Journal of Experimental Psychology: Applied
2009, Vol. 15, No. 1, 25–34

1076-898X/09/$12.00 DOI: 10.1037/a0015085

A Perceptual Process Approach to Selecting Color Scales for Complex Visualizations

Leonard A. Breslow and J. Gregory Trafton
Naval Research Laboratory, Washington, DC

Raj M. Ratwani
Naval Research Laboratory, Washington, DC
and George Mason University

Previous research has shown that multicolored scales are superior to ordered brightness scales for supporting identification tasks on complex visualizations (categorization, absolute numeric value judgments, etc.), whereas ordered brightness scales are superior for relative comparison tasks (greater/less). We examined the processes by which such tasks are performed. By studying eye movements and by comparing performance on scales of different sizes, we argued that (a) people perform identification tasks by conducting a serial visual search of the legend, whose speed is sensitive to the number of scale colors and the discriminability of the colors; and (b) people perform relative comparison tasks using different processes for multicolored versus brightness scales. With multicolored scales, they perform a parallel search of the legend, whose speed is relatively insensitive to the size of the scale, whereas with brightness scales, people usually directly compare the target colors in the visualization, while making little reference to the legend. Performance of comparisons was relatively robust against increases in scale size, whereas performance of identifications deteriorated markedly, especially with brightness scales, once scale sizes reached 10 colors or more.

Keywords: data visualization, graph comprehension, eye tracking, visual search

Supplemental materials: http://dx.doi.org/10.1037/a0015085.supp

The design of a complex visualization must be guided by the consideration of the uses to which it will be put. A visualization that is effective for some tasks may not be effective for others (Wickens, Merwin, & Lin, 1994). Two basic uses of color codes in data visualizations are (a) the identification of a represented object's type or features (e.g., a temperature range, political party affiliation, body tissue type, etc.) and (b) the determination of quantitative relations among objects, including relative comparison (greater/less), ordering, and maxima/minima. In addition, users combine these functions, for instance when discerning trends across categories.

Various features of color can support each of these functions to different extents. For instance, hue has been found to be highly effective for nominal coding (e.g., coding of object categories, absolute quantitative values and qualitative features), more effective than brightness and several noncolor attributes (Christ, 1975). On the other hand, brightness is well-suited to coding relative quantitative relations (Spence, Kutlesa, & Rose, 1999). Controlled comparisons have demonstrated that hue coding is more effective than brightness coding for supporting absolute value judgments, whereas brightness coding is superior to hue coding for supporting relative quantitative judgments (Merwin & Wickens, 1993; Phillips, 1982).

However, many questions remain concerning this Color Code × Task interaction. How exactly do hue and brightness support each of these tasks? How can we make scales that improve people's task performance? How robust is this interaction effect? To begin to answer these questions, we argue, it is important to analyze the processes by which people solve these tasks. Previous process explanations have been limited to the informal observations that hue scales are more discriminable than brightness scales, whereas brightness scales are more amenable to perceptual ordering than hue scales (Phillips, 1982). This hypothesized analysis is summarized in Table 1. One method of exploring the process by which people use visualizations is by examining their eye movements so as to relate eye movement patterns to performance measures such as accuracy and response time. Also, patterns of response times provide evidence of process. Previous research on the Color Code × Task interaction either did not examine response times (Phillips, 1982) or did not examine eye movements (Merwin & Wickens, 1993; Phillips, 1982) and so were unable to shed much light on process.

We propose that one of the critical aspects of the process of using visualization colors is how the legend is used perceptually and cognitively. Other researchers have shown how important the legend is to graph interpretation (Carpenter & Shah, 1998; Peebles & Cheng, 2003; Trafton, Marshall, Mintz, & Trickett, 2002). For example, Carpenter and Shah showed that people needed to look multiple times at the legend to remember the information. Trafton et al. replicated and extended that work and showed that even skilled forecasters frequently referred to the legend a great deal to answer relatively easy questions about complex meteorological visualizations. Clearly, the more time spent looking back and forth

Leonard A. Breslow and J. Gregory Trafton, Naval Research Laboratory, Washington, DC; Raj M. Ratwani, Naval Research Laboratory, Washington, DC, and Department of Psychology, George Mason University.

Correspondence concerning this article should be addressed to Len Breslow, NRL, Code 5515, Washington, DC 20375. E-mail: len.breslow@nrl.navy.mil

Table 1
Hypothetical Explanations of Task × Scale Type Interaction

	Multicolored scale	Brightness scale
Identification task	**Easy:** highly discriminable	**Hard:** less discriminable
Comparison task	**Hard:** no/little perceived ordering	**Easy:** clear perceived ordering

between the legend and the main visualization, the more difficult and time consuming the task will be. The legend may distract the user from the visualization and the information it encodes (Gillan, Wickens, Hollands & Carswell, 1998).

Hypothesized Process Explanations

Identification of color-coded objects in a visualization depends on reference to a legend except when the color-value associations are symbolic in nature (e.g., blue represents water) or when the number of associations is so small and the colors are sufficiently memorable and distinct that they can be memorized. However, although it is sometimes possible to memorize the legend colors, a great deal of research in both graph comprehension (Carpenter & Shah, 1998; Peebles & Cheng, 2003) and interface design (Gray & Fu, 2004; Gray, Sims, Fu, & Schoelles, 2006) suggests that people use their perceptual system to reduce memory load. Thus, use of the legend is usually central to identification tasks with color-coded visualizations.

Because visual search is central to the process of legend use, research on visual search for colors (Nagy, 1999; Nagy & Sanchez, 1992) may provide some direction to understanding the process of locating colors in legends, even though that research is typically conducted under highly controlled conditions that are not identical to those in legend search. This research has consistently pointed to the impact of the discriminability of colors on the efficiency of visual search (Carter, 1982; Smallman & Boynton, 1990). Moreover, discriminability can affect the search process itself. When the target and distractor stimuli are similar, people engage in a serial search that increases in duration with the size of the set of distractors. In contrast, when the target and distractor stimuli are sufficiently distinct, people sometimes perform a faster parallel search, which is relatively insensitive to set size. This trend has been observed for visual search for colors (Nagy & Sanchez, 1992) as well as for visual search more generally (Wolfe, 2003). We will assess the use of parallel search by comparing performance on color scales of different sizes.

Turning now to quantitative tasks, we will focus in this paper on the task of making relative comparisons between two represented objects: that is, determining which is greater/less on some quantitative dimension. Whereas the process differences in identification tasks were perceptual in nature, the process differences in relative comparison tasks appear more cognitive. One can imagine two distinct strategies for making a relative comparison between two locations on a visualization. In one strategy, one locates the colors on the legend matching each of the two locations' respective colors, as in identification tasks, and then compares the legend entries, either on the basis of their relative spatial positions on the legend or on the basis of the quantitative values associated with each. By the second strategy, one compares the colors of the two visualization locations directly (e.g., judging which is darker) and thereby infers the relative difference between their respective numerical values without reference to the legend. Clearly, the second strategy is more efficient.

To support the direct comparison strategy, scale colors must be perceived as ordered. Research generally suggests that brightness provides an effective encoding of order (Spence et al., 1999). Saturation is also sometimes used for this purpose, but it appears to be inferior to brightness (Ware, 1988). Hue codes, especially rainbow/spectral codes, are frequently used in practice, but their use is not recommended by experts (Brewer, 1994) and provide poorer performance for ordering judgments. A spectrum or rainbow represents a physical ordering, but not a perceptual ordering. Moreover, spectral hue orderings conflict with the brightness ordering, as yellows and greens in the interior of the ordering are typically perceived as brighter than the other colors (Lehmann, Kaser, & Repges 1997).

In the experiments to be reported, we assessed the processes underlying the Task × Scale type interaction. In Experiment 1, we assessed participants' cognitive-perceptual processes by examining their eye movement patterns, performance accuracy, and reaction times. Specifically, we wished to determine whether certain patterns of eye movements characterized high task performance, as assessed by high accuracy and low response times, as well as to determine which color scales promote performance. In general, we predict that superior performance is associated with less frequent saccades to the legend. This should be true for both relative comparison tasks and identification tasks. In the case of relative comparison, faster comparisons are hypothesized to be associated with direct comparison between locations on the visualization, requiring few saccades to the legend. Experiments 2 and 3 specifically addresses the question of whether legend search is serial or parallel with different types of color scales by examining the effect of scale size on performance. On the basis of previous research (Nagy & Sanchez, 1992), we would expect that parallel search is more often used with multicolored scales and serial search is more often used with brightness scales because multicolored scales are generally more discriminable than brightness scales.

Experiment 1

This experiment was designed to explore both multicolored and brightness scales on both an identification task and a relative comparison task while examining eye movements.

Method

Participants

Sixteen undergraduate psychology students from George Mason University participated in this study for partial course credit. None of the participants were color blind. All participants had normal or correct-to-normal vision. The experiment lasted approximately 30 min.

Apparatus

A Power Macintosh G4 (Dual 1 GHz), equipped with a 20-inch (viewable) ViewSonic P225fb capable of running at 120 Hz, running custom software, was used to present the stimuli, control the timing of experimental events, and record participants' response times. This computer was networked to a Dell Pentium 4 machine that collected eye-tracking data in conjunction with an Eyelink 2 system (SR Research Ltd. Mississauga, Ontario, Canada). Latency between the machines is 10 ms. The Eyelink 2 tracker has 250-Hz temporal resolution and gaze position accuracy of less than 0.5° and uses an infrared video-based tracking technology to compute the center and size of the pupils in both eyes. An infrared system also tracked head motion. Even though head motion was measured, the head was stabilized by means of a chin rest. The chin rest was located 61 cm from the monitor.

Materials

A stimulus consisted of a 10 × 10 color grid and a legend (see Figures 1 to 3). The colors on each stimulus were taken from a seven-color scale. Each color was represented approximately equally in the grid, with 14 instances of five of the colors and 15 instances of the remaining two colors (which two was determined randomly). The legend of the scale colors and their respective numeric values was presented vertically on the right side of the grid, a typical location of legends in complex visualizations.

Two scale types were used, multicolored and brightness, with two instantiations of each type. The two multicolored scales were Rainbow and COAMPS. The two brightness scales were Grayscale and Greenscale. Rainbow was constructed by using the built-in "rainbow" set of hues from the R statistics/graphics environment (R Development Core Team, 2007). The COAMPS scale came directly from one of the displays of the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) meteorological modeling system (Hodur, 1997). The Grayscale was created by varying luminance in equal steps from black to white. The Greenscale came from Spence et al. (1999; called "Brightness" by those authors). It had approximately equal intervals on the Munsell value dimensions with hue and chroma held constant. All stimuli were created using the R environment (R Development Core Team, 2007). The specifications of the colors in all the color scales are available on the web in supplemental materials.

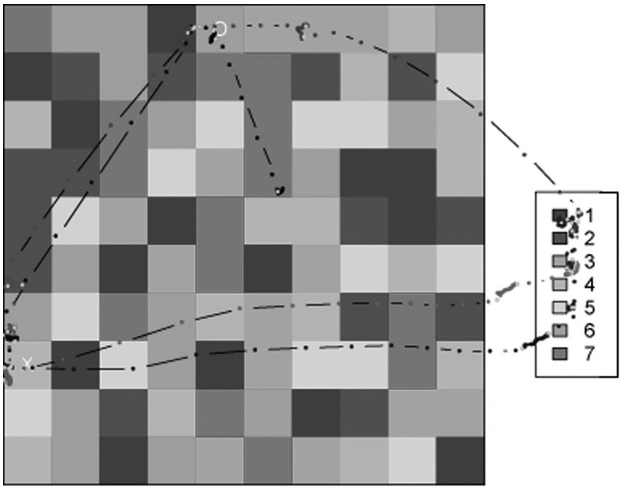

Figure 2. An example of eye movements on a multicolored scale in the comparison task. (A color version of this figure is available on the web in supplemental materials.)

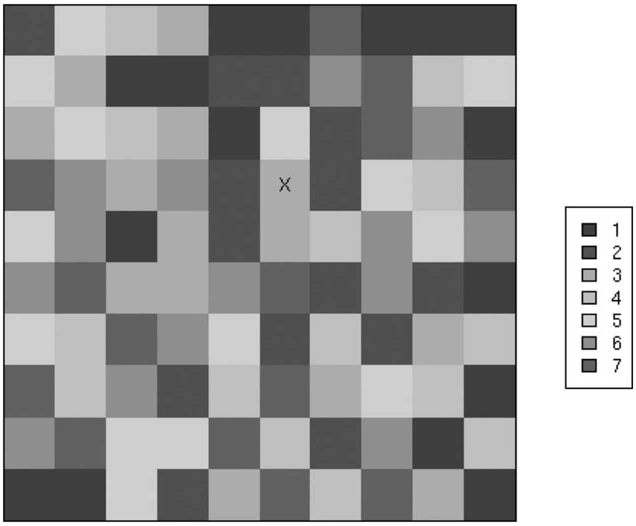

Figure 1. An example of the stimuli used in the current experiments. All stimuli consisted of a 10 × 10 grid of randomly distributed colored squares. The colors shown in this example are from the identification task and the multicolored COAMPS scale. (A color version of this figure is available on the web in supplemental materials.)

On each trial of the identification task, an "X" appeared in one cell of the grid to mark the target color to be identified. Each of the scale's seven colors was the target color on two trials.

On each trial of the comparison task, an "X" and an "O" appeared in two cells of the grid to mark the target colors to be compared. A graph was generated for each of the 21 possible pairwise comparisons for the seven colors in the scale. The assignment of the "X" or "O" to the greater value was determined randomly.

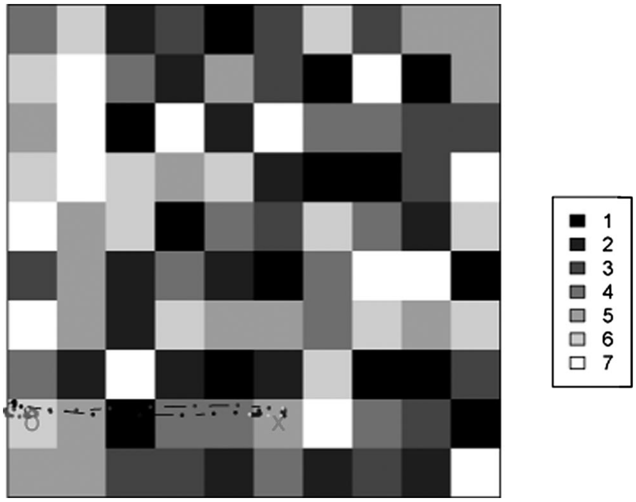

Figure 3. An example of eye movements on a brightness scale in the comparison task. (A color version of this figure is available on the web in supplemental materials.)

The location of the target(s) on the grid and the arrangement of colors on the grid were determined randomly for each subject.

Fifty-six graphs for the identification task (7 colors × 2 repetitions × 4 scales) and 84 graphs for the comparison task (21 pairwise comparisons × 4 scales) were created, for a total of 140 graphs.

Procedure

Participants were tested individually. To minimize the search requirements of the task, the location of the targets (the "X" for the identification task and both the "X" and the "O" for the comparison task) was presented on a blank screen prior to the presentation of the graph. The trial proceeded to the graph only when the participant was fixating within 2° of the target(s), and the measurement of response time began at the same time. On the identification task, the participant determined the numerical value of the color that the "X" was on, hit a key (causing the response time to be recorded) and then said the response (a no. from 1 to 7) out loud, which was recorded and scored later for accuracy. Verbal responses were used because if participants needed to look at the keypad to enter a numerical response the eye-tracker could lose calibration. On the comparison task, participants indicated whether the numerical value represented by the "X" or the "O" was greater on the legend by pressing the "z" or the "/" key (labeled with an "X" or an "O", respectively), causing the response time to be recorded. After a response was made, the next trial started.

Stimuli were counterbalanced according to task (identification or comparison) and block randomized by scale. All stimuli were presented in a random order within each block. Each participant was given a brief training session with only three colors (black, white, red) before each task began.

Results

As significant Task × Scale Type interactions were found for both accuracy, $F(1, 15) = 73.94$, $MSE = .007$, $p < .0001$, $\eta_p^2 = .83$; and response times (RT), $F(1, 15) = 22.81$, $MSE = 94.06$, $p < .001$, $\eta_p^2 = .60$; the results will be reported separately for each task. Accuracy was measured as the proportion of trials on which the subject made a correct response.

Identification Task

Accuracy and RT. Table 2 indicates that participants were faster to respond to graphs that were coded with multicolored scales than to graphs coded with brightness scales, $F(1, 15) = 16.8$, $MSE = 89.98$, $p < .001$, $\eta_p^2 = .53$. This effect was not due to a speed–accuracy trade-off as participants were also more accurate on multicolored graphs than on brightness graphs, $F(1, 15) = 61.2$, $MSE = .01$, $p < .001$, $\eta_p^2 = .80$.

Eye movement analysis. Only saccades that landed within 75 pixels of the center of a target or at least 75 pixels within the boundary of the legend were counted as fixations. The item closest to the point of fixation was classified as the fixated item. Consecutive fixations on the same item were considered a single gaze. Eye movements were classified as saccades if they either (a) exceeded a speed of 30°/s and an acceleration of $8{,}000°/s^2$, or (b) exceeded an acceleration of $8{,}000°/s^2$ and a distance of 0.2°.

Table 2
Results From Experiment 1

	Multicolored scale		Brightness scale	
	M	*SD*	*M*	*SD*
Identification task				
Reaction time[a]	1.28	.47	1.71	.78
Accuracy	0.93	.04	0.63	.15
Comparison task				
Reaction time[a]	1.76	.66	1.46	.58
Accuracy	0.89	.08	0.95	.04

[a] Given in seconds.

Table 3 shows the results. Participants spent longer looking at the target with brightness scales than with multicolored scales, $F(1, 15) = 27.7$, $MSE = 11.90$, $p < .001$, $\eta_p^2 = .65$. Also, participants spent more time looking at the legend with brightness scales than with multicolored scales, $F(1, 15) = 9.2$, $MSE = 12.40$, $p < .01$, $\eta_p^2 = .38$. Finally, participants made more saccades to the legend with brightness than with multicolored scales, $F(1, 15) = 10.7$, $MSE = .08$, $p < .01$, $\eta_p^2 = .42$.

Comparison Task

Accuracy and RT. As Table 2 indicates, participants were faster to respond to graphs that were coded with brightness scales than to graphs that were multicolored, $F(1, 15) = 23.9$, $MSE = 29.75$, $p < .001$, $\eta_p^2 = .61$. This effect was not due to a speed–accuracy trade-off as participants were also more accurate on brightness graphs than on multicolored graphs, $F(1, 15) = 15.6$, $MSE = .002$, $p = .001$,$\eta_p^2 = .51$.

Eye movement analysis. As a comparison of Figure 2 and Figure 3 suggests, the perceptual process was quite distinct for different scales on the comparison task. To explore the differences systematically, average gaze duration on targets and legend was examined. Interestingly, as Table 3 suggests, the time spent looking at the two targets was greater for brightness scales than for multicolored scales, $F(1, 15) = 15.1$, $MSE = 2.81$, $p < .05$, $\eta_p^2 = .50$. In contrast, the time spent looking at the legend was much greater for multicolored scales than for brightness scales, $F(1, 15) = 40.5$, $MSE = 13.97$, $p < .0001$, $\eta_p^2 = .73$.

Participants using brightness scales looked back and forth between the two targets much more frequently than when using the multicolored scale, $F(1, 15) = 33.0$, $MSE = .03$, $p < .0001$, $\eta_p^2 = .69$. In contrast, participants made many more saccades to the legend when using a multicolored scale than when using a brightness scale, $F(1, 15) = 30.7$, $MSE = .07$, $p < .0001$, $\eta_p^2 = .67$.

Discussion

The results replicated the previously found Task × Scale Type interaction (Merwin & Wickens, 1993; Phillips, 1982); that is, performance was superior using multicolored scales on the identification task and superior using brightness scales on the relative comparison task. Eye-movement analyses suggested process similarities and differences among the four task-scale conditions.

Table 3
Results From Experiment 1 Eye Tracking Data

	Multicolored scale		Brightness scale	
	M	*SD*	*M*	*SD*
Identification task				
Target fixation duration[a]	0.48	.14	0.68	.25
Legend fixation duration[a]	0.68	.26	0.80	.31
Saccades to legend	2.40	.60	2.70	.70
Comparison task				
Target fixation duration[a]	0.99	.29	1.06	.31
Legend fixation duration[a]	0.57	.35	0.30	.29
Saccades between targets	1.10	.30	1.50	.40
Saccades to legend	1.40	.80	0.90	.80

Note. Target data for the comparison task include both targets.
[a] Given in seconds.

For the identification task, more time was spent looking at the target and legend and there were more saccades to the legend with a brightness scale than with a multicolored scale. These results suggest that multicolored scales afford more efficient search of the legend. The possibility that participants solved the task by memorizing the color-number associations appears unlikely as they devoted approximately half of the trial response times to examining the legend with both of the respective scale types (compare Table 2 and Table 3, identification task). The greater efficiency of visual search with multicolored scales is more likely due to the greater discriminability of those scales' colors relative to the brightness scales (Carter, 1982; Smallman & Boynton, 1990).

The comparison task showed a completely different pattern of results. When using a brightness scale, participants looked back and forth between the two target colors, apparently determining which one was lighter or darker. With the multicolored scales, participants needed to look at the legend to compare the colors, either by determining which was higher (or lower) on the legend or by reading and comparing the numbers associated with each color. These differences in process are suggested by reaction times as well, as direct comparison between targets is more efficient than legend-based search and comparison.

The findings shed light on salient process differences between the two tasks. Most significantly, the comparison task is clearly not solved by repeating the process used in the identification task twice over, once for each target, and then comparing the results of the two identifications. If this "double identification" process were adopted, response times and legend saccades for comparisons would be at least twice as great as those for identifications. However, on the contrary, the quantity and duration of fixations on the legend are lower on the comparison task than on the identification task, especially for the brightness scales. Similarly, the response times are lower on the comparison task than on the identification task for the brightness scales and, though response times are higher for the multicolored scales on the comparison task, the increase in time (38%) is far below the roughly 100% increase predicted by the double identification hypothesis.

The strategy used to solve comparison tasks in place of the double identification strategy likely differs depending on scale type. In the case of the brightness scales, the eye tracking data point to a difference in cognitive strategy between the identification and comparison task. The great decrease in rate and duration of legend fixations, together with the relatively high rate of saccades between the targets, suggests that participants typically solved the comparisons by means of direct comparison between the target colors with little reference to the legend.

The case of the multicolored scales is less clear. The increase in response times and legend fixations between identification and comparison tasks suggest that participants continued to search the legend when solving comparison problems. However, the finding that the response time increase was far less than the doubling predicted by the double identification hypothesis suggests a much more efficient search process is being used. Perhaps participants are able to adjust their visual search strategy to the task demands, adopting a parallel search strategy for comparison tasks, but performing serial search for identification tasks. This seems reasonable, as identification requires the participant to make a more precise determination of the matching color on the legend, so as to determine the corresponding numerical value. In contrast, the comparison task only requires the participant to determine the positions of legend colors with sufficient precision to judge which of two is higher (or lower) on the legend. One common method to differentiate between serial and parallel visual search is by determining whether search time is influenced by the set size—that is, the number of "distractors" or in the present context the number of colors in the scale. Parallel search, in contrast to serial search, is relatively unaffected by set size (Nagy & Sanchez, 1992).

Of course, strategies other than parallel search may account for the surprisingly rapid comparisons made with multicolored scales. For instance, participants may employ logical strategies, such as responding after identifying one of the comparison colors as an endpoint in the scale without locating the other color. Alternately, they may memorize the values/positions of some of the colors. In the next two experiments, we will compare performance on scales of different sizes in hopes of resolving this question: in Experiment 2 for the identification task and in Experiment 3 for the comparison task.

Experiment 2

This experiment was designed to compare multicolored and brightness scales on an identification task across different numbers of scale colors. Previous researchers have made explicit suggestions on the number of scale colors that should be used in visualizations, ranging from 4 to 10 and have suggested that having more than 12 colors on a scale hurts discriminability and adds confusion (Krebs & Wolf, 1979; Rice, 1991). A quick examination of most complex visualizations, however, shows many violations of these suggestions. In practice, users want enough precision to differentiate quantitative values at a level useful for a particular domain. For meteorological forecasting, for example, large temperature ranges (greater than 5 degrees) are not usually informative to the forecaster. Thus, having many colors on a scale to show a large range of quantitative values seems to be an important concern of users of complex visualizations.

Method

Participants

Thirteen undergraduate psychology students from George Mason University participated in this study for partial course credit. None of the participants were color blind. All participants had normal or correct-to-normal vision. The experiment lasted approximately 25 min.

Materials

The stimuli consisted of 10 × 10 color grids and a legend. The color grids and legends were constructed in the same way as in Experiment 1, with scales consisting of 4, 10, and 20 colors for each type of scale. The same four scales were used as in Experiment 1: the ordered brightness scales, Greenscale and Grayscale, and the multicolored scales, COAMPS and Rainbow. A total of 136 graphs was constructed (4 scale types × 34 [= 20 + 10 + 4] graphs per scale type).

Procedure

The procedure was the same as the identification task in Experiment 1.

Results

As Figure 4 suggests, participants responded faster to graphs that were multicolored than to the brightness graphs, $F(1, 12) = 5.1$, $MSE = 397.20$, $p < .05$, $\eta_p^2 = .30$. Mean response times with multicolored scales were 1.9 s, 2.7 s, and 3.4 s on 4-, 10-, and 20-color scales, respectively; with the brightness scales they were 2.4 s, 3.0 s, and 3.5 s on 4-, 10-, and 20-color scales, respectively. Linear trend analysis revealed that response time increased with scale size, $F(1, 12) = 30.47$, $MSE = 764.29$, $p < .0001$, $\eta_p^2 = .72$. There was no interaction between type of scale and number of colors, $F(1, 12) = 2.62$, $MSE = 254.34$, $p > .10$, $\eta_p^2 = .18$.

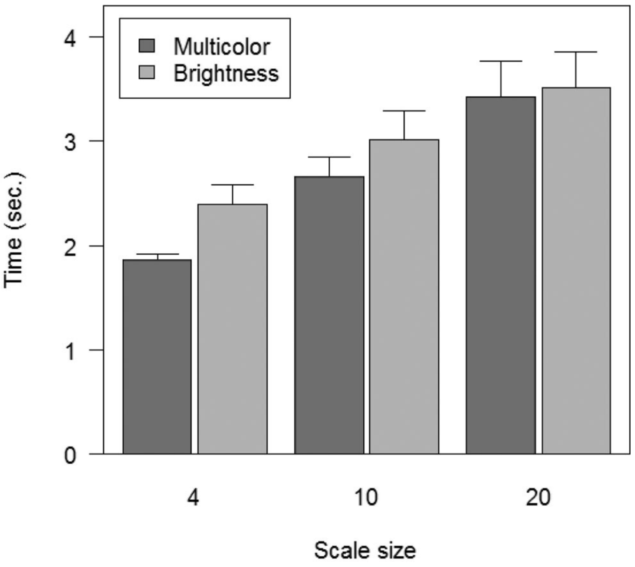

Figure 4. Experiment 2 reaction time for the identification task for different sized scales. Error bars are standard error of the mean. (A color version of this figure is available on the web in supplemental materials.)

As Figure 5 indicates, accuracy showed a somewhat similar pattern. Mean accuracies with the multicolored scales were .98, .81, and .59 with 4-, 10-, and 20-color scales, respectively, whereas accuracies with the brightness scales were .87, .44, and .19 with 4-, 10-, and 20-color scales, respectively. Participants were much more accurate on multicolored graphs than on brightness graphs, $F(1, 12) = 140.9$, $MSE = .01$, $p < .0001$,$\eta_p^2 = .92$. Linear trend analysis revealed that accuracy decreased markedly as more colors were on the scale, $F(1, 12) = 399.1$, $MSE = .009$, $p < .0001$, $\eta_p^2 = .97$. However, the accuracy decreased at a faster rate for brightness colors than for multicolored colors as shown by a significant interaction, $F(1, 12) = 44.98$, $MSE = 0.006$, $p < .001$, $\eta_p^2 = .79$. As can be seen in Figure 5, this differential effect of set size on accuracy occurred primarily between set sizes 4 and 10.

Discussion

These results replicated the finding of superior identification with multicolored scales as compared to brightness scales across highly unequal scale sizes. The only exception was the comparable response times for the largest scale size. Identification performance degraded markedly with increased scale size, especially for brightness scales. These results, together with those of Experiment 1, suggest that multicolored scales of 10 or more colors should not be used for identification tasks, and that brightness scales in general are ill-suited for such tasks.

The linear relation between response times and scale size for both scale types suggests that participants conducted serial searches of the legend, even with the more discriminable multicolored scales. However, faster response times with multicolored scales suggest that their superior discriminability permitted a faster serial search relative to brightness scales (Carter, 1982; Smallman & Boynton, 1990).

An alternate explanation for the effect of scale size on performance may be that it reflects capacity limitations in working memory (Cowan, 2001; Miller, 1956). This explanation assumes that participants attempt to memorize the color-number associations. However, eye movement analyses in Experiment 1 with seven-color scales and previous research (Carpenter & Shah, 1998; Peebles & Cheng, 2003; Trafton et al., 2002) provide evidence that people rely on perception (e.g., looking at the legend) rather than working memory when using complex visualizations. Recall that participants in Experiment 1 devoted approximately half of their trial response times to examining the legend when performing identification tasks with each scale type.

These conclusions will be further elaborated in the context of the next experiment that used a similar parametric methodology for the comparison task.

Experiment 3

This experiment was designed to explore both multicolored and brightness scales on a comparison task with different-sized scales.

Method

Participants

Twenty-two undergraduate psychology students from George Mason University participated in this study for partial course

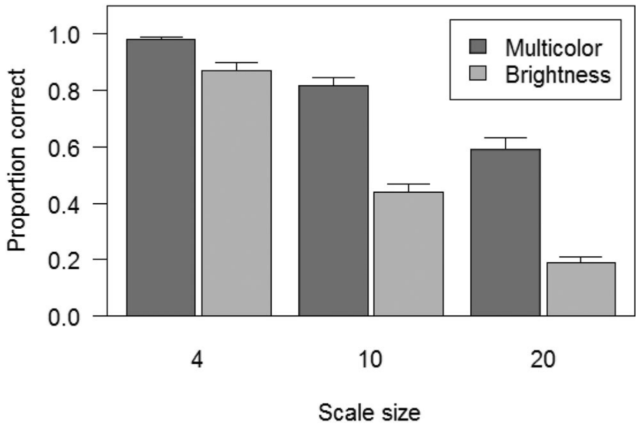

Figure 5. Experiment 2 accuracy for the identification task for different sized scales. Error bars are standard error of the mean. (A color version of this figure is available on the web in supplemental materials.)

credit. None of the participants were color blind. All participants had normal or correct-to-normal vision. The experiment lasted approximately 45 min.

Materials

The stimuli consisted of 10 × 10 color grids and a legend. The color grids were constructed in the same way as in Experiment 1, with scales consisting of 4, 10, and 20 colors for each of the four scales tested. The same four scales were used as in the preceding experiments: Greenscale and Grayscale brightness scales and COAMPS and Rainbow multicolored scales. A total of 964 graphs was constructed (4 scale types × 241 [= 6 + 45 + 190] graphs for each scale type), but each participant was assigned only one brightness and one multicolored scale, or a total of 482 trials. This assignment was counterbalanced so that an equal number of participants were tested on each scale.

Procedure

The procedure was the same as in the comparison task in Experiment 1.

Results

As Figure 6 suggests, participants were faster to respond with the brightness scales than with the multicolored scales, $F(1, 21) = 53.027$, $MSE = 543.04$, $p < .001$, $\eta_p^2 = .72$. More surprising, response times actually decreased slightly with increased scale size, $F(1, 21) = 4.46$, $MSE = 56.33$, $p < .05$, $\eta_p^2 = .18$. There was a significant interaction between type of scale and number of colors, $F(1, 21) = 13.3$, $MSE = 45.62$, $p < .001$, $\eta_p^2 = .39$: The brightness scales had a relatively constant reaction time, whereas multicolored response times decreased slightly (by .27 s) as the number of colors rose from 4 to 10.

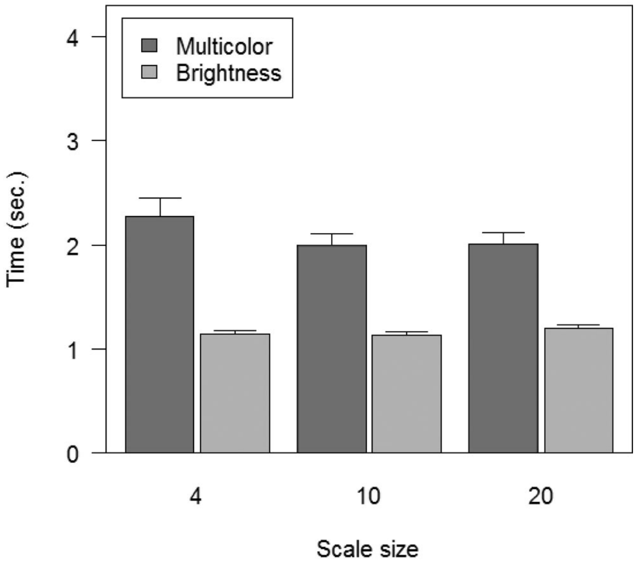

Figure 6. Experiment 3 reaction time for the comparison task for different sized scales. Error bars are standard error of the mean. (A color version of this figure is available on the web in supplemental materials.)

Figure 7 shows that accuracy followed a somewhat similar pattern. Participants were more accurate on the brightness graphs than on the multicolored graphs, $F(1, 21) = 19.8$, $MSE = 23$, $p < .0001$, $\eta_p^2 = .49$. Overall, accuracy decreased as more colors were on the scale, $F(1, 21) = 17.4$, $MSE = 20.7$, $p < .0001$, $\eta_p^2 = .40$, but there was a significant Scale Type × Scale Size interaction, $F(1, 21) = 6.2$, $MSE = 38$, $p < .05$, $\eta_p^2 = .23$. Brightness scales' accuracy mirrored the response time data, staying relatively constant across different scale sizes, whereas the multicolored scales' accuracy diverged from their response times, decreasing slightly as the number of colors on the scale increased. Thus, the slight decrease in response times for the multicolored scales may reflect a speed–accuracy trade-off.

Discussion

In contrast to the identification task, comparison task performance was relatively robust in the face of increases in scale size, especially for the brightness scales (.97, .97, and .95, for 4, 10, and 20 colors, respectively). Even the multicolored scales had fairly

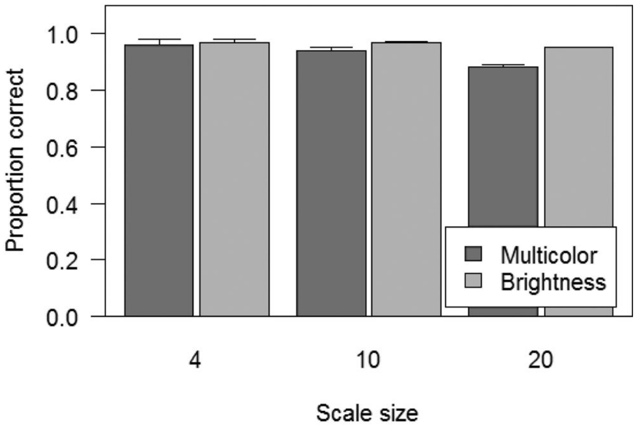

Figure 7. Experiment 3 accuracy for the comparison task for different sized scales. Error bars are standard error of the mean. (A color version of this figure is available on the web in supplemental materials.)

high accuracies, except with the largest scales (.96, .94, .88, respectively). Likewise, the typically found superiority of the brightness scales over the multicolored scales on relative comparison tasks was a robust finding over a wide range of scale sizes. The sole exception was accuracy on the smallest, 4-color scales, probably reflecting a ceiling effect. Taken together with Experiment 2, these results demonstrate the robustness of the Scale Type × Task interaction demonstrated in Experiment 1 with 7-color scales.

The absence of increased response times with increased scale size has important, but differing, implications for the brightness and multicolored scales. For the brightness scales (1.1 s, 1.1 s, and 1.2 s for 4-, 10-, and 20-color scales, respectively), this finding supports the conclusion from Experiment 1, based on the analysis of eye movements, that participants compared the two targets on the grid directly with little reliance on the legend. This stability suggests that the same process of comparing the two target stimuli characterizes the process for scales of differing sizes. Although the overall discriminability of the brightness values of the scale colors diminishes with increased scale size, the reduced discriminability imposed no added penalty and was clearly sufficient to support fast accurate responses for all scale sizes.

In the case of the multicolored scales (2.3 s, 2.0 s, and 2.0 s for 4-, 10-, and 20-color scales, respectively) the response time trend with comparison tasks stood in marked contrast to that in Experiment 2 with identification tasks. In the latter, responses markedly slowed with increased scale size, whereas in the present study with comparison tasks, the response times sped up slightly with increased scale size. This finding suggests that participants conducted a parallel search of the legend when making comparisons. Although it remains possible that participants also made deductive conclusions after locating only one endpoint color on the legend or sometimes responded based on remembered information, neither of these processes would be unaffected by increased scale size. An endpoint color, whether located visually or memorized, would account for a diminishing proportion of the pairwise comparisons as scale size increased (e.g., 50% with 4 colors vs. 10% with 20 colors). Likewise, memorization would become more difficult as the set of colors to memorize increased. In contrast, relatively constant speeds across different set sizes are commonly characterized in the visual search literature as evidence of parallel search, whether studying color (Nagy & Sanchez, 1992) or visual search more generally (Wolfe, 2003).

General Discussion

We have replicated and elaborated on previous findings (Merwin & Wickens, 1993; Phillips, 1982) of a Scale × Task interaction for color scales: Namely, multicolored scales were faster and more accurate than brightness scales on absolute-value identification tasks, whereas brightness scales were faster and more accurate than multicolored scales on relative comparison tasks. We have replicated this interaction with a broad range of scale sizes, thus providing evidence for its robustness. In addition, we have offered a process explanation to explain the interaction. The hypothetical process explanation offered on the basis of previous research in Table 1 may now be expanded into the analysis summarized in Table 4.

Table 4
Expanded Explanation of Task × Scale Type Interaction (New Features Italicized)

	Multicolored scale	Brightness scale
Identification task	**Easy:** highly discriminable; *fast serial search of legend*	**Hard:** less discriminable; *slow serial search of legend*
Comparison task	**Hard:** no/little perceived ordering; *parallel search of legend*	**Easy:** clear perceived ordering; *direct comparison of targets*

Experiment 1 explored this effect with seven-color scales while analyzing eye movements. This analysis indicated that people spent more time looking at the legend with brightness scales than with multicolored scales when performing identifications, but when making comparisons they spent roughly half the time looking at the legend with brightness scales in comparison to multicolored scales. At the same time, they made more saccades between targets to make comparisons using brightness scales, suggesting that they were making a direct comparison between the target colors without relying very much on the legend.

Experiments 2 and 3 explored the impact of having more and fewer colors in the scale and replicated the basic findings from Experiment 1. In addition, Experiment 2 showed that accuracy of identification decreased and reaction time increased as the number of colors on the scale increased. The fact that response times increased with increased scale size indicates that people did not perform a parallel visual search of the legend, even with the more discriminable multicolored scales. Instead, they seemed to perform a serial search, though more efficiently with the more discriminable multicolored scales than with the brightness scales.

In contrast, Experiment 3 showed that on comparison tasks, accuracy and response times remained relatively constant for both brightness scales and multicolored scales with increased scale size. The finding for brightness scales suggests that the process of direct comparison between target colors used with these scales is robust against the decreasing discriminability that accompanies increased scale size.

The results for multicolored scales suggests, interestingly, that when making comparisons using the legend, people conducted a parallel search of the legend, in contrast to the situation with identification tasks using the same scales. The eye movement data indicated that people used the legend to solve both identification and comparison tasks when using multicolored scales. However, the findings of Experiments 2 and 3 suggest that their process of using the legend is very different in the two tasks. In the identification task, they must retrieve the precise value corresponding to a color in the legend; this requirement might constrain them to perform a serial search. In the comparison task, in contrast, a more approximate localization of the colors on the legend may often suffice to determine which is higher/greater, and so a more efficient parallel search may be employed. These results suggest that the use of serial versus parallel search is not only determined by the nature of the color scale, as other researchers have emphasized (Nagy & Sanchez, 1992), but by task demands as well.

Following the location of approximate legend positions, people can compare those positions spatially to arrive at a response, without needing to extract numeric values from the legend. This spatial strategy is comparable to the comparison strategy hypothesized by Gillan and Lewis (1994) for graphs.

Practical Implications

If reliance on a legend is associated with inferior performance, perhaps it would be best to eliminate the legend where possible and place the numerical values directly within the visualization. Indeed in designing graphs, that is the recommended practice, except where it produces too much visual clutter (Kosslyn, 1994). Thus, digits on maps result in far superior performance in comparison to a brightness representation of quantity for identification and only slightly inferior performance to brightness for relative comparisons (Phillips, De Lucia, & Skelton, 1975), but at the cost of high clutter. Lines on maps, such as contour lines or isotherms, reduce clutter compared to digits but at the cost of much inferior relative comparisons compared to brightness coding (Phillips et al., 1975). Thus, considerations of performance on some tasks must be balanced against concerns about clutter, which can impair performance on other tasks.

Because visual search of the legend is central to most of the conditions we examined, efforts to facilitate legend search should enhance the usability of color-coded visualizations. The function of the legend in such visualizations is comparable to the function of the *y*-axis in graphs as depicted in the componential model of graph interpretation proposed by Gillan and Lewis (1994). These authors proposed that placing labeled *y*-axes on both sides of a graph may facilitate the search for the values of data points in the graph. Similarly, it is possible that placing more than one copy of the legend in the margins of a visualization may facilitate the extraction of values.

Our results also highlighted the limitations of certain color scales for identification tasks. Specifically, accuracy of identifications is significantly impacted once the size of multicolored scales reaches 10 colors or more, whereas brightness scales appear ill-suited in general for identification tasks. In contrast, both scale types supported high accuracy on comparison tasks and performance was fairly robust in the face of large increases in scale size. Brightness scales distinguished themselves from multicolored scales primarily by the faster speed of response they supported.

These experiments suggest that there is no perfect scale for every task. If it is possible to know in advance what type of task people will perform on a specific visualization, then the decision about which type of scale to use is very clear: use a multicolored scale for identification tasks and a brightness scale for comparison tasks. However, if one wishes to support both identification of absolute values and relative comparisons with a single color code, then the decision of what type of scale to use may be governed by one's priorities. For example, arguing that "relative height is more important than absolute height for children using atlases" (p. 1143), Phillips (1982) recommended using brightness scales for school atlases despite their inferior accuracy for absolute height judgments. However, if high accuracy on both tasks is one's prime concern, then multicolored scales would be preferred, despite the slow performance we found they afford on relative comparison tasks.

Contributions of Color Process Analysis

What practical contributions does a color process analysis offer? First, by explaining why a particular type of color scale is useful for a particular task, a process analysis points to ways in which research may be applied in practice to the design of color scales. If the superiority of multicolored scales for identification tasks lies in the discriminability of the colors, then one would do well to explore exactly how discriminable the colors should be and perhaps develop color-difference standards to guide designers of color scales to produce easy-to-use scales. Although the strength of brightness scales lies in their encoding of order, it may also prove to be important for efficient comparison that the brightness of their colors be easily discriminable across the scale. Research that may be applied to this includes findings that scales with perceptually equal brightness differences conform to a logarithmic function of luminosity L (Whittle, 1992), rather than to equal intervals of L. Finally, evidence that stimulus size and background color/brightness affects the gamut of recognizable color/brightness differences (Brown & MacLeod, 1997; Carter & Carter, 1988) should inform guidelines for the construction of the visualization itself. In sum, process analyses point to relevant bodies of research that may inform application design.

Second, process analyses may suggest new types of color scales. To illustrate this, we may address a question that the present research begs: Can a color scale be constructed that is useful for both identification and comparison tasks by being both multicolored and ordered by brightness? Spence et al. (1999) examined this question only with regard to relative comparison (as well as another quantitative task) but not identification tasks. Reviewing the literature, they found conflicting evidence as to whether multicolored brightness scales are as effective as monochrome brightness scales for quantitative tasks. They offered the perceptual linearity hypothesis to resolve this question:

> for a coding assignment to be perceptually linear, it must be possible to form an additive weighted combination of the Cartesian coordinates of each color in perceptual space such that the combination correlates maximally with a linear sequence of numbers. (p. 397)

Spence et al. (1999) hypothesized that perceptual linearity is a requirement for a multicolored brightness scale applied to quantitative tasks. Perceptual linearity is only possible if luminosity is more highly weighted than the two hue dimensions, since luminosity is a linear color attribute whereas hue is a circular attribute. They offered generally supportive evidence that a multicolored perceptually linear scale was as effective as a monochrome brightness scale on quantitative tasks. It would be interesting to test their multicolored perceptually linear scale on identification tasks. Encouraging evidence from the visual search literature (Nagy, 1999) suggests that the brightness variation in the scale should not interfere with the hue-based search of the legend that is required for identification tasks. Further, Jameson, Kaiwi, and Bamber (2001) demonstrated the effectiveness of codes varying in both hue and brightness on classification tasks. A multicolored ordered-brightness scale that was effective for both identification and relative comparison tasks may make it possible to avoid the sort of compromises discussed above for school atlases.

Conclusions

We have replicated previous research showing that multicolored scales are superior to ordered brightness scales for supporting identification tasks with complex visualizations, whereas ordered brightness scales are superior to multicolored scales for supporting relative comparisons, using scales of different sizes. In addition, we have examined the processes by which the tasks are solved with either scale type. By studying eye movements and by comparing performance on scales of different sizes, we found evidence that (a) people perform identification tasks by means of a serial visual search of the legend, whose speed is sensitive to the number of scale colors and the discriminability of the colors, and (b) people perform relative comparison tasks using different processes for multicolored versus brightness scales. With multicolored scales, they perform a parallel search of the legend whose speed is relatively insensitive to the size of the scale, whereas with brightness scales, people usually compare the colors on the visualization directly, with little reference to the legend. This analysis may help guide scale selection and visualization design by pointing to relevant bodies of research to apply.

References

Brewer, C. A. (1994). Color use guidelines for mapping and visualization. In M. A. MacEachren & D. R. F. Taylor (Eds.), *Visualization in modern cartography.* Tarrytown, NY: Elsevier.

Brown, R., & MacLeod, D. (1997). Color appearance depends on the variance of surround colors. *Current Biology, 7,* 844–849.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4,* 75–100.

Carter, R. (1982). Visual search with color. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 127–136.

Carter, R., & Carter, E. (1988). Color coding for rapid location of small symbols. *Color Research and Application, 13,* 226–234.

Christ, R. E. (1975). Review and analysis of color coding research for visual displays. *Human Factors, 7,* 542–570.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–185.

Gillan, D. J., & Lewis, R. (1994). A componential model of human interaction with graphs: 1. Linear regression modeling. *Human Factors, 36,* 419–440.

Gillan, D. J., Wickens, C. D., Hollands, J. G., & Carswell, C. M. (1998). Guidelines for presenting quantitative data in HFES publications. *Human Factors, 40,* 28–41.

Gray, W. D., & Fu, W. T. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science, 28,* 359–382.

Gray, W. D., Sims, C. R., Fu, W. T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review, 113,* 461–482.

Hodur, R. M. (1997). Coupled ocean/atmosphere mesoscale prediction system. *Monthly Weather Review, 125,* 1414–1430.

Jameson, K. A., Kaiwi, J. L., & Bamber, D. (2001). Color coding information: Assessing alternative coding systems using independent brightness and hue dimensions. *Journal of Experimental Psychology: Applied, 7,* 112–128.

Kosslyn, S. M. (1994). *Elements of graph design.* New York: Freeman.

Krebs, M. J., & Wolf, J. D. (1979). Design principles for the use of color in displays, *Proceedings of the Society for Information Display, 20,* 10–15.

Lehmann, T. M., Kaser, A., & Repges, R. (1997). A simple parametric equation for pseudocoloring grey scale images keeping their original brightness progression. *Image and Vision Computing, 15,* 251–257.

Merwin, D. H., & Wickens, C. D. (1993). Comparison of eight color and gray scales for displaying continuous 2D data. *Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting,* 1330–1334. Santa Monica, CA: Human Factors and Ergonomic Society.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Nagy, A. L. (1999). Interactions between achromatic and chromatic mechanisms in visual search. *Vision Research, 39,* 3253–3326.

Nagy, A. L., & Sanchez, R. (1992). Chromaticity and luminance as coding dimensions in visual search. *Human Factors, 34,* 601–614.

Peebles, D., & Cheng, P. C. H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors, 45,* 28–46.

Phillips, R. J. (1982). An experimental investigation of layer tints for relief maps in school atlases. *Ergonomics, 25,* 1143–1154.

Phillips, R. J., De Lucia, A., & Skelton, N. (1975). Some objective tests of the legibility of relief maps. *Cartographic Journal, 12,* 39–46.

R Development Core Team. (2007). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rice, J. F. (1991). Ten rules for color coding. *Information Display, 3*(9), 12–14.

Smallman, H. S., & Boynton, R. M. (1990). Segregation of basic colors in an information display. *Journal of the Optical Society of America A, 7,* 1985–1994.

Spence, I., Kutlesa, N., & Rose, D. L. (1999). Using color to code quantity in spatial displays. *Journal of Experimental Psychology: Applied, 5,* 393–412.

Trafton, J. G., Marshall, S., Mintz, F. E., & Trickett, S. B. (2002). Extracting explicit and implicit information from complex visualizations. In B. Meyer & H. Narayanan (Eds.), *Diagrammatic representation and inference* (pp. 206–220). Berlin: Springer-Verlag.

Ware, C. (1988). Color sequences for univariate maps: Theory, experiments, and principles. *IEEE Computer Graphics and Applications, 8,* 41–49.

Whittle, P. (1992). Brightness, discriminability and the “crispening effect.” *Vision Research, 32,* 1493–1507.

Wickens, C. D., Merwin, D. H., & Lin, E. L. (1994). Implications of graphics enhancements for the visualization of scientific data: Dimensional integrity, stereopsis, motion, and mesh. *Human Factors, 36,* 44–61.

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Science, 7*(2), 70–76.

Received November 16, 2007
Revision received November 24, 2008
Accepted December 3, 2008 ■