Cognitive Models of the Influence of Color Scale on Data Visualization Tasks

Leonard A. Breslow, Naval Research Laboratory, Washington, D.C.,
Raj M. Ratwani, George Mason University, Fairfax, Virginia, and
J. Gregory Trafton, Naval Research Laboratory, Washington, D.C.

Objective: Computational models of identification and relative comparison tasks performed on color-coded data visualizations were presented and evaluated against two experiments. In this context, the possibility of a dual-use color scale, useful for both tasks, was explored, and the use of the legend was a major focus. **Background:** Multicolored scales are superior to ordered brightness scales for identification tasks, such as determining the absolute numeric value of a represented item, whereas ordered brightness scales are superior for relative comparison tasks, such as determining which of two represented items has a greater value. **Method:** Computational models were constructed for these tasks, and their predictions were compared with the results of two experiments. **Results:** The models fit the experimental results well. A multicolored, brightness-ordered dual-use scale supported high accuracy on both tasks and fast responses on a comparison task but relatively slower responses on the identification task. **Conclusion:** Identification tasks are solved by a serial visual search of the legend, whose speed and accuracy are a function of the discriminability of the color scales. Comparison tasks with multicolored scales are performed by a parallel search of the legend; with brightness scales, comparison tasks are generally solved by a direct comparison between colors on the visualization, without reference to the legend. Finally, it is possible to provide users a dual-use color scale effective on both tasks. **Application:** Trade-offs that must typically be made in the design of color-coded visualizations between speed and accuracy or between identification and comparison tasks may be mitigated.

INTRODUCTION

A designer of a data visualization needs to be guided by the tasks to which the visualization will be applied. For example, observational studies of professional and student meteorologists and scientists have shown that they use visualizations to perform many identification tasks (e.g., determining the temperature at a specific location) and many comparison tasks (e.g., determining where the highest temperature or humidity is) (Trafton et al., 2000; Trafton, Marshall, Mintz, & Trickett, 2002; Trafton, Trickett, & Mintz, 2005; Trickett & Trafton, 2007).

In the case of color-coded visualizations, experimental research has highlighted the relative strengths and weaknesses of unordered multicolored versus ordered brightness scales for different tasks. In particular, multicolored scales have been found to be well suited for identification tasks, whereas ordered brightness scales have been found to be well suited for relative comparison tasks, but neither of these scale types is effective on the other's forte (Breslow, Trafton, & Ratwani, 2009; Merwin & Wickens, 1993; Phillips, 1982).

In this article, we are concerned with the particular identification task of determining the *absolute* quantitative value of a color-coded item and the comparison task of determining the *relative* quantitative relation (greater or less) between two color-coded items in a visualization. As an example, Figure 1 shows a color-coded weather map. A possible identification

Address correspondence to Leonard A. Breslow, Naval Research Laboratory, Code 5515, 4555 Overlook Ave. SW, Washington, DC, 20375; breslow@aic.nrl.navy.mil. ***HUMAN FACTORS,*** Vol. 51, No. 3, June 2009, pp. 321-338. DOI: 10.1177/0018720809338286. Copyright © 2009, Human Factors and Ergonomics Society.

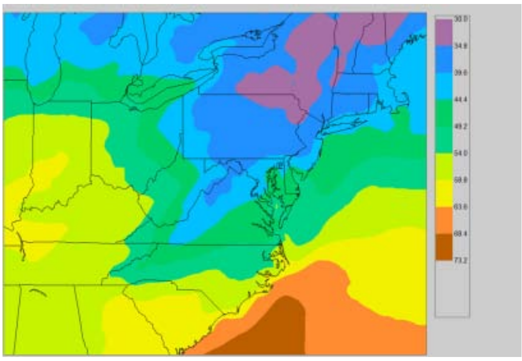

Figure 1. Example of color-coded visualization: weather map displaying temperatures.

task might be to determine the temperature of New York City, whereas a possible comparison task might be to determine which city is warmer, New York or Washington, D.C. As this figure illustrates, many practical applications suffer from the respective weaknesses of both unordered multicolored and ordered brightness codes. Although the color scale coding this figure is partially ordered by some combination of the color attributes hue, saturation, and brightness, the scale provides no overarching perceptual order. As a result, it is not always easy to compare two colors on the map to determine which location is warmer without referring to the legend. At the same time, matches to the legend are impaired by similarities between certain adjacent colors on the legend, thus impeding identification. In what follows, we study fairly “pure” examples of unordered multicolored and ordered brightness scales to experimentally differentiate the effects of these respective scale types.

In addition, a main focus is on the role of the legend in the use of visualizations. Whereas a great deal of attention has focused on the influence of color on the perception and understanding of visualizations (Hoffman, Detweiler, Lipton, & Conway, 1993), much less attention has been devoted to the effects of color on the use of the legend, despite observational evidence that novice and expert users make extensive use of the legend in the performance of practical applications of visualizations (Carpenter & Shah, 1998; Peebles & Cheng, 2003; Trafton et al., 2002). To concentrate on legend use, we use simplified visualizations and experimental procedures that reduce the search demands on visualization processing.

Breslow et al. (2009) proposed a general process model, summarized in Table 1, to account for the Color Scale × Task interaction discussed earlier, that is, the superiority of unordered multicolored scales on identification tasks and of ordered brightness scales on relative comparison tasks. The model accounts for the findings primarily in terms of the nature and efficiency of the visual search of the legend that each task–code combination supports. On the identification task, people perform a visual search of the legend to locate the legend color matching the target item in the visualization, as evidenced by eye movement and other data (Breslow et al., 2009). The efficiency of visual search for colors is highly sensitive to the discriminability of the set of colors being searched (Carter, 1982; Smallman & Boynton, 1990). As multicolored scales are typically more discriminable than brightness scales, search is faster with multicolored than with brightness scales (Nagy & Sanchez, 1992).

In some cases, high discriminability enables people to perform a faster parallel search for colors, as evidenced by relatively invariant search times across color scales of highly unequal sizes (Nagy, 1999; Nagy & Sanchez, 1992). In the context of color-coded visualizations, such findings suggest the use of parallel search in the comparison task, in contrast with the serial search used in the identification task (Breslow et al., 2009). Although less precise than serial search, parallel search is adequate to the task of determining relative quantity because that task does not require as precise a localization of each legend color as the identification task does; one must simply locate the two legend colors correctly enough to determine which is spatially higher or lower on the legend.

Finally, the model suggests that one performs the relative comparison task using a different process with brightness scales than with multicolored scales. With brightness scales, people often bypass the legend and instead make a direct comparison between items on the visualization in terms of their relative brightness. Evidence for this hypothesis was provided by the analysis of people’s eye movements while performing the tasks in addition to patterns of

TABLE 1: Process Explanation of the Task × Scale Type Interaction (processes in italics)

Task	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*

accuracies and response times across conditions (Breslow et al., 2009).

In this article, we further evaluate and elaborate on this general process model by embedding it in running cognitive models, whose predictions are compared to experimental findings. Specifically, we model performance of the two tasks within the ACT-R (Adaptive Control Of Thought–Rational) Version 6.0 cognitive architecture (Anderson, 2007; Anderson, Bothell, Byrne, Lebiere, & Qin, 2004). A cognitive architecture is useful for evaluating the general model because its design is informed by a large body of research on perception, attention, and cognition. Thus, the ability of a running model, operating within the constraints of a psychologically based architecture, to accurately simulate the performance of human participants bolsters our confidence in the presuppositions of the model. At the same time, the great specificity of the running model can suggest testable hypotheses that go beyond one's initial presuppositions, leading to further research that in turn may motivate refinements in the model. The specific features of ACT-R that make it suitable to model the tasks of concern to us will be discussed later.

In this article, we present empirical results replicating the Task × Scale interaction and evaluate the ability of the operational models we present to predict those results. In this context, we also consider the question of whether it is possible to create a "dual-use" scale that is effective for both identification and comparison tasks.

IS A DUAL-USE SCALE POSSIBLE?

A further goal of the present research is to consider whether it is possible for a hybrid scale that is both multicolored and ordered by brightness to incorporate the respective strengths of unordered multicolored scales and ordered brightness scales on both identification and comparison tasks. Such a scale would enable visualization designers to bypass trade-offs they are typically compelled to make between accuracy and speed on one or another of the visualization tasks. For instance, studying methods of color coding absolute and relative altitudes in school maps, Phillips (1982) concluded that brightness codes should be preferred to multicolored codes. Even though he found that brightness codes afforded inferior accuracy on absolute value judgments, they afforded superior accuracy on relative comparison tasks, which Phillips considered of greater practical importance than absolute value judgments for children using maps. However, if accuracy is one's primary concern, then multicolored scales are preferable, because they support high accuracy on both identification and comparison tasks—but at the cost of slower performance on comparisons (Breslow et al., 2009). Certainly it would be desirable to provide a brightness-ordered multicolored scale that was effective for both comparison and identification tasks in hopes of mitigating, if not eliminating, such compromises.

The possibility of providing such scales is raised by the research of Spence and his colleagues (Spence, Kutlesa, & Rose, 1999), who proposed the principle of perceptual linearity (PL) to characterize the requirements for a scale, whether monochrome or multicolored, that is useful for relative comparison tasks. It ensures that the colors in a scale lie along a curve in color space that is linear both in perceived luminosity (i.e., brightness) and in hue. PL is possible only if the scale colors' luminosity coordinates in three-dimensional color space are more highly weighted than the two coordinates defining their hues, because hues, unlike luminosity, maintain a circular ordering in color space. Thus, a monochrome brightness scale is perceptually linear, as it is represented by a straight vertical line in perceptual color space (by convention, the vertical axis represents luminosity and the horizontal axes represent hue and saturation). However, a multihue brightness scale may also be perceptually linear if brightness is more highly weighted than the hue dimensions; such a scale is represented by a vertical spiral that does not complete a revolution in horizontal space.

In this manner, Spence et al. (1999) constructed a multihue brightness scale called HSB (Hue-Saturation-Brightness). The HSB scale behaved as brightness scales typically do on a relative comparison task. Specifically, HSB and a monochrome brightness (green) scale were superior to an unordered multicolored scale in terms of both accuracy and response times. Although the multicolored HSB scale was comparable to the monochrome brightness scale in terms of the accuracy of responses, it afforded slower response times than the monochrome scale. Spence et al.'s HSB scale is tested in the current research.

Breslow, Trafton, McCurry, and Ratwani (in press) evaluated the PL hypothesis by comparing PL-ordered scales, one monochrome and one multicolored (specifically HSB), to unordered multicolored scales as well as to scales specified by their Motley algorithm. The Motley scales were ordered by brightness but otherwise maximally discriminable and unordered by hue and thus not perceptually linear. The finding that the Motley scales were as effective as the PL scales in supporting relative comparisons stands as evidence that PL is not a *necessary* condition for effective comparisons. However, PL may still be a *sufficient* condition, as the PL scales were as efficient as the Motley scales and both were faster than unordered multicolored scales.

Breslow et al. (in press) further considered whether PL might provide a guideline for creating dual-use scales effective for identification as well as for comparison tasks. Thus, they tested multicolored PL scales in the context of identification tasks, in essence testing the hypothesis that multicolored PL is a sufficient condition for effective identification. Unfortunately, their findings were generally negative, as the PL multicolored scale (again, HSB) was inferior to both the unordered multicolored scales and to the Motley scales for identification in terms of both accuracy and speed. One explanation for this may be that the hues of adjacent colors in the multicolored PL scale are too similar to provide the discriminability required to effectively match the visualization's colors to the legend. By including both multicolored and monochrome PL scales in the experiments and model simulations to be reported, we hope to determine whether similar principles apply to both and, more generally, to determine the requirements for a dual-use scale.

In the following two sections, we describe two experiments designed to assess the influence of color scale on performance of identification and comparison tasks, extending previous work primarily by the addition of a perceptually linear multicolored scale. The experiments differ in whether a gray mask is displayed between trials. Then, in the subsequent two sections, we present the computational models hypothesized to account for users' performance on each of the two tasks.

EXPERIMENT 1

The first experiment compared performance in four conditions, defined by two tasks with color-coded visualizations (identification vs. comparison) and by two scale types (brightness vs. multicolored) used to color those visualizations. The brightness scales included Spence et al.'s (1999) HSB scale, a multicolored brightness scale designed according to their PL principle.

Method

Participants. Eighty-six undergraduate psychology students from George Mason University participated in this study for partial course credit. Participants were determined to be color normal with the Pseudoisochromatic Plates Ishihara Compatible (PIPIC) 24-plate test (Good-Lite Co., Elgin, IL). All participants had normal or corrected-to-normal vision. The experiment lasted approximately 45 min. Participants were assigned randomly to one of four conditions defined by scale type (brightness or multicolored) and task (identification or comparison), resulting in 21 participants in each condition, except the multicolored-comparison condition, which had 23 participants.

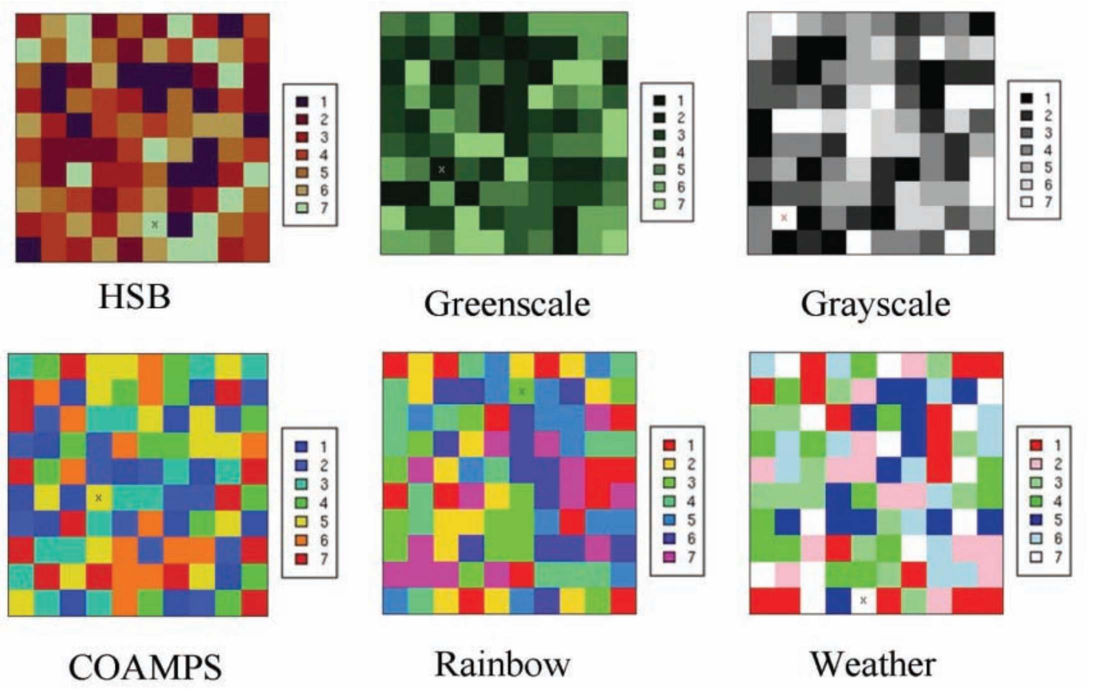

Figure 2. Experimental stimuli. HSB = Hue-Saturation-Brightness; COAMPS = Coupled Ocean/Atmosphere Mesoscale Prediction System.

Apparatus. Eye track data were collected using a Tobii 1750 operating at 60 Hz with spatial resolution of 0.5° (1,024 × 768 resolution). The graphics card used was a Nvidia Quadro FX 1400. A fixation was defined as a minimum of five consecutive eye samples (approximately 20 ms each) within 10 pixels (approximately 2° of visual angle), calculated in Euclidian distance. The target letters and the legend were defined as areas of interest.

Materials. A stimulus consisted of a 10 × 10 grid and a legend (see Figure 2). Each cell on the color grid subtended 2.54° of visual angle. 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). To the right of the grid was the legend, displaying the scale colors and their associated numbers, listed vertically downward from 1 to 7. Each color cell on the legend subtended 1.27° of visual angle and each number subtended 1.69°.

Two scale types were used, multicolored and brightness, with three instances of each type. The multicolored scales—Rainbow, COAMPS (Coupled Ocean/Atmosphere Mesoscale Prediction System; Hodur, 1997), and Weather—were not ordered by either brightness or hue. The brightness scales—Grayscale, Greenscale, and HSB—were ordered by brightness, and HSB was ordered by hue as well. The brightness scales all conformed to Spence et al.'s (1999) PL hypothesis. The Rainbow multicolored scale was constructed by using the built-in "rainbow" set of hues from R (R Development Core Team, 2007). The COAMPS scale came directly from one of the displays of the COAMPS meteorological modeling system (Hodur, 1997). The Weather scale came from the *Washington Post* daily weather map for May 2003.

Turning to the brightness scales, the Grayscale was created by varying luminance in equal steps from white to black. Both the Greenscale and HSB scales came directly from Spence et al. (1999). The Greenscale (called Brightness by Spence et al., 1999) colors were separated by approximately equal intervals on the Munsell value (brightness) dimension with hue and chroma held constant. HSB varied linearly in Munsell value (brightness), hue, and chroma (saturation), with brightness and saturation varying in opposite directions.

Values for sRGB and CIELAB of the colors in each scale may be found at http://www.nrl.navy.mil/aic/iss/pubs/color.support/supplement.pdf. All stimuli were created with the use of R (R Development Core Team, 2007). The experiment was presented with the use of E-Prime (Schneider, Eschman, & Zuccolotto, 2002).

Procedure for identification condition. On each identification task trial, an *X* appeared in one cell of the grid to mark the target color to be identified. Both the location of the target *X* and the arrangement of colors on the grid were determined randomly on each trial. Each of a scale's seven colors was the target color on 6 trials, resulting in 42 trials per scale, or a total 126 test trials for the three scales in a condition.

Participants were tested individually and were seated approximately 43 cm from the computer monitor. To minimize the requirement for visual search of the grid, the location of the target (*X*) was presented on a blank screen prior to each trial. After the participant pressed the space bar, the colored test stimulus was presented. The participant's task was to indicate the numerical value associated with the color that the *X* was on by entering the appropriate value (1 through 7) on the keypad. Response time was recorded as the duration between the appearance of the colored test stimulus, which replaced the blank screen, and the participant's response. After a response was made, the next trial started. Stimuli were presented in block-randomized order, blocked by scale. The order of the three scale blocks was randomized.

Prior to the experiment, the participants were given brief training with only three colors and a 3 × 3 grid. Next, they were introduced to the legends for the three scales, followed by the test trials. They were instructed to respond as quickly and as accurately as possible.

Procedure for comparison condition. For each of the 21 possible pairwise comparisons among the seven colors in a scale, two stimuli were generated, each one having an *X* and an *O* in different colored cells on the grid. On one of these two stimuli, the *X* had the greater value, and on the other stimulus, the *O* had the greater value, as determined by the legend numbers. The location of both targets was generated randomly, and each participant received a different random set of stimuli. Thus, a total of 42 (21 pairwise comparisons × 2) different stimuli were created for each of the three scales, for a total of 126 test trials.

Participants were tested individually and were seated approximately 43 cm from the computer monitor. They were instructed to respond as quickly and as accurately as possible. To minimize the requirement for visual search of the grid, the locations of both targets (the *X* and the *O*) were presented on a blank screen prior to the colored stimulus, which was displayed after the participant hit the space bar. The participants' task was to determine whether the numeric value associated with *X* or *O* was greater on the legend and then to respond by pressing the *z* or the slash key (labeled with an *X* or an *O,* respectively). Response time was recorded as the duration between the appearance of the colored stimulus, which replaced the blank screen, and the participant's response. After a response was made, the next trial started.

Block randomization and training were similar to those in the identification task, with necessary modifications appropriate to the comparison task.

Results

Accuracy and response time. Significant interactions between task (identification vs. comparison) and scale type (i.e., multicolored vs. brightness) were found for both accuracy, $F(1, 82) = 74.69$, MSE = .002, $p < .001$, and response time, $F(1, 82) = 34.71$, MSE = 526,053, $p < .001$. Similarly, when the six individual scales were analyzed separately, rather than by scale type, Task × Scale interactions were found for accuracy, $F(1, 82) = 75.6$, MSE = .01, $p < .001$, and response time, $F(1, 82) = 35.6$, MSE = 1,572,575, $p < .001$. In view of these interaction effects, the results for each task will be reported separately.

Identification task. Accuracy and response time data from the identification task are displayed in Figures 3 and 4, respectively. Performance on the multicolored scales was significantly more accurate, $F(1, 40) = 145.4$, MSE = .002, $p < .001$, and faster, $F(1, 40) = 24.4$, MSE = 489,603, $p < .001$, compared with the brightness scales. Mean accuracies for multicolored scales and brightness scales were .98 and .83, respectively, and mean

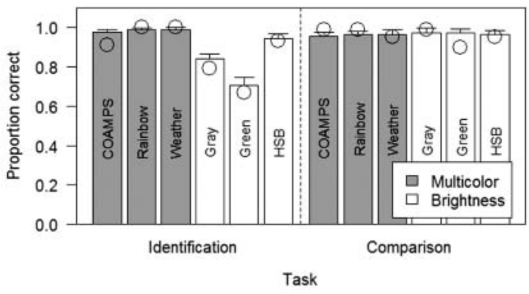

Figure 3. Response accuracy in Experiment 1. Error bars represent 95% confidence intervals. Circles represent predictions by the Adaptive Control of Thought–Rational model.

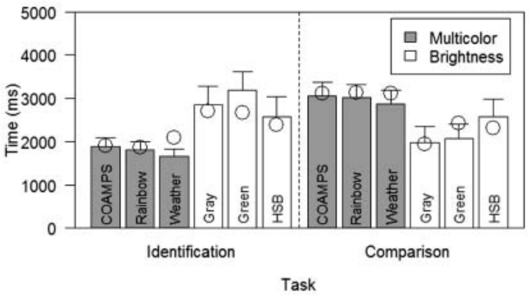

Figure 4. Response times in Experiment 1. Error bars represent 95% confidence intervals. Circles represent predictions by the Adaptive Control of Thought–Rational model.

response times were 1,785 ms and 2,851 ms, respectively. Analyses of the six individual scales also revealed significant effects for accuracy, *F*(1, 40) = 145.9, MSE = .01, *p* < .001, and response time, *F*(1, 40) = 25.5, MSE = 1,459,354, *p* < .001.

Post hoc comparisons, assessed by the Tukey honestly significant display (HSD) test, were conducted relative to the HSB brightness scale developed by Spence et al. (1999). The HSB scale afforded more accurate responses (*M* = 95%) than the other two brightness scales, Grayscale and Greenscale (*M* = 84% and *M* = 71%, respectively; *p* < .05) and comparable accuracy to the three multicolored scales (*p* > .10), although this might be attributable to a ceiling effect, as accuracy levels were all high for the multicolored and HSB scales. In contrast, response time for HSB was similar to the other brightness scales (*p* > .10) and, like them, was slower than the multicolored scales (*p* < .05).

Comparison task. Accuracy and response time data from the comparison task are displayed in Figures 3 and 4, respectively. In contrast to the identification task, performance on the brightness scales (*M* = 2,206 ms) was significantly faster, *F*(1, 42) = 11.9, MSE = 560,768, *p* < .01, than on the multicolored scales (*M* = 2,985 ms); for individual scales, *F*(1, 42) = 11.9, MSE = 1,680,405, *p* < .01. However, accuracies on the two scale types or individual scales were not significantly different (*p* > .10; *M* = .97 for brightness and *M* = .96 for multicolored). The absence of a significant effect of accuracy may be attributable to a ceiling effect, as accuracies were very high.

Tukey HSD comparisons involving the HSB scale were conducted on comparison response times. HSB supported response times intermediate between the other brightness scales, on one hand, and the multicolored scales, on the other hand, but did not differ significantly from any other scale (*p* > .10). In contrast, the other two brightness scales allowed faster responses than any of the multicolored scales (*p* < .05). Post hoc comparisons were not conducted on the accuracy data because the main effect of scale was not significant.

Eye movements. Proportions of trials on which participants fixated the legend are shown in Figure 5. Analysis by scale type revealed that participants fixated the legend considerably less often on the comparison task when using brightness scales than on any of the other three task–scale type conditions. A significant Task × Scale Type interaction effect, *F*(1, 70) = 48.7, MSE = .02, *p* < .001, characterized the proportion of trials on which participants looked at the legend. Significant main effects of task, *F*(1, 70) = 74.1, MSE = .02, *p* < .001, and scale type, *F*(1, 70) = 82.6, MSE = .02, *p* < .001, were also found. Proportion of trials with legend fixations averaged .31 in the brightness-comparison condition, .81 in the multicolored-comparison condition, .86 in the brightness-identification condition, and .89 in the multicolored-identification condition.

Analyses of the individual scales, rather than scale type, mirrored these trends, with a significant Task × Scale interaction, *F*(1, 70) = 48.8, MSE = .06, *p* < .001, and significant effects of task, *F*(1, 70) = 74.2, MSE = .06, *p* < .001, and scale, *F*(1, 70) = 82.6, MSE = .06, *p* < .001.

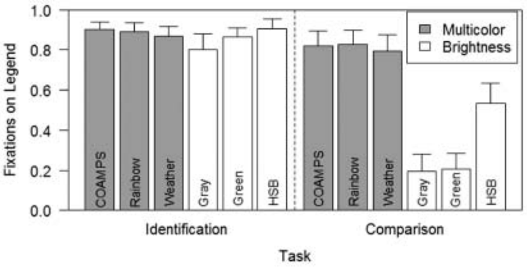

Figure 5. Proportion of trials on which the legend was fixated, Experiment 1. Error bars represent 95% confidence intervals.

Analyses of frequency, rather than proportions, of legend fixations revealed trends identical to these.

However, post hoc analyses revealed a more complicated picture, particularly with regard to the HSB scale on the comparison task, where it occupied an intermediate position between the other brightness scales and the multicolored scales. On that task, there was a significant effect of scale, $F(1, 41) = 90.5$, MSE = .09, $p < .001$, with participants referring to the legend with the HSB scale more often than with the other brightness scales and less often than with the multicolored scales (Tukey HSD comparisons, $p < .05$). In contrast, in the identification task, the effect of scale was not significant ($p > .10$).

Discussion

The results of Experiment 1 replicated previous research in demonstrating that multicolored scales support superior accuracy and speed compared with brightness scales on identification tasks, whereas brightness scales support generally superior performance to multicolored scales on comparison tasks. The sole exception was the absence of significant differences in accuracy on the comparison task. This may reflect a ceiling effect, as accuracies were high with both scale types in the comparison task. However, response time and legend fixation variables are less susceptible to ceiling or floor effects, and results for those variables conformed to predictions.

The eye movement analysis lent further support to the hypothesis that people usually solve the comparison task with brightness scales by performing a direct comparison between the targets, without reference to the legend.

The results lent mixed support to the PL hypothesis. The perceptually linear multicolored HSB scale generally displayed results intermediate between the other brightness scales and the multicolored scales in accuracy, response time, and legend fixations, although the differences were not always significant. On the identification task, HSB displayed accuracies higher than the inferior scale type, brightness, but slower response times than the superior, multicolored type. On the comparison task, the HSB results did not differ significantly in accuracy or response time from the other scales but did differ significantly from them in terms of legend fixations, being intermediate between the other brightness scales and the multicolored scales in this respect.

Many color researchers insert a gray or neutral filler between color trials to prevent the impact of previous trials on the current trial, such as color carryover effects, color preview effects, and so on (Braithwaite, Humphreys, & Hodsoll, 2003). For this reason, we replicated Experiment 1, inserting a gray mask between trials.

EXPERIMENT 2

Experiment 2 was similar to Experiment 1 except that a gray mask was displayed at the start of each trial to neutralize possible effects of the preceding trial on participants' perception of the stimuli (Braithwaite et al., 2003).

Method

Participants. Participating in this study for partial course credit were 49 undergraduate psychology students from George Mason University. Participants were determined to be color normal with the PIPIC 24-plate test (Good-Lite Co., Elgin, IL). All participants had normal or corrected-to-normal vision. The experiment lasted approximately 45 min. Participants were assigned randomly to one of four conditions defined by scale type (brightness or multicolored) and task (identification or comparison), resulting in 12 participants in each condition, except the multicolored-identification condition, which had 13 participants. Data from 1 participant in the brightness-identification condition were eliminated from the analysis, because response times were extremely high—approximately 3 times higher than the next slowest participant on one of the scales.

Procedure. The procedure was identical to that used in Experiment 1, except that a uniform gray screen was displayed for 1,000 ms at the start of each trial, followed by a stimulus of the same sort used in the previous experiment.

Results

Accuracy and response time. Significant interactions between task (identification vs. comparison) and scale type (multicolored vs. brightness) were found for both accuracy, $F(1, 43) = 55.9$, MSE = .001, $p < .001$, and response times, $F(1, 43) = 38.1$, MSE = 345,607, $p < .001$. Similarly, interactions were found when analyzing the six scales separately, rather than by scale type, for both accuracy, $F(1, 43) = 55.9$, MSE = .003, $p < .001$, and response time, $F(1, 43) = 38.1$, MSE = 1,036,822, $p < .001$. Thus, the results for each task will be reported separately.

Identification task. Accuracy and response time data from the identification task are displayed in Figures 6 and 7, respectively. Responses with the multicolored scales were more accurate, $F(1, 22) = 48.7$, MSE = .001, $p < .001$, and faster, $F(1, 22) = 26.5$, MSE = 253,039, $p < .001$, than those with the brightness scales. Mean accuracy for multicolored and brightness scale types was .97 and .86, respectively, and means response times were 1,760 ms and 2,821 ms, respectively. Analyses of the six individual scales revealed significant effects for accuracy, $F(1, 22) = 48.7$, MSE = .004, $p < .001$, and response time, $F(1, 22) = 26.5$, MSE = 759,116, $p < .001$.

Post hoc comparisons, assessed by the Tukey HSD test, were conducted relative to the HSB brightness scale. As in Experiment 1, HSB afforded more accurate responses than the other two brightness scales, Grayscale and Greenscale ($p < .05$), and comparable accuracy to the three multicolored scales ($p > .10$). However, whereas in the first experiment, response time for HSB was no better than for the other brightness scales and slower than the multicolored scales, in Experiment 2, HSB was faster than the Greenscale brightness scale only ($p < .05$) and slower than the Weather multicolored scale only ($p < .05$).

Comparison task. Accuracy and response time on the comparison task are summarized in Figures 6 and 7, respectively. Responses with the brightness scales were more accurate, $F(1, 22) = 11.4$, MSE = .0002, $p < .01$, and faster, $F(1, 22) = 14.4$, MSE = 434,822, $p < .01$, than with multicolored scales. Mean accuracies for brightness and multicolored scales were .98 and .96, respectively, whereas mean response times were 1,909 ms and 2,929 ms, respectively. Analyses of the six individual scales revealed significant effects for accuracy, $F(1, 22) = 11.4$, MSE = .001, $p < .01$, and response time, $F(1, 22) = 14.4$, MSE = 1,304,467, $p < .01$.

Tukey HSD comparisons involving the HSB scale were conducted on comparison response times. As in Experiment 1, HSB afforded response times intermediate between the other brightness scales and the multicolored scales but did not differ significantly from any other scales ($p > .10$). In contrast, the other two brightness scales allowed faster responses than any of the multicolored

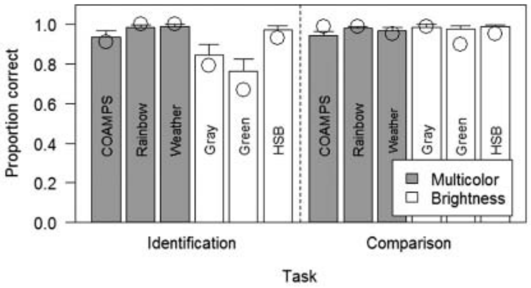

Figure 6. Accuracy in Experiment 2. Error bars represent 95% confidence intervals. Circles represent predictions by the Adaptive Control of Thought–Rational model.

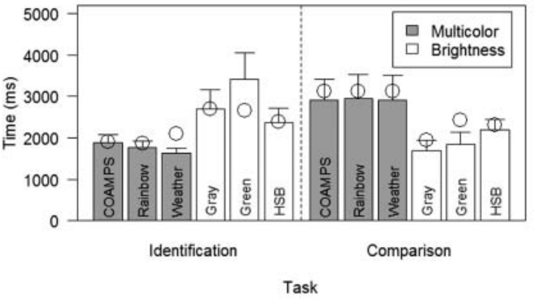

Figure 7. Response times in Experiment 2. Error bars represent 95% confidence intervals. Circles represent predictions by the Adaptive Control of Thought–Rational model.

scales ($p < .05$). In terms of accuracy, HSB, like the other brightness scales, was significantly more accurate than the COAMPS multicolored scale only ($p < .05$).

Discussion

Experiment 2 replicated the general findings of Experiment 1, revealing the Scale × Task interaction found by previous researchers (Breslow et al., 2009; Merwin & Wickens, 1993; Phillips, 1982). Most importantly, the Task × Scale interaction found in all of these experiments was unaffected by the interposition of a gray mask between trials. Indeed, Experiment 2 was more similar to previous research than was Experiment 1 in that significant accuracy differences were found between the two scale types on the comparison task in Experiment 2, in contrast with Experiment 1.

Again, the perceptually linear multicolored HSB scale generally displayed results intermediate between the other brightness scales and the multicolored scales in accuracy and response time, although the differences were not always significant. The HSB scales performed quite well on the identification task, in which it generally surpassed the inferior, brightness scales. In contrast, on the comparison task, it generally failed to surpass the inferior, multicolored scales.

We now describe the computational models hypothesized to predict the performance of participants on the identification and comparison tasks in the two experiments. The discussion is divided into two parts: the description of the cognitive architecture, followed by the description of the models.

COGNITIVE ARCHITECTURE

We modeled task performance using the ACT-R 6.0 cognitive architecture (Anderson, 2007; Anderson et al., 2004). ACT-R is a hybrid symbolic-subsymbolic production-based system consisting of a number of modules. It interfaces with the outside world through the visual, aural, motor, and vocal modules. Central processing is simulated by intentional, imaginal, temporal, and declarative modules.

ACT-R is well suited to model the visualization tasks of concern, as it represents both parallel and serial processing in visual search (Nagy, 1999; Nagy & Sanchez, 1992), specifically, parallel processing for preattentive visual localization of the next candidate object to consider and serial processing for reflective consideration of each candidate object located. However, like most other cognitive architectures, ACT-R lacks a means for representing colors precisely, for instance, by specific red, green, and blue values in RGB color space. Instead, ACT-R represents colors using a limited set of color constants (red, green, blue, etc.). What is more, ACT-R is unable to compare colors in a perceptually realistic manner. Rather, colors are compared in an all-or-none manner through the matching of color-constant symbols. However, this approach is not adequate when comparing colors that are represented more precisely. For instance, two colors with the RGB representations 255, 0, 0 and 254, 0, 0 are perceptually indistinguishable examples of red even though their representations are different.

To fill this gap, we modified the ACT-R architecture to enable it to represent colors more precisely and to compare colors both preattentively and attentively in terms of either their overall similarity or their relative brightness. To do this, we modified ACT-R's vision module to enable it to process precisely specified colors preattentively during visual search and added a new visual analysis module (VAM) to handle attentive, high-level color processing. Both modules used the same measures to determine both overall *color similarity* and the degree and direction of *brightness difference* between colors. The two modules differed in the thresholds for similarity and difference: The vision module's preattentive visual location submodule (an ACT-R "buffer") used more lenient criteria than the VAM's attentive processor.

The VAM's high-level visual processing of color is an example of what has been called "color cognition" (Derefeldt, Swartling, Berggrund, & Bodrogi, 2004). It represents an unusual case of high-level cognition for ACT-R. Whereas ACT-R generally represents higher-level cognition symbolically, cognition of color may not always be symbolic. Indeed, the VAM computes color similarity using the same formulas used in the subsymbolic visual localization submodule but with different thresholds. Similarly, visual search may be influenced by color categories, although it remains unclear whether these "categories" are symbolic or instead represent perceptual discontinuities in

TABLE 2: Responses of the Visual Analysis Module to CIE-DE2000 Color Differences

Color Difference	Response	Add to Candidates List?
< Strict threshold	Same	Yes
> Strict threshold < Loose threshold	Different	Yes
> Loose threshold	Different	No

Note. CIE-DE2000 = Commission Internationale de l'Eclairage Difference Equation 2000 (CIE, 2001).

color space (Smallman & Boynton, 1990; Yokoi & Uchikawa, 2005).

We will describe in turn the new capabilities we added to ACT-R to enable it to determine the overall similarity and the relative brightness of two colors.

Determining Color Similarity

When assessing color similarity, the VAM receives requests to compare two colors, specified either by two RGB colors in the format "COLOR-rrr-bbb-ggg" or by two visual objects whose colors are to be compared; in either case, one color-object is designated as the target and the other as the candidate. In the models to be described, visual objects, rather than colors, are input into the VAM. In ACT-R, visual objects are either objects displayed on a computer screen, such as shapes or letters, or simulations of such objects.

The measure for determining perceptual color similarity and difference is the recently created CIE-DE2000 (Commission Internationale de l'Eclairage Difference Equation 2000; CIE, 2001), which is especially well suited to measuring small color differences. VAM's responses to color difference scores relative to its two thresholds are outlined in Table 2. Strict and lenient thresholds are specified by user-modifiable parameters to ACT-R models; both thresholds are stricter than the threshold used by the vision module's preattentive visual location submodule, which is also a model parameter. As indicated in Table 2, if the score is below the strict threshold, then the module responds with the symbol *same*; if not, it responds *different.* In the case where it has responded *different*, VAM behaves differently based on the loose threshold: If the difference score is below the loose threshold, then the candidate colored object is added to the module's candidates list; however, if the score exceeds this threshold, nothing further is done. Also, a colored object passing the strict threshold is added to the candidates list. If colors, rather than colored objects, are input to the VAM, they are not added to the candidates list, because color memory span is very short (cf. later discussion of color memory). Colored objects are input into VAM in the models to be reported. The candidates list may be used later by the VAM to guess a matching color. For instance, when the preattentive color comparison mechanism fails to find a new match, a model may request that VAM guess a color. In response, VAM randomly selects one of the colored objects on its candidates list to output; or if the list is empty, it outputs *NULL* to indicate that it cannot select a color. As the candidates are objects, the vision module is then able to return attention to the location of the selected object, as location is stored as a feature of visual objects. After VAM responds to a guess request, it clears its candidates list. Also an explicit request may be made to VAM to clear its candidates list.

Determining Brightness Difference

The assessment of brightness difference between two colors proceeds as follows. We adopted as our measure of perceived brightness the L* value of the CIELAB (L*a*b*) representation of color (Fairchild & Pirrotta, 1991). The measure of brightness difference used by VAM was the signed difference between the L* values of the two colors being compared. If the absolute value of the difference exceeds the more stringent criterion, the VAM outputs *darker* or *lighter*, depending on the sign of the difference. If the absolute difference exceeds only the more lenient criterion, the VAM outputs *maybe-darker* or *maybe-lighter.* If the absolute difference does not exceed even the lenient criterion, VAM outputs *not-different.* This capability is used only in our model of the color comparison task, which uses only the lenient criterion.

In sum, our extension to ACT-R introduces five new parameters: two each for the determination of color similarity and brightness difference by VAM and one for the visual location of colors. Four of these parameters are relevant to our models. The fit between our models and the data was sensitive to the settings of these four parameters, a situation that is typical of computational modeling research. The important question is the generality of applicability of the parameters to tasks different from those reported here. As the body of research on computational modeling with ACT-R has grown, certain parameters have been found to be highly stable, whereas others need to be adjusted often. Certainly, the discovery of stable parameters is one contribution to knowledge that the modeling effort makes over time. However, even some variable parameters (e.g., corresponding to the participant's state of arousal) may be understood in terms of bodies of research in cognitive psychology. Future research with VAM and the modified ACT-R vision model will shed light on the nature of the parameters used here.

COGNITIVE MODELS

We first discuss how we modeled color memory for purposes of visual search, followed by descriptions of the models of the two tasks, and finally by a discussion of the success of the models in predicting the data from our experiments.

Color Memory

Whereas ACT-R assumes that items persist in visual short-term memory (VSTM) for several seconds (default = 3 s), the memory for colors decays more rapidly, beginning at about 100 to 200 ms (Vandenbeld & Rensink, 2003). Because the modifications to ACT-R that would be required to add an additional memory mechanism specifically for color would be extensive, we chose instead to accommodate color memory within our cognitive models. We therefore proceeded on the assumption that color memory has a span of around 200 ms, similar to that of iconic memory (Sperling, 1960). This limitation was reflected in our modeling of visual search by the requirement that the target color be reencoded following each comparison to a candidate legend color. In contrast to this, object locations (i.e., target location, legend locations already examined during search, and legend locations on the VAM's candidates list) are retained across several legend comparisons, a duration reflecting the use of VSTM.

Identification Task Model

According to the model of the identification task, people conduct a serial search of the legend to determine the legend color matching the target color in the visualization. Evidence of legend search is provided by eye tracking data in Experiment 1 and in Breslow et al. (2009). The hypothesis that legend search is serial in nature is supported by findings that response times increase with increased scale size (Breslow et al., 2009). Differences in predicted performance among the various color scales are determined by the relative discriminability of the colors in each scale, as assessed by the CIE-DE2000 color difference metric that we incorporated in the modified ACT-R architecture. According to this metric, multicolored scales are more discriminable than the HSB scale, which is more discriminable than the other, monochrome, brightness scales.

We now describe a run of the identification task model. First, the *X* appears on a blank screen, and the model locates and encodes the letter's location. Once the colored stimulus appears, the model encodes the background color of *X* in the encoded location on the grid.

Next, the model searches the legend for the color that matches the target color. In general terms, the preattentive visual location submodule selects each new legend color to consider using a parallel search, but candidate colors are selected serially in this manner and then are compared by VAM to the target color retrieved from memory. More specifically, ACT-R's visual location submodule selects a legend color whose similarity to the target color falls within its similarity threshold. It does not select the same legend color more than once. The selected legend color is then encoded, the target color is retrieved, and the model then calls on VAM to compare the two colors (really, colored objects). If the measured difference between the colors is below the strict threshold, the VAM responds *same* (see Table 2). If VAM responds that the colors are the same and this is the first legend color examined, then the selected legend color is accepted as a match: The model proceeds to locate and encode the number associated with the legend color and then to press the corresponding number key.

However, if either the model responds *different*, or if it responds *same* and the examined legend color is not the first legend color examined, then the model continues the legend search process. This differential treatment of colors that closely match the target (i.e., those deemed same), depending on whether they are examined first or not, prevents the model from always choosing the correct match. This is necessary, as the correct color will always be examined at some point in the legend search and as there is often only a single color judged to be same. If search always terminated at this point, the model would predict perfect accuracy, which is not observed. On the other hand, when there are two or three legend colors judged to be same, the model would predict accuracies of 50% or 67%, respectively, which are lower than the observed accuracies.

Serial search in the model recruits VAM's candidates list, from which a "match" is finally selected. If VAM responds *same* when comparing a legend color to the target, or if it responds *different* and the match score exceeds VAM's lenient threshold, then VAM adds the legend color object to its candidates list; otherwise, it is not added. If the legend search process has not terminated, the model reexamines and reencodes the target's color, because the color is quickly forgotten. It does not need to relocate the target's location, because that is remembered.

The examination of legend colors iterates until the visual location submodule fails to find another legend color to compare to the target (i.e., a legend color that has not yet been examined and whose similarity to the target color falls below the submodule's threshold). At that point, the model directs the VAM to guess a color. The VAM selects a colored object at random from its candidates list. If the list is empty, VAM responds *NULL* and the model selects a legend color at random. In either case, the model's attention is directed to the position of the selected colored object. The model then locates and encodes the number associated with the selected color. Finally, the model's motor module presses the corresponding number key.

Without the candidates list, it was difficult to model the serial search of the legend in a manner consistent with the empirical findings. One alternative to this approach would be to recruit the existing capability of ACT-R's visual location module to randomly select a not-yet-examined object and to iteratively select objects in this way until a colored object satisfying a color similarity threshold was found. In this way, a single threshold, rather than strict and lenient thresholds, would suffice. However, this approach failed to provide a good fit to both the response time and accuracy data, regardless of the threshold selected. Clearly, more research will be needed to elucidate the nature of legend search, which has received little attention to date.

Comparison Task Model

As the general model outlined in Table 1 handles brightness scales and multicolored scales very differently, the process models of the two tasks are likewise distinct and so are be described separately.

According to the model for brightness scales, people first attempt to directly compare the target colors with regard to relative brightness and refer to the legend only when direct comparison fails. Support for this was provided by eye tracking data in Experiment 1 and in Breslow et al. (2009), which demonstrated that people referred to the legend much less frequently in this condition than in the other task–scale type conditions. What is more, the faster responses for comparisons made with brightness scales relative to multicolored scales would be difficult to explain on the basis of legend search, given that people have a harder time searching the legend with brightness scales than with multicolored scales, as evidenced by their relatively poor performance on identification tasks.

The model for brightness scales first locates the *X* and the *O* on the blank screen and stores the letters' locations. Once the colored stimulus appears, the model examines and encodes the background colors of the two letters at their respective remembered locations. Next, the model directs the VAM to determine the relative brightness of the two colors. If the brightness difference of the colors surpasses the lenient threshold, the model executes the motor response of pushing the key of the letter with the higher value.

If the VAM is unable to determine the relative brightness difference of the colors (i.e., they are not sufficiently distinct in brightness), it searches for the legend colors matching *X* and *O*, in turn, much as in the identification task model, with iterations of the visual location submodule's parallel search followed by similarity assessments by VAM. However, in contrast with the identification model, the numbers associated with the two selected legend colors are not located. Instead, the model determines its response by determining the relative spatial position of the two colors in the legend; that is, the spatially lower color is determined to be "greater." Finally, the motor module presses the appropriate key, *X* or *O*, indicating the greater value.

The model for multicolored scales posits that participants make comparisons by searching for the target colors on the legend but that, in contrast with the identification task, the search is parallel rather than serial. Empirical support is provided by evidence that response times do not increase with increases in scale size (Breslow et al., 2009). Breslow et al. (2009) also note that response times on comparison tasks are far shorter than would be expected if people performed the same sort of search as they do in identification tasks, repeated twice in the comparison tasks.

The model for multicolored scales, like that for brightness scales, first locates the *X* and the *O* on the blank screen, storing the letters' locations and, once the colored stimulus appears, examines the background colors of the two letters. Then it departs from the brightness model by immediately searching the legend for each of the two target colors in turn. This visual search differs from the search used in the identification task model in that it employs only the initial parallel search of the visual location submodule, without the subsequent use of the VAM. Thus, there is a single iteration of visual location search for each of the two target colors. Even though such a search process may appear imprecise, it results in high accuracies that provide a good match to the data. Finally, the multicolored model terminates by comparing the spatial locations of the two legend colors and responding on the basis of that comparison.

Model Fit

Experiment 1. Figures 3 and 4 show that the predictions of the identification and comparison models fit the data fairly well. Model data points were means of 10 runs of the model, and matches to the data were calculated on a per-scale (rather than scale type) basis, considering the six scales separately. The identification task model's predictions matched the data closely: Root mean square deviation (RMSD)/root mean square of successive difference (RMSSD) measures of fit (Schunn & Wallach, 2005) were .04/4.24 for accuracy and 292.20/2.31 for response times, and r^2 was .96 for accuracy and .86 for response times.

The comparison task model predictions also matched the data fairly closely: RMSD/RMSSD was .03/3.73 for accuracy and 211.9/1.23 for response times; the r^2 for response times was .83. We did not assess r^2 for accuracy predictions, because there was no significant effect of accuracy in the comparison task data and because r^2 is not a good measure of model fit when the data are in a very narrow range (Schunn & Wallach, 2005).

Experiment 2. Figures 6 and 7 also show that the predictions of the identification and comparison task models fit the data quite well. RMSD/RMSSD measures of fit (Schunn & Wallach, 2005) on the identification task were .05/3.11 for accuracy and 356.3/3.57 for response times, with r^2 of .98 for accuracy and .78 for response times. RMSD/RMSSD measures of fit on the comparison task were .04/8.49 for accuracy and 295.0/2.16 for response times, with r^2 of .93 for response times. Again, we did not determine r^2 for accuracy even though, unlike in Experiment 1, there was a significant effect of accuracy on the comparison task, because the data varied within a very narrow range (Schunn & Wallach, 2005). The appropriate measure of model fit in this situation, RMSSD, was very good.

Overall. A comparison of the model–data fits in Experiment 1, shown in Figures 3 and 4, with those of Experiment 2, shown in Figures 6 and 7, reveals that few deviations from model prediction were found in both experiments. The most striking consistent deviation in accuracy fit was observed for Greenscale, whereas the only consistent deviations in response time fit were observed for Greenscale and for Weather in the identification task only. These deviations may reflect limitations in the color difference measures, which are continually being improved (Fairchild & Pirrotta, 1991; Sharma, Wu, & Dalal, 2005). In general, there were many more deviations, though usually small, from the accuracy predictions than from the response time predictions. The variable significance of accuracy effects in Experiment 1 and the small range of variation in both experiments suggest that most of the deviations in accuracy predictions reflect measurement error.

Modeling the HSB scale. Turning to the HSB scale, an examination of Figures 3 and 4 and Figures 6 and 7 reveals that the models predicted the behavior of the HSB scale well, with the exception of small deviations in accuracy in Experiment 2. In the case of the comparison tasks in both experiments, the behavior of HSB is well predicted when it is treated as a brightness scale but not when HSB is treated as a multicolored scale (not shown). Most notably, the predicted response time of HSB is much higher than observed when it is treated as a multicolored scale. This provides some support for our classification of HSB as a brightness scale and the consequent application of the brightness scale comparison model, which attempts to compare the target colors directly on the visualization before consulting the legend.

GENERAL DISCUSSION

The findings largely replicated previous evidence (Breslow et al., 2009; Merwin & Wickens, 1993; Phillips, 1982) concerning the Color Scale × Task interaction: Unordered multicolored scales are superior to brightness-ordered scales for identification tasks, whereas the reverse is true for relative comparison tasks. The sole exception was the absence of a difference in accuracy between the two scale types on the comparison task in Experiment 1, presumably because of a ceiling effect. The accuracies found by Breslow et al. (2009), using similar procedures, were already high for both conditions. That both accuracy and response time were higher in Experiment 1 relative to their study suggests that the current participants traded speed for accuracy. In contrast, in Experiment 2, brightness scales afforded superior comparison accuracies to multicolored scales, although pairwise comparisons revealed that the effect was significant only relative to the COAMPS multicolored scale. Aside from this small difference, results were the same for both experiments. The use of a gray mask, then, appeared to be immaterial to the Scale × Task interaction.

Dual-Use Scales

The present experiment lent mixed support for Spence et al.'s (1999) PL hypothesis. Surprisingly, on the task most similar to ones those authors used, the relative comparison task, we found less support for the hypothesis. Whereas the perceptually linear HSB scale assumed an intermediate position between the other brightness scales and the multicolored scales in terms of both accuracy and response time, the HSB scale often did not differ significantly from any other scale on either measure. This is somewhat surprising, given that the HSB scale and one of the other two brightness scales we used (Greenscale) were used by Spence et al. They found the HSB scale to be slower than but equally as accurate as the Greenscale as well as faster as and more accurate than a multicolored scale on a relative comparison task. The difference in findings may be attributable to the greater complexity of the stimuli they used: Their stimuli were surfaces generated by mathematical functions and then distorted in one of several ways.

Although the PL principle was not intended for identification tasks, the HSB scale showed some promise in supporting those tasks, as it afforded the same benefits in terms of accuracy as the multicolored scales and, like them, was more accurate than the other brightness scales. Again, the HSB scale was intermediate in response times between the other brightness scales and the multicolored scales, with significance of differences from these latter differing across the two experiments.

The evaluations of the HSB scale on identification and relative comparison tasks that have been conducted to date in the present research, Breslow et al. (in press) and Spence et al. (1999), may be summarized as follows. First, HSB never afforded better performance in terms of accuracy or response time to the "superior" scale type for the respective tasks (i.e., brightness scales for relative comparison tasks or multicolored scales for identification tasks). However, on a maxima-minima task, Spence et al. (1999) found that the HSB scale enabled response times faster relative to the superior brightness scales and, as usual, relative to the inferior multicolored scales. HSB also afforded accuracy equivalent to the brightness scales and superior to the multicolored scales on this task.

Second, the HSB scale never afforded worse performance than the inferior scale type for either task. Third, HSB often enabled performance equivalent to the superior scale type and better than the inferior scale type in terms of accuracy and response time, although there are exceptions to each of these generalizations on each task. Fourth, the HSB scale consistently supported high response accuracy.

On the basis of these findings, the following recommendations may be offered to visualization designers desiring a color scale useful for both identification and relative comparison tasks. When ensuring high accuracy on both tasks is a priority, either a PL multicolored scale or an unordered multicolored scale would be a good choice, as both support high accuracies on both tasks. Furthermore, if speed of relative comparisons is also a priority, then a PL multicolored scale should be preferred, whereas if rapid identification is a higher priority, then unordered multicolored scales should be preferred. Of course, these conclusions must necessarily be tentative, pending replication with a greater diversity of scales, tasks, and visualizations. In particular, only one multicolored PL scale (HSB) and two monochrome PL scales (Greenscale and Grayscale) are included in the factorial experiments conducted to date.

Finally, the introduction of the Motley algorithm by Breslow et al. (in press) places these conclusions in a different light. The scales whose specifications were generated by Motley were multicolored and ordered by brightness, like the PL HSB scale, but unlike HSB, the Motley scales were *unordered* by hue, as they were designed to be highly discriminable by hue. This discriminability was probably the source of their superior behavior on identification tasks. At the same time, the monotonic exponential function defining the brightness of the Motley scale colors yielded comparison performance comparable to HSB and to a monochrome brightness-ordered scale. The success of the Motley scales, which violate the PL principle, constitutes evidence that PL is not a necessary property of scales supporting relative comparison tasks. Although the present research demonstrates that the PL principle is not a sufficient condition either, it nevertheless suggests that multicolored PL scales may be useful in certain applied contexts.

Computational Models

The ACT-R models of the two tasks represent a computational elaboration of the general process model proposed by Breslow et al. (2009), outlined in Table 1. The experimental results matched the predictions of the ACT-R models quite closely, thus supporting the general process model. On the identification task, the model predicts that people conduct a serial search of the legend, whose speed is largely a function of the discriminability of the colors in the scale. Because multicolored scales are typically more discriminable than brightness scales, the search for legend colors is faster with multicolored scales than with brightness scales (Nagy, 1999; Nagy & Sanchez, 1992).

The models' handling of short-term memory is an example of how cognitive modeling can suggest hypotheses for further research. Our models assume that the target color(s) must be reencoded frequently. Although this assumption is supported by findings that color memory decays rapidly (Vandenbeld & Rensink, 2003), it bears further empirical evaluation in the context of visualization processing. In contrast to color memory, memory for spatial locations is assumed to persist across several legend comparisons, consistent with ACT-R modeling of VSTM, which is based on previous empirical findings (Ratwani & Trafton, 2008).

Turning to the color comparison task, very different solutions are adopted for multicolored and brightness scales. For multicolored scales, the model demonstrates that a fast parallel search of the legend is sufficient to produce highly accurate responses even though parallel search is less precise than serial search. Whereas in the identification task, participants must attempt to locate the precisely matching legend color to identify its associated numerical value, in the comparison task, they need only locate the legend colors with sufficient precision to compare their relative spatial positions (Breslow et al., 2009).

Whereas previous research (Nagy, 1999; Nagy & Sanchez, 1992) has explained the use of serial versus parallel search for colors in terms of characteristics of the stimuli, the present research suggests that task demands can also play a role in determining the sort of visual search adopted.

Finally, the hypothesized strategy for brightness scales of directly comparing the target colors with infrequent reference to the legend enabled the model to predict the experimental findings. Empirical support was provided by eye tracking data, which demonstrated that people refer to the legend much less frequently in this condition than in the other conditions, as well as by accuracy and response time patterns across the tasks (Breslow et al., 2009).

Finally, our modeling of these tasks highlighted needed improvements in the ACT-R architecture, some of which we implemented. Improvements that we implemented included the ability to represent the full gamut of colors that may be displayed on a display device and to make perceptually realistic color comparisons, including overall difference and brightness difference. Although we argued that ACT-R would benefit from a more psychologically realistic model of color memory, we did not modify the architecture for this purpose but instead designed our models to simulate color memory.

CONCLUSION

In sum, we have provided support for a process model of how and whether people use a legend when using color-coded visualizations for two types of task—namely, identification and relative comparison tasks. We have also explored the necessary and sufficient conditions for a color scale that is useful for both types of task. As the primary focus of the current research has been on the effect of color scale choice on the use of legends in visualizations, our experimental procedures minimized visual search and other demands on processing the visualization itself. Our conclusions should therefore be tested on real-world instances of color-coded visualizations, where those processing demands have not been reduced as they have in the experiments reported here.

ACKNOWLEDGMENTS

This work was supported by the U.S. Office of Naval Research to J. Gregory Trafton: Grant numbers N0001405WX2011, N0001405WX30020. The views and conclusions contained in this document should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Navy. We wish to thank J. Malcolm McCurry for programming the experiments and Andrew Blumenfeld and Kristina O'Connell for running the experiments.

REFERENCES

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060.

Braithwaite, J. J., Humphreys, G. W., & Hodsoll, J. (2003). Color grouping in space and time: Evidence from negative color-based carryover effects in preview search. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 758–778.

Breslow, L. A., Trafton, J. G., McCurry, J. M., & Ratwani, R. M. (in press). An algorithm for generating color scales for both categorical and ordinal coding. *Color Research and Application*.

Breslow, L. A., Trafton, J. G., & Ratwani, R. M. (2009). A perceptual process approach to selecting color scales for complex visualizations. *Journal of Experimental Psychology: Applied*, *15*, 25–34.

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

CIE. (2001). *Improvements to industrial colour-difference evaluation* (Publication CIE 142-2001). Vienna: Commission Internationale de l'Eclairage.

Derefeldt, G., Swartling, T., Berggrund, U., & Bodrogi, P. (2004). Cognitive color. *Color Research and Application*, *29*, 7–19.

Fairchild, M. D., & Pirrotta, E. (1991). Predicting the lightness of chromatic object colors using CIELAB. *Color Research and Application*, *16*, 385–393.

Hodur, R. M. (1997). Coupled Ocean/Atmosphere Mesoscale Prediction System. *Monthly Weather Review*, *125*, 1414–1430.

Hoffman, R. R., Detweiler, M. A., Lipton, K., & Conway, J. A. (1993). Considerations in the use of color in meteorological displays. *Weather and Forecasting*, *8*, 505–518.

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

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

Nagy, A. L., & Sanchez, R. 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.

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

Ratwani, R. M., & Trafton, J. G. (2008). Spatial memory guides task resumption. *Visual Cognition*, *16*, 1001–1010.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools.

Schunn, C. D., & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), *Psychologie der Kognition: Reden und Vorträge anlässlich der Emeritierung von Werner Tack* (pp. 115–154). Saarbrueken, Germany: University of Saarland Press.

Sharma, G., Wu, W., & Dalal, E. N. (2005). The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research and Application*, *30*, 21–30.

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.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–30.

Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (2000). Turning pictures into numbers: Extracting and generating information from complex visualizations. *International Journal of Human Computer Studies*, *53*, 827–850.

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.), *Diagramatic representation and inference* (pp. 206–220). Berlin: Springer-Verlag.

Trafton, J. G., Trickett, S. B., & Mintz, F. E. (2005). Connecting internal and external representations: Spatial transformations of scientific visualizations. *Foundations of Science*, *10*, 89–106.

Trickett, S. B., & Trafton, J. G. (2007). *The use of spatial cognition in graph interpretation.* In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1563–1568). Austin, TX: Cognitive Science Society. Retrieved April 15, 2009, from http://www.cogsci.rpi.edu/csjarchive/proceedings/2007/docs/p1563.pdf

Vandenbeld, L. A., & Rensink, R. A. (2003). The decay characteristics of size, color, and shape information in visual short-term memory. *Journal of Vision*, *3*, 682.

Yokoi, K., & Uchikawa, K. (2005). Color category influences heterogeneous visual search for color. *Optical Society of America A*, *22*, 2309–2317.

Leonard A. Breslow is a cognitive scientist at the Naval Research Laboratory in Washington, D.C. He received his PhD in psychology from the University of California, Berkeley, in 1981.

Raj M. Ratwani is a postdoctoral fellow at the Naval Research Laboratory in Washington, D.C. He received his PhD in psychology from George Mason University in 2008.

J. Gregory Trafton is a cognitive scientist at the Naval Research Laboratory in Washington, D.C. He received his PhD in psychology from Princeton University in 1994.

Date received: October 31, 2008
Date accepted: April 21, 2009