An Algorithm for Generating Color Scales for Both Categorical and Ordinal Coding

Leonard A. Breslow,[1]* J. Gregory Trafton,[1] J. Malcolm McCurry,[1] Raj M. Ratwani[1,2]

[1]Naval Research Laboratory, Code 5515, 4555 Overlook Ave., SW, Washington, DC 20375

[2]George Mason University, Fairfax, VA

Received 21 July 2008; revised 17 November 2008; accepted 5 December 2008

Abstract: Previous research has shown multihue scales to be well-suited to code categorical features and shown lightness scales to be well-suited to code ordinal quantities. We introduce an algorithm, Motley, that produces color scales varying in both hue and lightness, intended to be effective for both categorical and ordinal coding, allowing users to determine both absolute and relative quantities efficiently and accurately. The algorithm first determines the lightnesses of scale colors to maximize perceived lightness differences and establish the lightness ordering, generating separate search spaces for each scale position. It then selects hues by heuristic search to maximize the discriminability of the scale. It produces scales that are ordered with respect to lightness but unordered with respect to hue and thus more discriminable than typical multihue lightness scales. In an experimental evaluation on human subjects, Motley's scales enabled accurate judgments of relative quantity, with response times superior to unordered multihue scales and comparable to ordered lightness scales, and enabled accuracy and speed of judgments of absolute quantity superior to lightness scales and comparable to multihue scales. †Published 2009 Wiley Periodicals, Inc. Col Res Appl, 35, 18–28, 2010; Published online 17 November 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/col.20559

Key words: color coding; color order systems; color categorization; visualizations; displays; computer graphics

INTRODUCTION

Color scales and lightness scales are commonly used to code graphic visualizations used in meteorology, cartography, radiology, economics, and other applications. Multicolored scales are effective in coding categorical features,[1] as are ordered lightness scales in coding ordinal quantities.[2] In certain applications it would be desirable to use a scale that was effective for both categorical and ordinal coding, for example, to be able to make both relative comparisons between the temperatures of two regions on a map, as well as to determine the temperature of a single region.[3] Unfortunately, there is evidence that multicolored scales are not well-suited for ordinal coding and lightness scales are not well-suited for categorical coding.[4–6] This has been explained in terms of multicolored scales lacking adequate ordering cues and lightness scales lacking adequate discriminability.[6] Figure 1 gives examples of multicolored scales (Rainbow and Weather) and lightness scales [Grayscale and HSB (hue-saturation-brightness)].

The goal of this research is to begin to determine the requirements of a scale that can be used effectively for both categorical and ordinal coding, allowing users to make both absolute and relative quantitative judgments. The possibility of making a dual-use multicolored lightness scale is suggested by evidence that color variation need not interfere with lightness coding,[2,7,8] nor lightness variation with color coding.[7,9]

The desirability of providing dual-use functionality of this sort is apparent upon reflection on the tasks performed using a display coded by an ordinal scale. Normally, one would like to make relative quantitative judgments easily using the display, such as binary comparisons (which of these two locations is warmer?), maxima/minima (which location is coldest?), and spatial trends (temperatures get hotter as one goes south). However, one

*Correspondence to: Len Breslow (e-mail: len.breslow@nrl.navy.mil).

Contract grant sponsor: US Office of Naval Research; contract grant numbers: N0001405WX2011, N0001405WX30020.

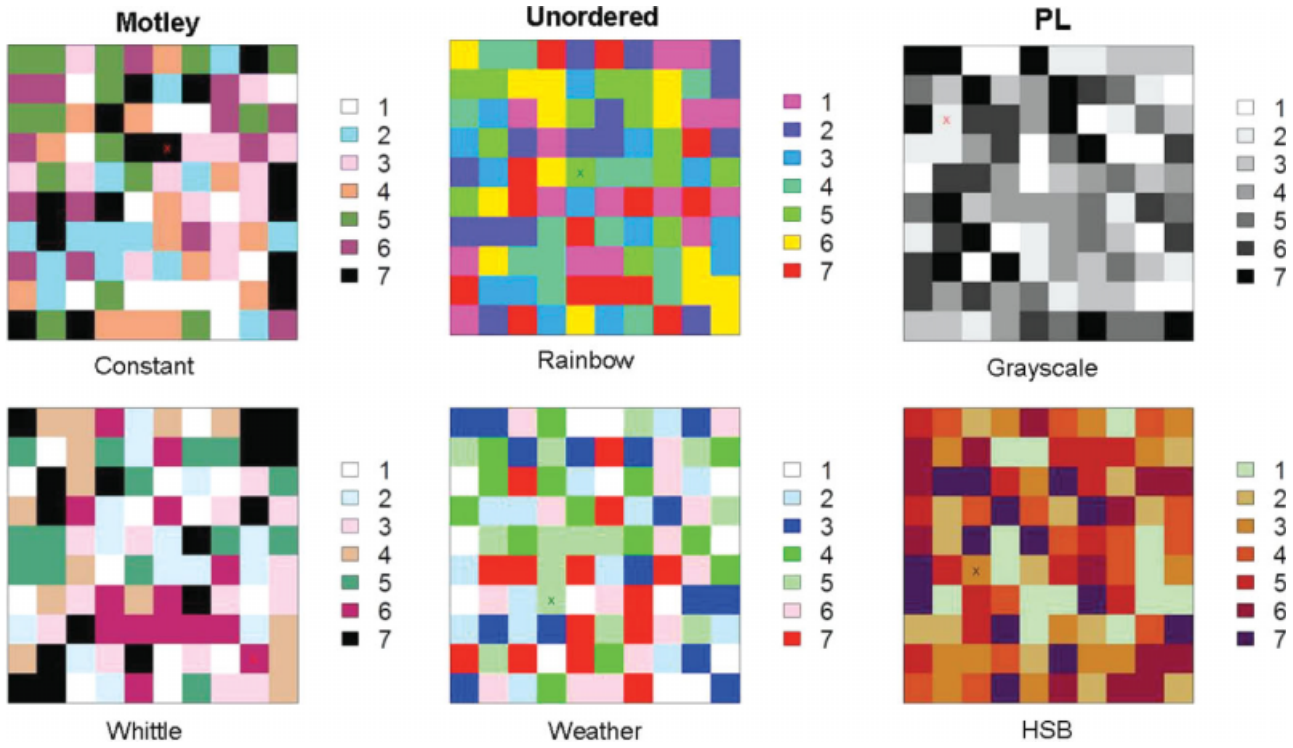

FIG. 1. Sample experimental stimuli. The printed colors necessarily differ somewhat from the colors on our monitor. Also, the RGB values we used are likely to appear slightly differently on another monitor. PL, perceptually linear; HSB, hue-saturation-brightness.

would also like to identify reliably and rapidly the absolute value of a particular display location, by matching its color to the legend. Such values may be specific numbers, but more often are ranges (e.g., 20°–30°) or ordinal categories (e.g., "somewhat agree"). To support such identifications, the ordinal scale would also need to have the discriminability of a categorical scale.

Unfortunately, the choice of a color scale to support both relative comparison and absolute value identification typically entails compromises. For example, in research on the representation of altitude in school atlases, Phillips[6] concluded that although lightness scales produced lower accuracies than multicolored scales for judging absolute heights, lightness scales are preferable, owing to their superiority for relative height judgments, "as relative height is more important than absolute height for children using atlases" (p. 1143). On the other hand, if accuracy on both tasks is one's primary concern, then multicolored scales would be preferred, even though they afford less efficient relative comparisons than lightness scales[4] (see also the research to be reported). Our goal is to produce a dual-use scale to obviate such compromises.*

We will evaluate the hypothesized requirements of a dual-use scale by implementing them in a computer algorithm, called Motley, designed to generate multihue lightness scales. Other color-scale selection tools and color-scale generation algorithms have been introduced.[2,11–13] Some tools provide an interactive environment for manual scale generation,[14,15] or a task-analytic retrieval from a database of color scales, guided by general principles supported in the literature.[16,17] These will not be considered further. Algorithms have been proposed to generate categorical scales, typically by heuristic search through a color space.[11,18] While we are aware of no algorithms for generating ordinal scales, various systematic approaches to generating them have been proposed. Most of these define guidelines for tracing a curve in a color space, along which the colors are selected.[2] We are aware of no algorithm or systematic approach for generating dual-use scales. In general terms, our approach is to make scales that are ordered by lightness and that are maximally discriminable (within the constraints imposed on lightness as well as saturation) in hue.

In what follows, we provide the motivations for the various features of the Motley algorithm, describe the algorithm, and then empirically evaluate two scales generated by Motley with human participants. In the evaluation, the Motley scales will be compared to traditionally-used lightness and multihue scales. We will attempt to replicate previous finding[4–6] with regard to the traditional scales: that is, the finding that multihue scales are superior to lightness scales for categorical coding and that lightness scales are superior for ordinal coding. Most importantly, the evaluation will determine whether the

*One solution for absolute judgments is to dispense with either colors or a legend by including numerical values directly on the display, supplemented perhaps by lines, such as contour lines, isothermal lines, etc. These are more effective than lightness coding for absolute judgments, but add clutter to the display and are inferior to lightness codes for relative comparisons.[4,10]

Motley scales are as effective as these traditional scales on their respective dominant coding task.

ORDINAL SCALES

Lightness has been found to be particularly well-suited for coding ordinal scales, because lightness provides a perceptually-based ordering.[2] Scales varying in saturation are also sometimes used, but are apparently not as effective as lightness scales in representing quantity.[8] However, hue does not lend itself to a perceptual ordering in the way lightness does. Some multicolored ordinal scales, such as ones representing the color changes undergone by a progressively heated object, rely on a symbolic rather than a perceptual ordering. Bipolar scales with two colors varying in lightness or saturation are most appropriate for scales having a central zero point.[19] The most commonly used ordered multicolor scale, the rainbow scale, suffers from the fact that the central yellows and greens in the scale are perceived as overly light;[20,21] thus the lightness ordering conflicts with the color ordering (See Rainbow scale in Fig. 1). While a rainbow spectrum represents a physical ordering of colors, it does not represent a perceptual ordering and, so, many experts advise against their use for coding quantitative data in most cases.[22]

Based on the examination of eye movements, Breslow *et al*.[4] suggested how relative comparisons are performed. They argued that quantitative comparisons are performed on lightness-coded visualizations by means of direct comparisons between locations on the visualization to determine which is darker or lighter. Consequently, the legend is often not consulted. In contrast, relative comparisons with multicolor-coded visualizations were found to involve the more laborious process of searching for each color on the legend and then comparing their respective legend values or positions. Thus, relative order is best represented by the graded lightness of colors in the scale and the optimization of a lightness scale should be directed towards facilitating the direct comparison of colors to determine easily which is darker/lighter, rather than facilitating visual search for colors.

Even if colors are not as effective as lightness for representing quantity, the question remains as to whether color variation aides or impedes the functionality of lightness scales. Some researchers have found that variation in color interferes with the lightness coding of quantity,[23] while others have found color facilitates lightness coding.[2,8] The answer to this question will be important for determining whether a dual-use multicolored lightness scale is possible.

CATEGORICAL SCALES

The opposite situation exists for categorical scales. Color has been found to be particularly well suited to categorical coding, superior to many other attributes including lightness.[1,8] Unless the code is memorized (and it's usually not[24,25]) or supported by familiar symbolic relations (e.g., blue = ocean), the user of a categorical code must look up the colors matching those in the display in an accompanying legend, thus performing visual search. Large perceptual differences among the colors have been found to be critical to efficient visual search for colors.[26,27] Indeed, in some conditions, increased color differences enable people to shift from a serial search to a parallel search[28,29] whose speed is largely insensitive to the number of different colors. A similar serial-to-parallel shift characterizes search for matching luminosities, but much larger differences are needed to support a parallel search for luminosities.[28] In sum, the optimization of a color scale for use in categorization depends upon the facilitation of visual search for the scale colors, which in turn depends upon the discriminability of the colors.

While scales of highly-discriminable hues are well-suited to categorical coding, the addition of lightness differences need not impede their effectiveness, since lightness variation does not interfere with the visual search for colors.[9,30] Thus, lightness variation added to a multihue scale should not interfere with its effectiveness as a dual-use scale.

GENERATING ORDINAL SCALES

Most systematic approaches to generating ordinal color scales involve tracing a line or curve in a color space.[2,12,31] This seems reasonable since most ordinal scales are defined by a linear trend in luminosity and since color spaces generally include a dimension to represent luminosity. By convention the luminosity dimension is usually the vertical axis, while the two horizontal axes represent hue. Thus, a vertical line may be used for selecting colors in a single-hue scale with progressive luminosities, while an upward spiral turns this into a multihue scale. Also, when the color space is perceptually uniform, equal vertical distances between colors on the curve represent equal perceived lightness differences, allowing the scale to map accurately to the quantities being represented.

For the purposes of creating a dual-use scale, we are concerned with the possibility of adding multicoloration to a lightness scale. In an effort to reconcile conflicting evidence as to whether hue variation enhances or interferes with the representation of quantity in a linear-luminosity scale, Spence *et al*.[2] introduced the hypothesis that an effective ordinal scale must be perceptually linear. Specifically, "... for a coding assignment to be PL, 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). Perceptual linearity is only possible if luminosity is more highly weighted than the two hue dimensions.

Spence *et al*. offered empirical support for the superiority of perceptually-linear (PL) scales over nonlinear

multicolored scales with human subjects tested on quantitative tasks. The authors' PL scales consisted of colors equidistant on vertical curves in the perceptually-uniform Munsell and CIELUV color spaces. Importantly, they provided evidence that a multicolored PL scale (specifically, the HSB scale in Fig. 1) is generally as effective as a monochrome PL scale in supporting quantitative tasks. While participants performed relative comparisons more slowly on the multicolored PL scale than on the monochrome scale, they were faster on the multicolored scale than the monochrome scale on a maximum/minimum task. The two PL scales did not differ in accuracy on either task. These findings suggest that multicolored lightness scales conforming to the PL principle will be effective for relative comparison tasks.

However the Perceptual-linearity hypothesis, like other curve-tracing approaches to scale construction, would appear to bode ill for the possibility of creating a multihue scale that is effective for categorical identification as well as relative comparison. Effective categorical coding depends on the discriminability of the colors in a scale,[26,28] but adjacent colors on a curve are similar to each other in both hue and luminosity (for examples, see HSB and Gray scales in Fig. 1). What is more, linearity can defeat the serial-to-parallel shift in search strategy that otherwise results from increased discriminability, and so nonlinearity is preferable for categorical codes, in contrast to ordinal codes.[32,33] The low discriminability of adjacent colors may account for the poor performance afforded by lightness scales on categorical identification tasks.[5,6] The experiment reported here evaluates an alternative to the Perceptual-linearity hypothesis; we hypothesize that a scale with ordered luminosities but colors that are otherwise maximally discriminable and unordered can be effective for both absolute and relative quantification tasks.

GENERATING CATEGORICAL SCALES

The chief requirement for an effective categorical scale is that the colors be highly discriminable. If a perceptually uniform color space is used, then a geometric approach may be applied to selecting a highly-discriminable set of colors. Thus, Healey[34] described a method for generating color scales by drawing the largest possible circle on the hue dimensions (in this case, uv in CIELUV space) within the gamut of the display device, while keep luminosity constant. The circle is then subdivided into the desired number of equal-sized arcs and rotated to satisfy other considerations such as color category memberships. Experimental subjects identified targets rapidly with scales up to seven colors in size.

In the case of nonuniform color spaces, distances must be computed independently for each pair of colors considered and so the only way to locate the most discriminable color set is by performing a search through the color space. This search space is very large and an exhaustive search of all possible scales within this space is NP-complete.[11] Thus, a heuristic search, rather than an exhaustive search, is typically performed. The first such algorithm was proposed by Carter and Carter.[18,35] Later, Campadelli *et al.*[36] proposed a neural network algorithm for selecting a high-contrast set of colors. This algorithm required parameter tuning and converged on a solution only 90% of the time. Campadelli *et al.*[11,37] then proposed an algorithm that did not suffer from those limitations and the color scales it produced were found to be superior to those produced by Carter and Carter's algorithm in terms of the sizes of the color differences within the scales. No empirical evaluation was conducted with human users.

To select a set of k colors V_k, Campadelli *et al.*'s[11] algorithm first randomly selects an initial set of k colors from the search space of n colors. Then, for each pair of colors i and j, such that $i \in V_k$ and $j \notin V_k$, evaluate

$$A = \sum_{l \neq i, l \neq j} \frac{1}{D_{il}^{\alpha}} \tag{1}$$

$$B = \sum_{l \neq j, l \neq i} \frac{1}{D_{jl}^{\alpha}} \tag{2}$$

where D_{ab} is the distance between colors a and b. If $A > B$, then substitute j for i in V_k. Repeat for each $i \in V_k$ and $j \notin V_k$. The final V_k is output.

The algorithm runs in time polynomial with respect to the size of the input n. α is set to 90. The algorithm is repeated 10 times, each with a different randomly-selected initial set, and the resulting color set with the greatest minimum distance among all the possible pairs of its colors is selected as the output.

PROPOSAL FOR GENERATING HYBRID ORDINAL-CATEGORICAL SCALES

We sought to compute color scales ordered by k levels of lightness. Within each lightness level, colors are selected to be maximally distinct from colors in other levels.

We adapted Campadelli *et al.*'s[11] algorithm as the basis for generating a scale that is maximally discriminable within the constraints imposed by a lightness ordering. Whereas their algorithm considers each color as a candidate for each position in the scale, ours assigns a separate search space to each scale position. Each position's search space is defined by a precomputed lightness and saturation. Thus, before search, the algorithm sorts the colors in the larger color space into the separate search spaces for each scale position; many colors do not qualify for inclusion in any of the spaces.

The main hurdle we encountered was to find a way to assign lightness values to the scale colors such that all colors would be clearly discriminable with regard to lightness. Luminosity was computed according to Fairchild and Pirrotta's[38] L^{**} adjustment to the L^{*} metric, designed to counteract the Helmholtz-Kohlrausch effect whereby

color affects perceived luminosity. Surprisingly, we found that equal intervals of L^{**} were not perceptually equal in the context of the highly discriminable sets of color our algorithm produced. We found that applying Weber's function to L^{**} also failed to produce the desired results. Finally, we found that power functions produced the best lightness discriminability and adopted a function found by Whittle[39] to characterize perceived lightness in cases where the "crispening effect" of background color on perceived lightness does not apply. The crispening effect is prevented in the context of the experiment to be reported, as in many practical applications, because colors within the visualizations have multicolored backgrounds and legend colors have thin black outlines.[39] We believe that other power functions[40] could serve equally well for our algorithm and that more research will be needed to determine the optimal luminosity function. In addition, future research will probably produce further improvements to the L^* luminosity measure in addition to that provided by the L^{**} adjustment.[41]

A further question was whether differences in lightness should be supplemented by differences in saturation. Levkowitz and Herman[23] recommended that saturation be either directly or inversely related to luminosity. However, since saturation differences may interfere with the perception of lightness differences,[38,42] we tested two variants of the algorithm, one in which saturation varied, following Whittle's function, in inverse relation to lightness, and another in which saturation was held constant. Further, as we found that high saturations tended to obscure luminosity differences,[42] we did not use highly-saturated colors.

The Motley Algorithm

The Motley algorithm may be outlined as follows. Before the execution of the algorithm, the space of colors must be generated. The algorithm proper consists of two major steps. In Step 1, the search space for each position in the color scale is generated. This step is primarily concerned with ensuring that the resultant scale colors are ordered and discriminable by lightness, but also controls saturation. Step 2 consists of a search through these color spaces and is primarily intended to maximize—within the constraints on each search space's luminosity and saturation—the color discriminability of the scale.

Step 0. Constructing the color space. To increase the efficiency of the algorithm (i.e., Steps 1 and 2), the color space is computed in advance and cached. The colors are defined by a modified CIELAB representation,[38] with L^{**}, a^*, and b^* dimensions divided into 50 equal-sized intervals.

Step 1. Generating the search spaces for each position. k search spaces are generated for each of the k ordinal positions in a scale consisting of k colors. Each search space is defined by a distinct lightness, and sometimes by a distinct saturation as well, as described in the following paragraphs. Colors at the ends of the scale may be prespecified, in which case the specified color is the only color in its search space. In the work reported here, $k = 7$ ordered positions, with position 1 set to black and position 7 is set to white.

For each position, a target lightness and target saturation are computed. The target lightness L is determined by Whittle's[39] function:

$$L = 5.27L^{**0.41} - 2.66 \tag{3}$$

The target saturation defining each search space is constrained in one of two ways in the following variants of the algorithm:

a. Whittle variant: Saturations are determined by Whittle's function [Eq. (3)], but varying in the opposite direction to the scale's lightness. Since both ends of a lightness scale have low saturation,[23] Whittle's function is applied only to the interior of the scale. Also, Whittle's function is scaled to a maximum saturation of 60, since it can be harder to discriminate the lightness of highly saturated colors.[42]
b. Constant variant: Saturation is set to a constant value s. We set $s = 45$, a relatively low saturation, as in the Whittle variant.

A candidate color is considered for admission into the search space whose target lightness and saturation are closest to its own. Focusing first on lightness, candidate color c is admitted into the search space for scale position j if its lightness L_c falls within $1/x$ of the difference between target lightness L_j and the target lightness of the closer of the two adjacent positions—i.e., either L_{j-1} or L_{j+1}, specifically L_{j-1} if $L_c < L_j$ or L_{j+1} if $L_c > L_j$. Thus, if $L_c < L_j$, c is admitted if the following is true:

$$\left|L_j - L_c\right| < \frac{1}{x}\left|L_j - L_{j-1}\right| \tag{4}$$

For example, suppose $x = 10$, $L_j = 30$, $L_c = 29$, and $L_{j-1} = 10$, then the algorithm would accept c into the search space of position j.

The selection of the value of x in Eq. (4) involves a trade-off between the demands of luminosity contrast and overall color discriminability. The higher the selected value of x, the tighter the fit between the lightness of the colors in position j's search space and j's target lightness, and thus the greater the lightness contrast between the ultimately-selected color and its neighbors in the scale. But since a higher x represents a stricter criterion for admission to the search space, the consequence is a reduction in the size of the search space. Since the subsequent search in Step 2, is designed to maximize the discriminability of the colors in the scale, the reduction in the search spaces tends to reduce the discriminability of the colors in the scale.

Another danger of an overly strict criterion x is that it can happen that no colors will be found for some of the

scale positions. In the event that the algorithm is not able to find any colors for one or more of the search spaces, Step 1 is repeated with the value of x reduced to half its previous value, thus loosening the criterion for entry into the search spaces, and with only the empty search spaces considered as candidates. This process iterates until all the color spaces are nonempty.

Identical considerations apply to the determination of whether a candidate color's saturation meets the requirements of a particular search space.

In the work reported here, x was set to 20 and only a single iteration of Step 1 was required to find colors for each of the search spaces, for both variants of the algorithm.

Step 2. Searching the color spaces. Before the search, an initial color scale is constructed by randomly selecting one color from each positional color space. This scale, referred to as the Current Scale, is progressively modified during the search. For the remainder of Step 2, the following substep is repeated until all the search spaces are empty:

Step 2a. Select and evaluate a candidate color. Randomly select a nonempty color space p. From this space, randomly select and remove a candidate color j. Compare color j to the color i currently occupying position p in the Current Scale. If color j is determined to enhance the discriminability of the Current Scale, according to Eqs. (1) and (2), then color j is substituted for color i in the Current Scale. Otherwise color j is not considered further. The CIE DE2000 metric[43] served as the measure of color distance D in our implementation of Eqs. (1) and (2).

Once the search process terminates, the Current Scale is output.

The algorithm, i.e. Steps 1 and 2, is repeated 10 times. From the 10 scales produced, the scale with the highest minimum pairwise color distance is selected as the final output; ties are broken by selecting the scale with the highest average pairwise color distance. Pairwise color distance is the set of distances between each pair of colors in the scale; in a seven-color scale, there are 21 pairwise comparisons.

The choices of CIELAB color space and of the CIE DE2000 distance metric are not essential elements of the algorithm. This color space and distance metric are especially suited to small color differences, but the scales being produced include both small and large color differences. CIE DE2000 is one of the most advanced color difference metrics in use.[43] However, a different color space and distance metric could be substituted for these.

EXPERIMENT

The Motley algorithm was evaluated by an experimental comparison of scales produced by the algorithm, ordered by lightness but unordered by hue, with both multihue scales, which are not ordered by lightness or hue, and perceptually-linear scales, ordered by both lightness and hue. The scales were tested on both absolute value identification and relative comparison tasks. Previous research has shown that unordered multihue scales are superior for value identification,[4–6] while perceptually-linear scales are superior for relative comparison tasks.[2,4–6] We evaluated scales created by the two Motley variants, the Whittle based saturation variant and the Constant saturation variant. Our hypothesis was that Motley scales would be as effective as unordered multihue scales on identification tasks and as effective as the perceptually-linear scales on relative comparison tasks.

Method

Participants. Thirty 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.[44] All participants had normal or correct-to-normal vision. The experiment lasted ~45 minutes. Subjects were assigned randomly and equally to either the identification or comparison condition.

Materials. A stimulus consisted of a 10×10 grid and a legend (see Fig. 1). Each cell on the color grid subtended a 2.54° visual angle. The colors on each stimulus were taken from a different 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°.

Three scale types were used—Unordered, Perceptually-linear, and Motley—with two instantiations of each type. The two multicolored Unordered scales, Rainbow and Weather, were not ordered by either lightness or hue. The two Perceptually-linear scales, the monochrome Grayscale and the multicolored HSB (for Hue-Saturation-Lightness in Munsell space), were both ordered by lightness and HSB was ordered by hue as well, in accordance with Spence *et al.*'s[2] Perceptual Linearity hypothesis. The two Motley instances were Whittle and Constant, according to whether the color saturations were determined by Whittle's function or were constant. They were ordered by lightness, but not ordered by hue. Rainbow was constructed by using the built-in "rainbow" set of hues from the *R* statistical computing environment.[45] The Weather scale came from the Washington Post daily weather map (May 2003). The Grayscale scale was created by varying luminance in equal steps from black to white. The HSB scale, created by Spence *et al.*,[2] varied linearly in Munsell value (brightness), hue, and chroma (saturation), with value and chroma varying in opposite directions. sRGB and CIELAB values of the colors in each scale are shown

in Table I. All stimuli were created using *R*.[45] The experiment was presented using E-Prime.[46]

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 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 six trials, resulting in 42 trials per scale, or a total 252 trials for all 6 scales.

Participants were tested individually and were seated ~43 cm from the computer monitor. To minimize search time, the location of the target "X" was presented on a blank screen before each trial. After the participant hit the space bar, the graph was presented. The participant's task was to determine the numerical value associated with the target color, and then enter the appropriate value (1–7) on the keypad. Response times were measured from the time the graph was presented until the participant responded. After a response was made, the next trial started. Stimuli were presented in block-randomized order, blocked by scale. Scale blocks were further blocked by scale type (Unordered, Perceptually-linear, Motley); the order of two scale blocks within a scale type and the order of the three scale type blocks were randomized.

Before the blocks for a given scale type, the participants were given brief training with only three colors and a 3-by-3 grid: (blue, green, red) for Unordered [RGB values (0 0 255), (0 255 0), and (255 0 0), respectively], green-scale for Perceptually-linear [RGB values (147 226 125), (75 125 68), (26 60 28)], and (blue, green, red) for Motley (RGB values (130 18 17), (145 181 130), and (192 241 254), respectively). Next they were introduced to the legends for the two scales in the block, followed by the test trials.

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" 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 locations of both targets were determined randomly, and each participant received a different random set of graphs. Thus, a total of 42 (21 pairwise comparisons × 2) different graphs were created for each scale, or a total of 252 trials.

Participants were tested individually and were seated ~43 cm from the computer monitor. To minimize search time, the location of both targets (the "X" and the "O") was presented on a blank screen before each trial. After the subject hit the space bar, the graph was presented. The participants' task was to determine whether the value of the "X" color or the value of the "O" color was greater on the legend, and then respond by pressing the "z" or the "/" key (labeled with an "X" or an "O" respectively). 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.

TABLE I. Colors in experimental scales in sRGB and CIELAB notation.

R	G	B	L*	a*	b*
Motley scales					
			Whittle		
255	255	255	100	0	0
192	241	254	92	−14	−12
239	189	220	82	22	−8
207	161	124	70	14	26
58	139	95	52	−34	16
162	17	87	36	58	0
0	0	0	0	0	0
			Constant		
255	255	255	100	0	0
86	250	248	90	−44	−12
251	172	244	80	38	−24
222	138	101	66	30	34
77	132	57	50	−30	34
142	49	102	36	44	−10
0	0	0	0	0	0
Unordered scales					
			Rainbow		
255	0	219	59	90	−43
73	0	255	33	70	−106
0	146	255	59	−1	−64
0	255	146	89	−70	37
73	255	0	89	−73	82
255	219	0	89	1	87
255	0	0	54	81	70
			Weather		
255	255	255	100	0	0
173	216	230	84	−12	−12
0	0	205	23	58	−95
0	255	0	88	−79	81
144	238	144	87	−43	36
255	192	203	84	24	4
255	0	0	54	81	70
Perceptually-linear scales					
			HSB		
171	214	160	82	−22	22
179	151	76	64	4	43
171	102	28	50	25	50
180	53	18	42	51	49
163	20	24	35	55	39
114	5	40	24	45	13
51	0	65	9	32	−28
			Grayscale		
255	255	255	100	0	0
213	213	213	85	0	0
170	170	170	70	0	0
128	128	128	54	0	0
85	85	85	36	0	0
43	43	43	18	0	0
0	0	0	0	0	0

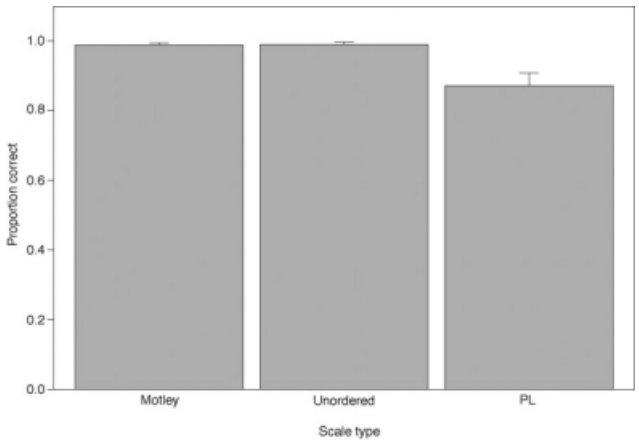

FIG. 2. Accuracy on Identification task (Error bars represent 95% confidence intervals.).

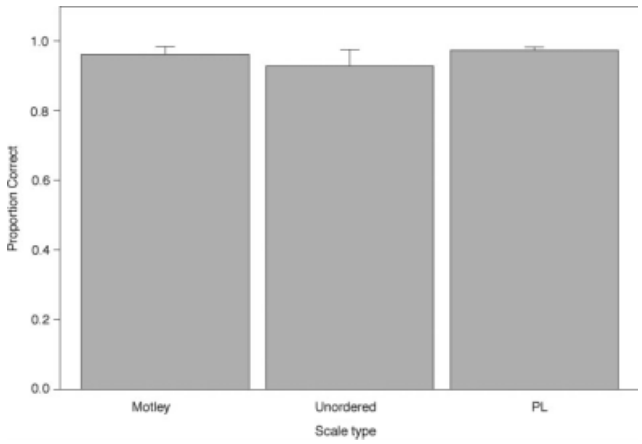

FIG. 4. Accuracy on Comparison task (Error bars represent 95% confidence intervals.).

Results

Accuracy, Response Time. Task by Scale Type analyzes revealed significant Task × Scale Type interaction effects for accuracy, $F(2,56) = 31.77$, $P < 0.001$, and for response time, $F(2,56) = 21.33$, $P < 0.001$. Given these interaction effects, we will discuss results for identification and comparison tasks separately. Data from one participant on the Perceptually-linear scales were missing.

Identification Task. As can be seen in Figs. 2 and 3, Motley scales performed similarly to Unordered scales and both of these were superior to Perceptually-linear scales on this task. Significant effects of Scale Type were found both for accuracy, $F(2,28) = 58.85$, $P < 0.001$, and for response time, $F(2,28) = 22.93$, $P < 0.001$. Tukey HSD pairwise comparisons confirmed that on both variables, Motley and Unordered scales were not significantly different from each other and both were faster and more accurate than Perceptually-linear scales ($P < 0.05$).

Further analyzes were conducted on the six scales, taken singly. Comparisons between the two Motley scales (Whittle saturation and Constant saturation) showed them to function equivalently in terms of both accuracy and efficiency, suggesting that the respective variants of the Motley algorithm they represent are equally effective. The Motley variants were each compared to the HSB scale, developed by Spence *et al*.,[2] to assess the effects of Motley's nonlinear hues compared to HSB's linear hues, determined by the perceptual-linear hypothesis. Comparisons between the two Motley scales and HSB mirrored the differences between their respective scale types with respect to response time and accuracy, with the Motley scales being more accurate and faster than the HSB scale, except that the Constant scale was only marginally ($P < 0.07$) more accurate than the HSB scale.

Comparison Task. As shown in Figs. 4 and 5, the Motley scales performed equivalently to the Perceptually-linear scales and generally better than the Unordered scales on this task. Again, significant effects of Scale Type were found for both accuracy, $F(2,28) = 3.58$, $P < 0.05$, and for response time, $F(2,28) = 9.56$, $P < 0.001$. Tukey HSD pairwise comparisons for response time showed performance on Motley scales to be equivalent to that on Perceptually-linear scales and both of these to be superior to Unordered scales ($P < 0.05$).

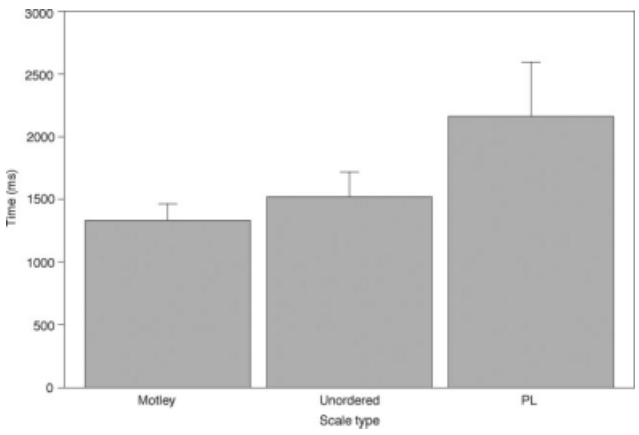

FIG. 3. Response time on Identification task (Error bars represent 95% confidence intervals.).

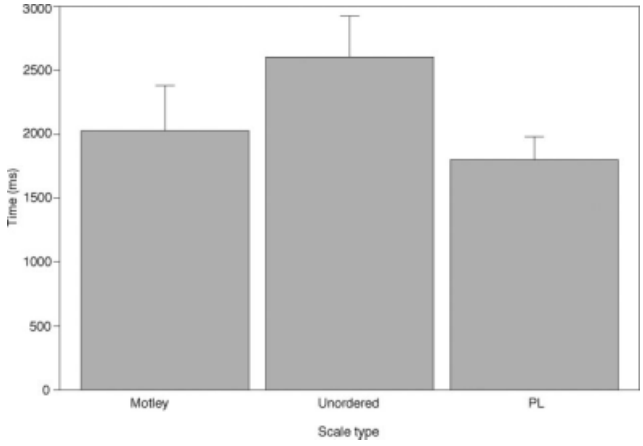

FIG. 5. Response Time on Comparison task (Error bars represent 95% confidence intervals.).

In terms of accuracy, none of the pairwise accuracy comparisons were significant; the Perceptually-linear scales were only marginally more accurate than Unordered scales ($P < 0.07$). The absence of differences was likely a ceiling effect given the high levels of accuracy in all three conditions. Similar ceiling effects in the accuracies of Perceptually-linear and Unordered scales on the comparison task have been found by some researchers,[47] but not by others.[4]

Analyzes examining the 6 scales separately again found that the two Motley scales did not differ. Comparisons between the Motley scales and the HSB scale mirrored the differences between their respective Scale Types, being similar to each other with respect to both response time and accuracy.

Discussion

The findings replicated previous evidence[4–6] concerning the traditionally-used multihue and perceptually-linear scales, namely that multihue scales are superior on absolute value identification tasks and perceptually-linear scales are superior on relative comparison tasks. The Motley algorithm successfully integrated the strengths of each of the two traditional scale types. Specifically, the scales produced by the algorithm elicited comparable performance to the multihue scales on the identification task and to the perceptually-linear scales on the comparison task, as well as superior performance to the perceptually-linear scales on the identification task and to the multihue scales on the comparison task, with the exception that accuracy was uniformly high on the comparison task.

The results do not lend support to an advantage of either of the Motley variants, Constant saturation or Whittle-based saturation, over the other. Thus, ordering by saturation in the Whittle variant does not appear to reinforce the effects of the lightness ordering. We suspect that for larger scales, of 10 or more colors, constructed with Whittle-based saturations, the lighter colors may be so desaturated as to be more difficult to discriminate. Further research is needed to resolve this question.

The experiment shed light on the Perceptual Linearity hypothesis, originally proposed to provide conditions for adding variegated coloration to lightness scales intended for relative comparison tasks.[2] The results cast doubt on the hypothesis that perceptual linearity is a necessary requirement for such scales. On the relative comparison task, neither of the Perceptually-linear scales was superior to the Motley-generated scales, which were unordered in hue.

With regard to the identification task, the HSB scale did improve accuracy relative to Grayscale (Tukey HSD, $P < 0.05$), whether that was due to the linearity of HSB's hues or simply to its multicoloration. However, the finding that people performed identifications more slowly with the HSB scale than with the Motley scales is consistent with the alternate hypothesis, that nonlinearity in hue is preferable for efficient identification.[32,33] However, since the Motley scales are highly discriminable and since highly discriminable color sets are nonlinear, the results do not clearly differentiate between the effects of nonlinearity and discriminability.

GENERAL DISCUSSION

We have provided evidence that a single color scale can be used effectively in both value identification and relative comparison tasks. We have also proposed an algorithm, Motley, that constructs such dual-use scales by promoting hue and lightness discriminability and a lightness ordering. The algorithm consists of two steps. In the first step, constraints on lightness and saturation are defined for each ordinal position in the scale, and search spaces are constructed for each position conforming to those constraints. The purpose of this step is to ensure that the scale colors will be ordered and discriminable by lightness. The second step, involving heuristic search within the search spaces is designed to maximize the color discriminability of the scale, adapting an algorithm proposed by Campadelli *et al*.[11]

The experimental evaluation of Motley with human participants provided support for the hypothesis that its scales incorporate the respective strengths of unordered multicolored scales and ordered lightness scales, making them well-suited for both absolute value identification and relative comparison tasks. The scales support efficient relative comparisons between color-coded regions in a visualization, as well as fast and accurate matching of a region to the legend to extract absolute values.

Theoretical Implications

The success of Motley's scales on quantitative tasks calls into question the Perceptual Linearity hypothesis concerning the conditions that allow multicoloration to support the lightness representation of quantity.[2] On the relative comparison task, the type of task for which the Perceptual Linearity hypothesis was designed, Motley's scales were equally effective even though they were not perceptual linear with regard to hue. However, the question the Perceptual Linearity hypothesis was introduced to address remains unanswered, namely why color sometimes supports and sometimes impedes the lightness representation of quantity. Also, while perceptual linearity may not be a necessary condition for relative comparison, the PL scales were no worse than the Motley scales on the relative comparison task and, like the Motley scales, were superior to the Unordered scales in response times. Finally, on the identification task, a task for which the Perceptually Linearity hypothesis was not designed, the PL scales were inferior to both the Unordered scales and the Motley scales.

Our results may also be viewed in terms of other research concerning the extent to which color and lightness support or interfere with one another on various tasks. On the relative comparison task, the addition of

color to lightness scales, as in the HSB or Motley scales, neither improved nor harmed performance relative to the monochrome grayscale.[2] On the identification task, the superior accuracy on the multicolored HSB and Motley scales relative to the grayscale suggests that the addition of color to lightness scales improves visual search.[48] This may reflect a more general phenomenon whereby variations in independent features of color facilitate visual search.[49]

Fairchild and Pirrotta's[38] L^{**} metric improved upon the CIELAB L^{*} metric with the goal of predicting the perceived lightness of chromatic objects. Our informal observation that equal intervals of L^{**} do not correspond to equal differences in perceived lightness among highly dissimilar colors received only indirect support in our experiment. Scales produced by the Motley algorithm, using an exponential function of L^{**} to determine lightness differences, served as well as for lightness comparisons as a perceptually-linear scale with lightness values equally spaced in Munsell space. Clearly, more systematic work is needed to explore the hypothesized phenomenon, for instance research comparing scales produced by various functions of L^{**}, both linear and nonlinear, in different chromatic contexts.

It is noteworthy that the Motley algorithm appears to have the serendipitous effect of producing sets of colors belonging to distinct color categories. The colors in the two Motley scales in Fig. 1 appear to belong to the categories white, cyan, pink, brown, green, purple, and black. The question of whether color category influences visual search and whether these categories are verbally mediated has been the subject of ongoing debate.[27,50] In any case, the colors in the PL HSB scale in Fig. 1 are clearly not differentiable on the basis of category; as characterized by Munsell hue, the sequence is green, green-yellow, yellow, yellow-red, red, red-purple, and purple. More generally, this is what one would expect from scales that are ordered by hue, as multicolored ordinal scales often are. Motley's scales, in contrast, are expressly unordered by hue. The extent to which color category contributes to the success of Motley scales is a subject for further research.

Practical Implications

Hybrid scales, such as those produced by Motley, with ordered lightness and highly-discriminable unordered colors, should introduce new alternatives into the choice of color codes for practical applications. The Motley scales were as efficient and accurate as the traditionally-used lightness scales on relative comparison tasks and as efficient and accurate as the traditionally-used multicolored scales on absolute value tasks. Thus, the sort of compromises of accuracy versus efficiency or of relative comparison versus absolute identification performance described in the introduction for the design of school atlases need not always be necessary. Of course, other considerations will still enter into the choice of color scales, including conventional color symbology and color-size illusions.[51]

We do not claim that Motley is the only or best algorithm for producing dual-use color scales. It does represent a "proof of concept" for the idea that ordered lightness together with highly discriminable—thus unordered—hues can offer a solution to the problem of producing color scales that are useful for both relative comparison and absolute value identification tasks.

ACKNOWLEDGMENTS

The authors thank Andrew Blumenfeld for running the experiment. They also thank Bill Kennedy and two anonymous reviewers for their valuable comments on this work.

The views and conclusions contained in this document should not be interpreted as necessarily representing the official policies, either expressed or implied, of U.S. Navy.

1. Christ RE. Review and analysis of color coding research for visual displays. Hum Factors 1975;7:542–570.
2. Spence I, Kutlesa N, Rose DL. Using color to code quantity in spatial displays. J Exp Psychology: Appl 1999;5:393–412.
3. Trafton JG, Kirschenbaum SS, Tsui TL, Miyamoto RT, Ballas JA, Raymond PD. Turning pictures into numbers: Extracting and generating information from complex visualizations. Int J Hum Comput Stud 2000;53:827–850.
4. Breslow LA, Trafton JG, Ratwani RM. A perceptual process approach to selecting color scales for complex visualizations. J Exp Psychology: Appl 2009;15:25–34.
5. Merwin DH, Wickens CD. Comparison of eight color and gray scales for displaying continuous 2D data. In: Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting. Santa Monica, CA: Human Factors and Ergonomic Society; 1993. p 1330–1334.
6. Phillips RJ. An experimental investigation of layer tints for relief maps in school atlases. Ergonomics 1982;25:1143–1154.
7. Brown R, MacLeod D. Color appearance depends on the variance of surround colors. Curr Biol 1997;7:844–849.
8. Ware C. Color sequences for univariate maps: Theory, experiments, and principles. IEEE Comput Graphics Appl 1988;8:41–49.
9. Nagy AL. Interactions between achromatic and chromatic mechanisms in visual search. Vision Res 1999;39:3253–3326.
10. Phillips RJ, De Lucia A, Skelton N. Some objective tests of the legibility of relief maps. Cartographic J 1975;12:39–46.
11. Campadelli P, Posenato R, Schettini R. An algorithm for the selection of high-contrast color sets. Color Res Appl 1999;24:132–138.
12. Robertson PK, O'Callaghan JF. The generation of color sequences for univariate and bivariate mapping. IEEE Comput Graphics Appl 1986;6:24–32.
13. Van Laar D, Flavell R. Two methods for producing discriminable colour sets for computer displays. SIGCHI Bull 1991;23:75–75.
14. Robertson PK. Visualizing color gamuts: A user interface for the effective use of perceptual color spaces in data displays. IEEE Comput Graph Appl 1988;8:50–64.
15. Robertson PK, Hutchins M, Stevenson DR, Gunn SBC, Smith D. Mapping data into colour gamuts: Using interaction to increase usability and reduce complexity. Comput Graphics 1994;18:653–665.
16. Brewer CA. A transition in improving maps: The ColorBrewer example. U.S. Report to the International Cartographic Association. Spec issue Cartography Geographic Inf Sci 2003;30:155–158.
17. Rogowitz BE, Treinish LA. An architecture for perceptual rule-based visualization. In: Proceedings of the IEEE Visualization Conference. Washington DC: IEEE Comp. Society; 1993. p 236–243.
18. Carter RC, Carter EC. High-contrast sets of colors. Appl Opt 1982;21:2936–2939.

19. Brewer CA. Spectral schemes: Controversial color use on maps. Cartography Geographic Inf Syst 1997;24:203–220.
20. Clarke FJJ, Leonard JK. Proposal for a standardized continuous pseudocolor spectrum with optimal visual contrast and resolution. Presented at the Third International Conference on Image Processing and its Applications. Warwick, UK; 1989. p 687–691.
21. Lehmann TM, Kaser A, Repges R. A simple parametric equation for pseudocoloring grey scale images keeping their original brightness progression. Image Vision Computing 1997;15:251–257.
22. Brewer CA. Color use guidelines for mapping and visualization. In: MacEuchren MA, Taylor DRF, editors. Visualization in Modern Cartography. Tarrytown, NY: Elsevier; 1994.
23. Levkowitz H, Herman GT. The design and evaluation of color scales for image data. IEEE Comput Graphics Appl 1992;12:72–80.
24. Carpenter PA, Shah P. A model of the perceptual and conceptual processes in graph comprehension. J Exp Psychol Appl 1998;4:75–100.
25. Peebles D, Cheng PCH. Modeling the effect of task and graphical representation on response latency in a graph reading task. Hum Factors 2003;45:28–46.
26. Carter RC. Visual search with color. J Exp Psychol: Hum Percept Perform 1982;8:127–136.
27. Smallman HS, Boynton RM. Segregation of basic colors in an information display. J Opt Soc Am A 1990;7:1985–1994.
28. Nagy AL, Sanchez RR. Chromaticity and luminance as coding dimensions in visual search. Hum Factors 1992;34:601–614.
29. Wolfe JM. Moving towards solutions to some enduring controversies in visual search. Trends Cognitive Sci 2003;7:70–76.
30. Brawn P, Snowden RJ. Can one pay attention to a particular color? Percept Psychophys 1999;61:860–873.
31. Shi XQ, Sallstrom P, Welander U. A color coding method for radiographic images. Image Vision Comput 2002;20:761–767.
32. Bauer B, Jolicoeur P, Cowan WB. Visual search for colour targets that are or are not linearly-separable from distractors. Vision Res 1996;36:1439–1466.
33. D'Zmura M. Color in visual search. Vision Res 1991;31:951–966.
34. Healey CG. Choosing effective colours for data visualization. In: Proceedings of the IEEE Visualization 96. New York: ACM Press; 1996. p 263–270.
35. Silverstein LD, Lepkowski JS, Carter RC, Carter EC. Modeling of display color parameters and algorithmic color selection. In: Proceedings of the SPIE, Advances in Display Technology VI. Bellingham, WA: SPIE; Vol. 624. 1986. p 26–34.
36. Campadelli P, Mora P, Schettini R. Color set selection for nominal coding by Hopfield networks. Visual Comput 1995;11:150–155.
37. Campadelli P, Schettini R. A system for the automatic selection of conspicuous color sets for qualitative data display. IEEE Trans Geoscie Remote Sensing 2001;36:2283–2286.
38. Fairchild MD, Pirrotta E. Predicting the lightness of chromatic object colors using CIELAB. Color Res Appl 1991;16:385–393.
39. Whittle P. Brightness, discriminability and the 'Crispening Effect'. Vision Res 1992;32:1493–1507.
40. Kalvin AD, Rogowitz BE, Pelah A, Cohen A. Building perceptual color maps for visualizing interval data. In: Rogowitz BE, Pappas TN, editors. Human Vision and Electronic Imaging V. Bellingham, WA: SPIE; Vol. 3959. 2000. p 323–335.
41. Wang M, Parker KJ, Spaulding KE, Yu Q, Miller RL. Perceived lightness difference with regard to spatial frequency and amplitude modulation. In: Rogowitz BE, Pappas T, editors. Proceedings of the SPIE, Human Vision and Electronic Imaging. Bellingham, WA: SPIE; Vol. 3959. 2000.
42. Zhang H, Montag ED. How well can people use different color attributes? Color Res Application 2006;31:445–457.
43. CIE Pub 142–2001. Improvements to industrial colour-difference evaluation. Vienna: CIE Central Bureau; 2001.
44. Waggoner T. Ishihara Compatible Pseudoisochromatic Plate (PIPIC) Color Vision Test, 24 Plate Edition. Elgin, IL: Good-Lite Co.
45. R development core team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2007.
46. Schneider W, Eschman A, Zuccolotto A. E-prime user's guide. Pittsburgh: Psychology Software Tools; 2002.
47. Breslow LA, Ratwani RM, Trafton JG. Cognitive models of the influence of color scale on data visualization tasks. Hum Factors 2009;51:321–338.
48. Nagy AL, Winterbottom M. The achromatic mechanism and mechanisms tuned to chromaticity and luminance in visual search. J Opt Soc Am A 2000;17:369–379.
49. Nagy AL, Thomas G. Distractor heterogeneity, attention, and color in visual search. Vision Res 2003;43:1541–1552.
50. Yokoi K, Uchikawa K. Color category influences heterogeneous visual search for color. Opt Soc Am A 2005;22:2309–2317.
51. Cleveland WS, McGill R. A color-caused optical illusion on a statistical graph. Am Statistician 1983;37:101–105.