80-column by 32 row display). Their major finding was no significant difference in reading time, but there was some evidence that gains resulting from practice at the task were smaller for the color-cued versions than for non-color-cued versions, hence they speculated that the color cues may have interfered. Baker, Belland and Cambre use Apple II low resolution graphics, which are also quite rudimentary by today's
standards. Hativa and Teper's use of a computer as an "electronic chalkboard" is not one that immediately springs to mind when one talks about using computers for instruction. Is it safe to generalize their findings to screen design? Simmers' subjects were partially-sighted, hence his results may not generalize to sighted students.

So, what do we know from the "scientific" literature about the use of color in screen design for instruction? Clearly, not much. There is an obvious need for considerable research to be conducted in the area of using color in the design of screen displays for instruction.

Yet there is no shortage of advice on how color should be used in screen design; a sizable literature exists. Recall that Table 1 lists 13 papers as being largely or exclusively summaries of empirical research. Furthermore, it lists 20 papers as being largely or exclusively non-empirical summaries. Much of the latter consists of experiential advice, which was examined more fully in Part I of this article. Experiential knowledge frequently has great value. The pooled wisdom of master practitioners, if it is consistent, informs the novice. What makes us uneasy about depending primarily on experiential knowledge, however, is that we often don't know how (or even by whom) certain generalizations were derived, what kinds of tasks they were derived from, and, particularly, when they were derived (as that often is related to the state of the art of computing equipment).

**Conclusion**

Generally speaking, we place high value on the results of empirical research, even while recognizing that empirical research may not be capable of providing all the answers, and that research results are colored by the way in which the research was conducted. At this stage, it appears that myth and legend may inform instructional screen design practice more than science. Hopefully, this article will serve as a wake-up call to researchers about the need for more investigation into an area in which it may seem, at first blush, we already know a good deal.

One level of research required is merely the replication of the best of earlier studies, on newer technology, and with more attention to reporting specifics of that technology. We need to test conclusions in light of improvements in display technology. For example, do earlier findings of "good" and "bad" (however they are defined) combinations of text on backgrounds hold when de-saturated colors are employed? Both propositions derived from physiological research (Murch, 1984) and experiential advice (e.g., Faiola, 1990; Faiola and DeBloois, 1988, among others) would lead one to conclude that de-saturated color, especially for backgrounds, is preferable to saturated color, but we have been unable to locate any research in which this hypothesis was tested in an instructional situation. (Our own observations have led us to speculate that beige or light gray might form the most pleasing and effective background against which to present text, for example, but we have not yet subjected that speculation to empirical verification.)
On another level, we now have the technological wherewithal to go beyond static displays on VDTs: we need to recognize that emerging multimedia technologies introduce new questions. What is the role played by compressed color video in displays? How can color be used effectively to present animated graphics? How do various compression strategies influence color? Does color really motivate, as is often claimed, or does it interfere and distract, as has also been alleged (Brockmann, 1991)?

The number of possible research questions involving the use of color in screen design for instructional purposes is very large. In investigating any of them researchers would do well to ensure that their tasks are relevant to the population to which they hope to generalize, that the equipment used is fairly contemporary, and that they provide a great deal of technical detail (with respect to that equipment and the way in which it was used, and with respect to the intended purpose of the displayed information) when reporting their results. Prospective interpreters and summarizers of research should also keep these imperatives in mind when teasing out generalizations. Finally, instructional designers seeking to apply generalizations to their work should act as a second level of filtering, by once again checking to see that the imperatives were applied at earlier stages.
References


<table>
<thead>
<tr>
<th>Table 1. Articles Dealing With Color Grouped According to Type of Information Included</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary Article Type</strong></td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
</tbody>
</table>
| Largely or Exclusively Summaries of Empirical Research | Brockmann, 1991  
Chapman, 1993  
Christ, 1975  
Dwyer, 1987  
Gillingham, 1988  
Hathaway, 1984  
Hutton, 1991  
Issacs, 1987  
Murch, 1987  
Rich, 1991  
Sawyer, 1985  
Shaw, 1991  
Tullis, 1991  
Winn, 1991 |
| Largely or Exclusively Empirical Research Studies | Anglin and Towers, 1993  
Baker and Layne, 1990  
Baker, Bellard, and Cambre, 19851  
Baker, Bellard, and Cambre, 19861  
Bruce and Foster, 1982  
Clausing and Schmitt, 1989  
Clausing and Schmitt, 1990  
D'Angelo, 1991  
Halwa and Tullis, 1988  
Holcomb, 1991  
Kerr, 1997  
Livingston, 1991  
McDonald, Molander, and Noel, 1988  
Olsson, Nilsson, and Rönningen, 1991  
Pace, 1984  
Pastoor, 1990  
Radl, 1980  
Simms, 1988  
Tullis, 1991  
Wright and Lickelough, 1998 |
| Largely or Exclusively Non-Empirical Summaries | Adkins and Pease, 1991  
Baecker and Buxton, 1987b  
Baker, 1983  
Brou, Sciascia, Linden, and Letvin, 1986  
Collery, 1985  
Durrett and Trezona, 1982  
England, 1984  
Faul, 1990  
Faul and DeBlois, 1988  
Goffitz, 1989  
Garner, 1991  
Heinzel, 1984  
Mithen and Lavin, 1992  
Murch, 1984  
Olson and Wilson, 1985  
Rambally and Rambally, 1987  
Reilly and Roach, 1986  
Shneiderman, 1992  
Steinberg, 1991  
Thorat and Smith, 1990  
Tufis, 1990  
Tullis, 1992  
van Nes, 1986  
Waller, Lefrere, and Macdonald-Ross, 1992 |

1 These two papers appear to report the results of the same study.
### Table 2. Tasks Employed in Recent Empirical Research

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Author(s)</th>
</tr>
</thead>
</table>

† Some papers report more than one study. In this table, multiple studies by the same author(s) are designated with numerals in square brackets.
‡ These two papers appear to report the results of the same study.
Table 3. Summary of Empirical Research Studies

<table>
<thead>
<tr>
<th>Category</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Significant Difference</td>
<td>Baek and Layne, 1988</td>
</tr>
<tr>
<td></td>
<td>Clausing and Schmitt, 1989</td>
</tr>
<tr>
<td></td>
<td>Clausing and Schmitt, 1990</td>
</tr>
<tr>
<td></td>
<td>D'Angelo, 1991</td>
</tr>
<tr>
<td></td>
<td>Kerr, 1987</td>
</tr>
<tr>
<td>Preferences Identified</td>
<td>Holcomb, 1991</td>
</tr>
<tr>
<td>Possibly Negative Results</td>
<td>Wright and Lickorish, 1988 [1]</td>
</tr>
<tr>
<td></td>
<td>Wright and Lickorish, 1988 [2]</td>
</tr>
<tr>
<td>Possibly Restricted Generalizability</td>
<td>Baker, Belland, and Cambre, 1985†</td>
</tr>
<tr>
<td></td>
<td>Baker, Belland, and Cambre, 1986†</td>
</tr>
<tr>
<td></td>
<td>Haiva and Teper, 1988</td>
</tr>
<tr>
<td></td>
<td>Simmers, 1989</td>
</tr>
</tbody>
</table>

† These two papers appear to report the results of the same study.