Using Color to Code Quantity in Spatial Displays

Ian Spence
University of Toronto

Natasha Kutlesa
University of Calgary

David L. Rose
University of Illinois at Urbana-Champaign

Participants made simple and complex judgments in 2 experiments that examined the use of color to code quantity in spatial displays. The coding assignments were chosen to evaluate the principle of perceptual linearity in color space. In Experiment 1, participants compared all possible pairs of colors used to represent magnitudes. Comparisons were made most rapidly with a scale that varied only brightness (B) and most accurately with a scale that covaried hue (H) with saturation (S) and brightness (H+S+B scale). In Experiment 2, clusters were identified fastest with the H+S+B scale, followed by brightness and bipolar scales, whereas a nonlinear, hue-only scale was slowest and produced the least accurate judgments. Coding assignments close to perceptual linearity were best for both simple and complex judgments in data visualization. However, hue conferred an advantage if the task involved segregation or classification.

The use of color to code quantity is widespread in statistical graphics, medical imaging, engineering displays such as aircraft instrument panels, and an unnumbered multitude of scientific visualizations. The use of color to represent quantity is not straightforward because our perception of light is not unidimensional, even though the physical variable of dominant wavelength is. Some experimental work has sought to find acceptable methods of assigning colors, but much practice is arbitrary and does not capitalize on what is known about human color vision. Nearly all experimental work has used reflected light (printed displays), whereas much contemporary visualization uses transmitted light (CRT, LCD, or projected displays). Only a few studies have investigated assignments on the basis of standard perceptual color spaces, such as the Munsell or CIELUV spaces, and, therefore, it is often not possible to generalize or make comparisons among studies. Unfortunately, many investigators have failed to control or measure study time, and, thus, the possibility of speed–accuracy effects compromises interpretation. Finally, the likelihood that performance may be task dependent has often been ignored, with contradictory results and conflicting recommendations reflecting the use of different tasks in different studies.

Coding Category by Color

Since the first statistical graphs were published (Playfair, 1786), color has been used successfully for separating, emphasizing, and identifying data. Color appeared even earlier in statistical maps (see Funkhouser, 1937), in which it was used to distinguish classes. The form reached its acme in the wonderful "cartes figuratives" of the French engineer Charles Joseph Minard (see Robinson, 1967) and his disciples (see, for example, Minis-
tère des Travaux Publiques, 1884). From a modern perspective, Brewer (1994) gave an excellent discussion of the use of color to code qualitative distinctions in cartography and other visualizations. Color can enhance the discriminability of objects in statistical graphs and may improve performance on recall, retention, search-and-locate, and decision tasks and may improve the comprehension of instructional materials (Benbasat, Dexter, & Todd, 1986; Christ, 1983; Gremillion & Jenkins, 1981; Hoadley, 1990; Lamberski & Dwyer, 1983; Vogel, Dickson, & Lehman, 1986). This use of color to code categories in graphs and charts or to segregate objects in displays is relatively uncontroversial, is consistent with known perceptual principles, and is widely practiced. Indeed, such use is optimal for data-analytic tasks involving segregation and classification (Cleveland, 1985; Lewandowsky & Spence, 1989). There is less agreement on how color should be used to code quantity.

Coding Magnitude by Color

Cartographers use color to represent progressive features, such as the height of terrain. A common assignment uses colors ranging from blue at sea level, to the greens for low areas with vegetation, through yellows and browns for hills, and finally to gray or white for mountains. The map thus constructed bears a strong resemblance to what a soaring observer might see looking down on the terrain. The correspondence makes ecological sense, and the general pattern has become firmly established, although there are countless variations in implementation.

In other contexts, colors are not chosen according to cartographic convention. A single hue is often used, varying lightness or brightness to denote magnitude, or, sometimes, the display is colored by two hues, each varying in brightness or saturation. The majority of applications, however, use multiple color assignments, and much scientific visualization uses rainbow scales that have colors that are ordered according to their position in the visible spectrum and that are assigned to magnitudes according to this sequence.

Although both continuous and discontinuous scales are found in practice, we consider only the latter. Contour and mosaic displays are more common than representations that vary color smoothly and continuously. The success of any discrete coding scheme may depend partially on the choice of cutpoints that effect the conversion to discrete representation. We do not explore the issue further, and it should not be assumed that our results necessarily generalize perfectly to continuous representations.

Color Spaces

The obvious and natural psychological dimensions of color are hue, saturation, and lightness or brightness. Hue refers to the chromatic attribute of a color, saturation is the amount of hue in a color, and lightness or brightness quantifies how much light is reflected or emitted, the former being used with reflected light and the latter with transmitted light. A set of colors can vary in hue (e.g., red), saturation (grayish red to deep red), and brightness (from very dim to dazzling).

Munsell Color Space

This system, devised by the American artist Albert Munsell in 1905, has undergone extensive evaluation, refinement, and revision. Munsell used the terms hue, value, and chroma to refer to the dimensions of his space, the latter two terms being approximately synonymous with lightness and saturation. The Munsell space uses a mixed coordinate system in which hue is specified in polar coordinates—angular location around the circumference of a circle. Value and chroma are specified in conventional cartesian fashion. The Munsell system is defined for a finite set of colors and is thus particularly appropriate for choosing a few discrete, widely spaced colors (Robertson, 1988). Because our study makes use of a small, discrete number of levels, we used the Munsell system. However, to realize the colors on a monitor, other color spaces served as intermediaries.

Red, Green, and Blue (RGB) Color Space

Engineers have developed descriptions that reflect the characteristics of display devices rather than our perception of color. In additive displays, such as CRTs, mixing red, green, and blue lights
produces colors. On a monitor, colors are specified by choosing the intensities (between 0% and 100%) produced by the electron guns that excite the red, green, and blue screen phosphors. The set of possible triples defines a three-dimensional cube known as the RGB space. Unfortunately, distances in RGB space are not proportional to perceptual dissimilarities among colors; colors that appear alike may be widely separated in RGB space, whereas others that seem quite distinct may have RGB coordinates that differ little. Thus, there is little psychological meaning associated with position in RGB space. This space should never be used as a basis for choosing coding assignments; nonetheless, RGB coordinates are required to realize a given color on a particular CRT.

Commission Internationale de l’Eclairage (CIE) Color Space

This system, devised in 1931 by the CIE, is based on color-matching data and is the accepted basis for the specification of color in industrial colorimetry. The CIE coordinate $Y$ measures luminance, whereas $x$ and $y$ contain information about both hue and saturation. The CIE space, unfortunately, distorts lines of constant hue and circles of constant saturation, but it may be transformed so that just-noticeable differences in color correspond to approximately equal distances. In 1976, two new spaces, CIELAB and CIELUV, were proposed; these are suitable for choosing colors for continuous or graded representations, according to Robertson (1988). For a review and critique of these and other color appearance models, see Fairchild (1998).

CIE coordinates may be mapped into RGB gun intensities, although, in practice, a monitor will only be able to display a subset of colors in CIE space (the gamut of the monitor). Because the relationship of Munsell space to CIE space has been established, transformations from one to the other can be made. The interested reader can find more detail in Wyszecki and Stiles (1982). Although it seems obvious that any comparative evaluation of coding assignments should be precise regarding the specification of colors, much previous empirical work fails to characterize colors either in terms of CIE or Munsell space.

Empirical Evaluations of Coding Assignments

Single-Hue Schemes

Cartographers and statisticians have enunciated standards for statistical maps to improve clarity, legibility, and aesthetic properties (e.g., Bertin, 1983; Brewer, 1994; Mersey, 1990; Memonier, 1993; Robinson, 1952, 1967), and many authorities have advocated varying lightness to code quantity. In cartography, McCarty and Salisbury (1961) showed that gradient scales — scales in which single colors are ranged from light to dark (as in gray scales) monochromatic gradient values of one hue and multichromatic gradient sequences of several hues — yielded better performance than spectral series, with no difference found among the gradient scales. Cuff (1972, 1973, 1974) also concluded that a simple progression of values of a single hue is the most effective way to represent quantitative information, but Miller (1974) and Mersey (1980) disputed this, contending that variation in hue can also play an important role.

Lewandowsky et al. (1993) recommended classed monochrome maps (shades of blue or black) for cluster identification tasks. This complemented work by Antes and Chang (1990), who analyzed eye movements and found shorter fixation times with monochrome maps. Merwin and Wickens (1993) also provided support for the hypothesis that monochrome scales best support accurate rank ordering tasks. Similarly, Levkowitz and Herman (1992) examined the usefulness of color scales for medical image data and showed that participants performed better with a gray scale than with color scales.

Two-Hue Schemes

Other investigators (e.g., Petchenik, 1983) have recommended sequences with two hues forming opposing extremes, with the intermediate elements varying in saturation and brightness (e.g., Brewer, 1994, 1996). The central element is often gray. It should be easier to identify three or four saturation or brightness levels of one hue and three or four saturation or brightness levels of another hue than to identify six or eight levels of either hue alone (Egeth & Pachella, 1969). However, brightness and saturation do not vary monotonically with the magnitudes represented; in
many bipolar scales, brightness is usually the greatest and saturation the least in the middle of the scale.

Performance on various tasks has generally not supported the use of bipolar color scales (Cuff, 1973, 1974; Lewandowsky et al, 1993; Mersey, 1990). However, Carswell, Kinslow, Pickle, and Herrmann (1995) found that bipolar sequences were more effective than gray scales for three different statistical map-reading tasks. Similarly, Lewandowsky and Behrens (1995) found that a bipolar sequence led to better performance in a cluster detection task than multiple-hue sequences.

Rainbow Scales

These multiple-hue sequences are spectrum based. Rainbow scales provide a color set that many find attractive, although some may argue that spectral assignments tend to be garish. Even if esthetic considerations are ignored, a rainbow scale would seem to be a poor candidate for encoding quantity.

The spectral hues red, orange, yellow, green, blue, indigo, and violet do not convey a sense of greater or lesser in the same way that varying brightness or saturation does. Furthermore, brightness and saturation do not vary monotonically with dominant wavelength; people see a bright yellow in the middle of the spectrum surrounded by darker reds and blues. It is not possible to achieve equal brightness by darkening the yellow and lightening the blue and retain the illusion of a rainbow. Even if brightness and saturation are held constant, hue is still a psychologically curved or even circular attribute, a truth known to Sir Isaac Newton as long ago as 1666. Notwithstanding, the rainbow scale has found supporters, and it is probably the most frequently encountered coding scheme. Why this should be so is mysterious because there are scales with much better psychometric properties (Robertson, 1988), and the view of most experts is opposed to the use of spectral assignments (Brewer, 1994), except perhaps when diverging quantitative data must be displayed (Brewer, 1997).

Nonetheless, Ware (1988) found that a rainbow sequence led to more accurate performance when compared with a red-to-green series, two kinds of gray scales, and a sequence varying in saturation. Ware also found that errors for gray scales (but not for rainbow scales) were in the direction predicted by simultaneous contrast—a consequence of lateral inhibition working through the opponency coding of signals in the retinal cells. He attributed the success of the rainbow scale to the fact that it does not vary monotonically with any of the opponent channels. Magnetic resonance imaging (MRI), positron emission tomography (PET), and computerized tomography (CT) images often use rainbow scales with advocates claiming enhanced detection of patterns and more rapid and accurate diagnostic interpretation (e.g., Alfano, Brunetti, Ciariello, & Salvatore, 1992; Brown et al., 1992). In other contexts, studies have sometimes found spectral assignments to be superior to hue-limited scales for some tasks (e.g., Brewer, 1997; Hastie, Hammerle, Kerwin, Croner, & Herrmann, 1995; Mersey, 1990; Ware, 1988).

Other Multiple-Hue Schemes

Hastie et al. (1995) showed that for intermediate complexity map-reading tasks, such as the identification of predominant disease type, and for elementary tasks, such as extracting the value (numerical or categorical) associated with a specific location, multiple-hue scales produced superior performance when compared with monochrome-ordered gray and blue formats. McCarty and Salisbury (1961) showed that gradient sequences of several hues yielded better performance than spectral assignments. Mersey (1990) obtained similar results with recall and recognition tasks, in which color sequences that incorporated both hue and value variations outperformed all others.

Contradictory Recommendations

This brief review shows that there is no strong consensus, even for a single task. Notwithstanding the successes reported by different authors for all varieties of scale, no unambiguous recommendations seem possible. Perhaps the only persistent theme is that monochromatic lightness and brightness scales perform well with a variety of tasks, but there is little agreement on the role of hue. In most studies, the choice of coding assignment was arbitrary (usually lacking any theoreti-
The Principle of Perceptual Linearity

It seems natural to require that the psychological sensation produced by a color be proportional to the number it represents. Regardless of the patterns of dimensional variation in the colors, the observer should experience sensations that are ordered and spaced proportionally to the numbers represented. Otherwise, colors that are perceptually similar may encode numbers that differ considerably, and, conversely, small numerical differences may be represented by markedly different colors. The correspondence between sensation and magnitude should be linear or at least monotonic. Nonmonotonicity is surely unacceptable. Most modern authorities agree that the colors used to represent numbers should lie along a line in perceptual color space (Brewer, 1994; Levkowitz & Herman, 1992; Robertson, 1988). Assignments should be linear and spaced proportional to the numbers represented.

Many coding assignments violate linearity, with rainbow scales being the most flagrant, but not the only, offenders. Other undesirable assignments are sometimes the result of choosing points in RGB space. Robertson (1988) gives a convincing demonstration of how bad an equal-interval linear scale in RGB space can be—the corresponding sequence in CIELUV space is neither equal interval nor linear. A simple way of ensuring linearity, recommended by many investigators, is the use of single-hue sequences that vary either in lightness or saturation. However, other directions in color space are possible and may even be preferable; some studies have demonstrated good results with schemes that are nonlinear in color space (e.g., bipolar schemes or multiple-hue schemes that covary brightness). To accommodate these variations, we suggest a slightly more general principle.

We propose that for a coding assignment to be perceptually linear, it must be possible to form an additive weighted combination of the Cartesian coordinates of each color in perceptual space such that the combination correlates maximally with a linear sequence of numbers. If the coordinates lie along a line in color space, the assignment is clearly (perceptually) linear, but other, less obvious sequences could also be perceptually linear. For example, a scale curving upward through color space would be perceptually linear because an additive combination that assigned almost all weight to brightness would correlate highly with a linear sequence. Conversely, a scale that varied in hue alone could not be perceptually linear because the locus of its coordinates in color space would define the circumference of a circle. No weighted combination of coordinates could possibly correlate highly with a linear sequence of numbers.

This definition of perceptual linearity acknowledges that observers may weight the dimensions of color space differentially. The principle is thus able to predict the good performance of bipolar scales, despite nonmonotonic variation in luminance. Psychologically, this makes sense because in this case, luminance does not carry consistent information with regard to quantity and, hence, is probably de-emphasized or ignored by the observer. The principle also predicts the much poorer performance of (circular) hue-only scales, in which variation in any direction in the subspace of constant brightness cannot be monotonic with quantity.

In our experiments, we chose the colors to be approximately equally spaced in Munsell space, and our four scales varied in their departure from perceptual linearity. One single-hue scale and another multiple-hue assignment were perceptually linear, even though the latter scale varied on all three dimensions, curving through color space. The third was a bipolar scale that deviated somewhat from perceptual linearity. The fourth scale varied in hue alone and clearly was not perceptually linear. Our hypothesis was that assignments close to perceptual linearity would perform best when simple judgments of quantity were made. If, however, the required judgment was more complex, for example involving classification or segregation, we expected that variation in hue would assume greater importance.

Experimental Considerations

There are many potential difficulties that can compromise the interpretation of experiments designed to assess the effectiveness of color coding schemes. We consider some of these
below. In each case, we have designed our experiments to accommodate the issue.

**Stimulus Familiarity**

Participants bring a great deal of knowledge of the world into the laboratory. With real data (e.g., infant mortality rate, blood flow, atmospheric pressure, vegetation) from real contexts (e.g., countries of the European Union, the human brain, the northern hemisphere, a remote-sensing Landsat image), participants' judgments may be affected by their familiarity with the data context. Although real-data displays constitute the natural habitat of data visualization, they introduce variation that can obscure effects due to color alone. Our experiments used artificially constructed data sets, both systematic and random. The random data sets removed all effect of context; there was no underlying structure that a participant might use to assist in making judgments.

**Representativeness**

Experiments conducted using a single data set (e.g., lung cancer rates for a single country) may contain peculiar patterns of variation that somehow influence the results. We tried to avoid this by using a variety of mathematical surfaces that exhibit considerable diversity in shape.

**Task Dependency**

It is clear (e.g., Hastie et al., 1993; Herrmann & Pickle, 1994; Lewandowsky et al., 1993; Merwin & Wickens, 1993; Pickle & Herrmann, 1995) that different results may be obtained using different tasks. Simple tasks, such as direct estimation, seem to produce one set of findings, whereas more complex tasks, such as cluster detection, yield another. We used both a simple and a complex task.

**Speed–Accuracy Trade-Off**

To ensure that participants do not trade off speed for accuracy, some account must be taken of processing time. Either presentation time must be controlled or participants must be timed. We chose the latter course of recording participants' latencies as they performed under identical instructions in all experimental conditions.

**Stimulus Specification**

If investigators do not specify the physical parameters of their stimuli, replication by others is impossible and comparison among studies is made unnecessarily difficult. Our specification of stimulus parameters is given in Munsell, CIE, and RGB coordinates (see Table 1), and the display characteristics of the monitor are in the public domain. The mathematical equations for the surfaces are given, and full details of how our stimulus objects were constructed are supplied.

Experiment 1

This experiment was designed to evaluate performance on a simple task (pairwise comparison of magnitudes) using four coding schemes varying in perceptual linearity.

**Method**

**Participants.** Twenty undergraduate psychology students at the University of Toronto (16 women and 4 men) volunteered to participate. They received course credit for their participation, and, as an incentive to be quick and accurate, the top three performers were rewarded for their efforts. Performance on a screening test using pseudo-isochromatic plates (*Pseudo-Isochromatic Plates*, 1986) indicated that none of the participants had any color vision anomalies. The experimental session lasted approximately 45 min.

**Stimuli.** The stimuli were presented on a 21-in. (53.34-cm) monitor (NEC MultiSync 5D, Itasca, IL) at 1024 × 768 resolution and 24-bit color (ATI Pro Turbo graphics card, Thornhill, Ontario, Canada). The monitor's brightness and contrast were adjusted to intermediate levels. The room was darkened with ambient illumination sufficient for using the keyboard and the mouse. Viewing distance to the screen was between 50 and 70 cm, with a visual angle of approximately 25°. The participant's head was not restrained, and no fixation point was used between trials.
Table 1

<table>
<thead>
<tr>
<th>Munsell</th>
<th>CIE</th>
<th>RGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hue</td>
<td>Value</td>
<td>Chroma</td>
</tr>
<tr>
<td>2.5 G</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>2.5 G</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>2.5 G</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>2.5 G</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2.5 G</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2.5 G</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2.5 G</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

H+S+B (perceptual linearity = 100%)

<table>
<thead>
<tr>
<th>Munsell</th>
<th>CIE</th>
<th>RGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hue</td>
<td>Value</td>
<td>Chroma</td>
</tr>
<tr>
<td>5.0 G</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>5.0 GY</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>5.0 Y</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>5.0 YR</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>5.0 R</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>5.0 RP</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>5.0 P</td>
<td>3</td>
<td>14</td>
</tr>
</tbody>
</table>

Bipolar (perceptual linearity = 95%)

<table>
<thead>
<tr>
<th>Munsell</th>
<th>CIE</th>
<th>RGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hue</td>
<td>Value</td>
<td>Chroma</td>
</tr>
<tr>
<td>5.0 R</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>5.0 R</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>5.0 R</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Neutral gray</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.0 PB</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>5.0 PB</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>5.0 PB</td>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>

Hue-only (perceptual linearity = 48%)

<table>
<thead>
<tr>
<th>Munsell</th>
<th>CIE</th>
<th>RGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hue</td>
<td>Value</td>
<td>Chroma</td>
</tr>
<tr>
<td>7.5 B</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>7.5 GY</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>7.5 YR</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>7.5 R</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>7.5 RP</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>7.5 P</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>7.5 PB</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

Note. The Commission Internationale de l'Eclairage (CIE) values are calculated, not measured, and the $Y$ values shown are luminance factors. The red, green, and blue (RGB) values show how the colors were realized on the NEC 5D monitor and are not appropriate for other displays. Perceptual linearity = proportion of variance in the sequence [1,2,3,4,5,6,7] accounted for by the CIELUV coordinates ($L^*, u^*, v^*$, not shown) of the colors in each of the four scales. R = red; G = green; B = blue; Y = yellow; P = purple; H+S+B = hue covaried with saturation and brightness.

Each stimulus was presented on a very light, neutral gray background (RGB 230/230/230).

Stimulus objects were created using Mathematica (Wolfram, 1991). Five surfaces were constructed (see Figure 1 for the surfaces, their functions, and ranges) and used to create contour and square mosaic displays—both variants are found in scientific visualizations. Random square mosaic displays were constructed by randomly permuting the elements of the prototype mosaic, thus preserving the coding distribution but destroying the spatial meaning of the original. This
manipulation was intended to remove any spatial cues that might assist participants; if a participant were able to construct a mental model of the surface, this could assist in making the judgment. The random square mosaic was used as the basis for the construction of a corresponding random contour display (see Figure 2 for an example of the four displays rendered in gray scale).

Stimulus colors were chosen with help of the Munsell color space (Munsell, 1976). We used a tabulation of the Munsell renotation system (Travis, 1991) and converted each Munsell (hue, value, chroma) triple to the corresponding CIE coordinates \((x, y)\) and luminance factor \((Y)\). Subsequently, the CIE coordinates were converted to the RGB values appropriate to the NEC MultiSync monitor display gamut (see Travis, 1991, pp. 214–270, for the mathematical transformations required). Some iteration of this process was necessary to ensure that the selected Munsell colors fell within the display gamut.

**Color assignment.** We designed four 7-level scales. Coding assignments with a discrete number of levels (ranging from about 5 to 11 or more) are found in the majority of scientific visualizations that use color. Representations that use continuous color coding are relatively uncommon. Scales with 7 levels are very common. The four scales were as follows:

1. **Brightness.** This scale had approximately equal intervals on the Munsell value dimension, with hue and chroma (saturation) held constant. The assignment followed a straight line in color space, and the hue that varied in brightness was green—this choice of hue was arbitrary.

2. **H+S+B.** Hue was covaried with saturation (chroma) and brightness (value). The Munsell colors were chosen such that the sequence was approximately equal interval and spiraled upward through the space. Although the hue angle varied through 216°, the locus of value/chroma formed an almost linear equally spaced sequence from 3/14 to 9/4, thus attenuating the influence of the curvature. The reader may visualize this scale by imagining a slightly longer than semicircular piece of wire with an axis passing vertically through the center of the implied circle. Pick up one end of the wire and stretch it vertically while simultaneously pulling it toward the vertical axis. The vertical stretch should about equal the diameter of the circle.

3. **Bipolar.** This scale had a sequence with a neutral gray center with two surrounding hues varying in saturation and brightness. We chose to vary both saturation and brightness because this is done in bipolar scales seen in many applications. The sequence followed a concave-downward dogleg across Munsell space.

4. **Hue-only.** This scale had a sequence varying in hue alone, with brightness (value) and saturation (chroma) held constant. Such a constant-brightness, constant-saturation scale is unlikely to be used in practice, but it isolates the effect of varying hue alone. Even though only one Munsell dimension was varied, the scale nonetheless spans two dimensions of Cartesian color space—the colors follow a circular path.

The Munsell, CIE and RGB coordinates of each assignment are shown in Table 1.

We may obtain a quantitative index of perceptual linearity by regressing a sequence of numbers against the coordinates of the colors in perceptual space. The squared multiple correlation measures the extent to which one can construct a weighted combination of the three dimensions that correlates perfectly with a linear sequence. We used the CIELUV Cartesian coordinates \((L^*, u^*, v^*)\) as regressors rather than assign numbers to the Munsell descriptors because Indow (1988) has shown that CIELUV space corresponds closely to Munsell space. The brightness and H+S+B scales account for all the variance in the sequence \([1,2,3,4,5,6,7]\)—to four digits of precision—whereas the bipolar scale accounts for 95% and the hue-only scale 48%. Thus, the first two scales are perceptually linear and the bipolar scale is not too far from being perceptually linear, whereas less than 50% of the variation in the hue-only scale is linear.

Eighty stimulus objects (4 Displays × 4 Scales × 5 Surfaces) were created and stored as bitmaps, permitting rapid display on a CRT. Vertical labeled coding scales to the left and right of the data object formed an integral part of the stimulus (see Figure 3 for an example). Legends were provided on both sides of the display to help avoid position bias in scale-lookup.
Figure 1. The five surfaces with their generating equations and contour display representations. The printed colors necessarily differ somewhat from the colors on our monitor; also, realizing our red, green, and blue values on another monitor is likely to produce a slightly different set of colors.

**Design and procedure.** A $4 \times 4 \times 5$ within-between design was used, with display serving as a between-participants variable and scale and surface serving as within-participants variables. Participants read instructions that stated that the objects they would see would be like maps and that the colors were used to indicate height. Participants were shown an example of a stimulus object with two superimposed white crosshairs (+), with 1-pixel thick arms in the left and the right half of the display (see Figure 3). They were instructed to compare the heights represented by the two colors under the crosshairs by pressing the left “Ctrl” or the right “Enter” key, according to which crosshair was on the higher region.

To minimize visual search effects, the pairs of crosshairs always appeared at approximately the same vertical position in the right and left halves of the display (with the vertical height varying randomly over trials). Participants were instructed to be as quick as possible without compromising accuracy. They were informed that their accuracy and response time scores would be ranked and that they would win $50, $30, or $20 if they were among the top three performers. After reading the instructions, the participants completed a training session that consisted of all possible 42 comparisons of the pairs of scale colors within one stimulus object. Each participant was given speed and error feedback at the end of the training session.

An experimental session followed the training session. Participants were randomly assigned to one of the four displays (contour, mosaic, random contour, or random mosaic). Twenty stimulus objects ($4 \text{ Scales} \times 5 \text{ Surfaces}$) were presented in random order. The randomization was restricted such that neither identical scales nor identical surfaces could appear successively. All seven levels of each scale were compared with each other pairwise, excluding identical-level comparisons, and each pair appeared twice, once in each
left–right arrangement, for a total of 42 (7 × (7–1)) comparisons. Thus, each participant made 840 (4 × 5 × 42) judgments. The order of comparisons was randomized for each stimulus object. At the end of the session, the purpose of the study was explained to each participant.

**Results**

Latencies for individual color pairs were used to measure speed of responding. Responses were scored 1 or 0 according to whether the correct member of the pair of colors had been chosen. The response time and error data were subjected to a four-way 1-between–3-within analysis of variance (ANOVA; Display × Surface × Scale × Pairs). All significant effects were explored using Student Newman-Keuls multiple comparison tests, with a .05 level of significance.

**Response time.** Two patterns were evident. First, the main effect of scale was highly significant, $F(3, 48) = 82.69, MSE = 3.76, p < .0001$, with the brightness scale producing the fastest response times. The brightness scale led to significantly faster response times (1.31 s) than the H+S+B scale (1.44 s), the bipolar scale (1.66 s), and the hue-only scale (1.93 s), respectively, with all scales differing from each other (shown in Figure 4), Student Newman-Keuls, $p < .05$.

The following interactions involving scale
Figure 3. A stimulus object with two vertical legends to the left and right and two superimposed crosshairs, indicating the colors to be compared. The printed colors necessarily differ somewhat from the colors on our monitor; also, realizing our red, green, and blue values on another monitor is likely to produce a slightly different set of colors.

were significant: Scale $\times$ Surface, $F(12, 192) = 18.57, MSE = 1.51, p < .0001$, and Scale $\times$ Surface $\times$ Display, $F(36, 192) = 12.67, MSE = 1.51, p < .0001$. These interactions were one or two orders of magnitude smaller than the scale effect itself. A graph of the larger of these interactions (Scale $\times$ Surface) exhibited some nonparallelism, but the ordinal pattern of response times for the scales was the same across all five surfaces.

The surface effect was significant, $F(4, 64) = 67.00, MSE = 0.78, p < .0001$. The five surfaces were responded to at different speeds ranging from 1.42 s to 1.73 s. The Surface $\times$ Display interaction was significant, $F(12, 64) = 16.05, MSE = 0.78, p < .0001$; however, the mean square was about one quarter of the surface effect, and it appears that most of the nonadditivity was due to slightly anomalous performance with the function contour display and Surface 3.

The only remaining significant effect was due to pairs, $F(41, 656) = 49.53, MSE = 0.56, p < .0001$. All interactions with pairs were of the same order of magnitude as the corresponding appropriate error terms; some were significant largely because of the large numbers of degrees of freedom involved, but none was of substantive interest or importance. Even the main effect of pairs is of little interest and was, of course, expected.

Errors. It is noteworthy that participants were extremely accurate, making an average of only
1.26 errors in the 42 comparisons made with each stimulus. The scale effect was by far the largest, $F(3, 48) = 8.07, MSE = 198.26, p < .0005$. The most accurate responses were made with the H+S+B scale (0.70 errors), which did not differ statistically from the brightness scale (0.78 errors). However, participants made significantly more errors with the bipolar scale (1.59 errors) and the hue-only scale (1.96 errors), which differed from each other (shown in Figure 4), Student Newman-Keuls, $p < .05$. The scales that required the longest response times (the bipolar and hue-only scales) also produced the greatest number of errors.

The following interactions with scale were significant: Scale × Surface, $F(12, 192) = 3.56, MSE = 59.20, p < .0001$, and Scale × Surface × Display, $F(36, 192) = 1.83, MSE = 59.20, p < .05$. Although significant, these interactions accounted for a very small proportion of variance, and their mean squares were one to two orders of magnitude smaller than the scale effect. The surface effect was significant, $F(4, 64) = 7.96, MSE = 57.08, p < .0001$, with Surface 2 producing more errors (1.84) than the other four surfaces (0.89 to 1.38 errors), Student Newman-Keuls, $p < .05$. The Surface × Display interaction was not significant, but again, the graph shows slightly anomalous performance with the function contour display and Surface 3. The pairs effect was significant, $F(41, 656) = 6.29, MSE = 67.70, p < .0001$; this was expected and is of little interest. All interactions involving pairs were of the same order of magnitude as their corresponding error terms.

**Discussion**

On the basis of the principle of perceptual linearity, we expected that the brightness scale and the H+S+B scale would perform better than the bipolar scale and that all three would outperform the hue-only scale. The results were consistent with our hypothesis. In terms of response times, all scales were significantly different from one another, with the brightness scale producing the fastest response times, followed very closely by the H+S+B scale, with the bipolar and hue-only scales trailing badly. In terms of accuracy, the H+S+B and the brightness scales were not significantly different from one another, and both produced fewer errors than the bipolar and hue-only scales. As predicted, participants were consistently slower and less accurate with the perceptually nonlinear scales (the bipolar and hue-only scales).

Although accuracy differences are small and performance is good with even the worst scales, the better scales (those closest to perceptual linearity) have much shorter response latencies—the brightness and H+S+B scales are responded to about 50% faster than the hue-only scale. This
speed advantage reflects ease of use. In data analysis, speed is extremely important: There are many implicit comparisons to be made when examining a scientific visualization, and the data analyst generally operates under time pressures of several sorts (see Lewandowsky & Spence, 1989). The combination of speed and accuracy in the brightness and H+S+B scales confers a significant advantage. Although the bipolar scale was almost as fast, the higher number of errors made with this scale renders it less attractive. It is possible that a bipolar scale whose central elements varied in saturation alone might have performed better.

Perhaps surprisingly, participants did not perform worse with random displays, suggesting that they did not make use of structure (the shape of the surface) to help determine the height of a point on the display when making paired comparisons. This failure of participants to take advantage of the structure of the surfaces was probably peculiar to simple tasks. Nevertheless, some surfaces produced slightly better overall performance (both in terms of response time and errors) than others, but there was no important change in the pattern of relative scale performance. Because the random and function surfaces produced similar responses with negligible nonadditivities, the small performance differences with different surfaces are probably the result of differing distributions of the seven coding levels for each surface rather than differences in shape.

The main effect of pairs and all significant interactions with this variable are of minor interest. We expect considerable variance to be associated with the different pairs because some are perceptually close and others distant. Furthermore, the significant high-order interactions account for a very small proportion of variance.

Our results are in good accord with the principle of perceptual linearity and with some findings in the literature. Better performance with the brightness scale and the H+S+B scale is in partial agreement with Mersey (1990), who found that spectral scales combined with changes in brightness were favored. The performance of the brightness scale is also consistent with results obtained by Merwin and Wickens (1993), Levkowitz and Herman (1992), and Lewandowsky et al. (1993). The shorter average latency associated with the brightness scale is also in accord with the Antes and Chang (1990) finding of shorter fixation times with monochrome maps.

Our results are also compatible with the assumption that ease of ordering of visual categories can affect performance (Herrmann & Pickle, 1994). Scales that are perceptually linear and hence easily ordered (such as the brightness scale and the H+S+B scale) give the user an advantage in scale comprehension and legend-display integration. In contrast, scales that are perceptually nonlinear possess no apparent natural order and probably require the user to refer more frequently to the legend, resulting in longer response times and lower accuracy, as we found with the hue-only scale.

Although this experiment has demonstrated the superior performance of perceptually linear assignments, it must be remembered that the participants’ task was elementary—to compare the heights represented by two color patches. In the real world of data displays, the observer’s principal task is to discover structure, identify clusters, or detect anomalous regions. The ability to discriminate and estimate levels rapidly and accurately will undoubtedly be helpful in more complex tasks, but other factors may assume greater importance.

If a task requires identifying, ordering, segregating, or clustering regions, variation in hue will probably confer an advantage. Hence, in Experiment 2, we examined the participants’ ability to find and select the highest and lowest regions in a display. The task was essentially one of cluster identification and is similar to noting a region of low rainfall on a weather map, discovering an area of increased blood flow in a PET scan, or detecting a region of high stress in the visualization of a metal component under test.

**Experiment 2**

This experiment was designed to evaluate performance on a complex task (location and selection of a high or low region) using four coding schemes varying in perceptual linearity.

**Method**

**Participants.** Twenty undergraduate psychology students at the University of Toronto (15
women and 5 men) volunteered to participate. They received course credit for their participation, and, as an incentive to be quick and accurate, the top three performers were rewarded for their efforts. Performance on a screening test using pseudo-isochromatic plates (Pseudo-Isochromatic Plates, 1986) indicated that none of the participants had any color vision anomalies. The experimental session lasted approximately 30 min.

**Stimuli.** The stimuli were the same as in Experiment 1.

**Design and procedure.** A $2 \times 4 \times 4 \times 5$ (Height $\times$ Display $\times$ Scale $\times$ Surface) within-participants design was used. The first variable, height, referred to the two types of cluster identification task—whether the participants were to select the highest or the lowest region. Scales, displays, and surfaces were the same as those used in Experiment 1. Participants read instructions that stated that the objects they would see would be like maps and that the colors were used to indicate height. A 1-pixel-thick circle that could be moved using the mouse was superimposed on the display (see Figure 5). Participants were instructed to position the circle so that it was located over the highest (or lowest) region. If more than one region was of equal height, they could choose either. After reading the instructions, participants completed two training sessions. Their task was to locate the highest (or the lowest) region as quickly as possible without compromising accuracy. After moving the circle using the mouse, the participant pressed the left mouse button (thus recording both the position and the response latency.) Participants were informed that their accuracy and response time scores would be ranked and that they would win $50, $30, or $20 if they were among the top three performers. Each participant was given speed and error feedback at the end of each training session.

The training sessions were followed by two experimental sessions, one for the highest and one for the lowest region identification. The two sets of 80 stimulus objects (4 Displays $\times$ 5 Surfaces $\times$ 4 Scales) were presented in random order; thus, each participant made 160 judgments. The randomization was restricted such that neither identical scales nor identical surfaces could appear in sequence. After the two experimental sessions, the purpose of the study was explained to each participant.

**Goodness of region selection.** After data collection, the goodness of each response was assessed using the following procedure:

1. A grid of 81 equally spaced points was superimposed on the circular area chosen by the participant. All 81 points fell within the circle, with the outermost points just inside the circumference of the circle (at least 1 pixel distant from the circle).
2. The height of the surface was evaluated at each of these 81 points and then the mean height was calculated over all 81 points.
3. An estimate of the probability distribution of average heights for similar circles located elsewhere in the display was obtained by repeating the computations described in Steps 1 and 2 above when the circles were located at all possible 121 equally spaced positions of an $11 \times 11$ grid that spanned the display.
4. The percentile position of the participant's choice was obtained by reference to the estimated distribution of average heights.

If the participant obtained a 100th-percentile score after attempting to choose the highest area in the display, this meant that it was not possible to choose another circular area with higher average height; the participant's choice was optimal. Conversely, a 0th-percentile score would imply that the participant had made the worst choice, having selected the lowest rather than the highest region.

**Results**

The response time data and percentile scores for each judgment were subjected to a $2 \times 4 \times 4 \times 5$ (Height $\times$ Display $\times$ Scale $\times$ Surface) within-participants ANOVA. All significant effects were explored further with the Student Newman-Keuls multiple comparison procedure, using a uniform .05 level of significance.

**Response time.** The main effect of scales was highly significant, $F(3, 57) = 6.49, MSE = 4.13, p < .0007$, with the H+S+B scale producing the fastest response time (3.60 s), followed by the bipolar scale (3.83 s), the brightness scale (3.87 s), and the hue-only scale (4.04 s), as shown in Figure 6. The latency for the H+S+B scale
Figure 5. A stimulus object with two vertical legends to the left and right and superimposed circle that the participant must move to the highest (or lowest) region. The printed colors necessarily differ somewhat from the colors on our monitor; also, realizing our red, green, and blue values on another monitor is likely to produce a slightly different set of colors.

differed from the other scales, which in turn did not differ, Student Newman-Keuls, $p < .05$.

The type of display had a large effect $F(3, 57) = 34.18, MSE = 69.19, p < .0001$. Participants required the least amount of time with the contour display (2.18 s) and the mosaic display (2.55 s), which did not differ statistically from each other but differed from the random-contour display (5.06 s) and random mosaic display (5.56 s), Student Newman-Keuls, $p < .05$. Three 2-way interactions were also found: Height $\times$ Scale, $F(3, 57) = 3.27, MSE = 2.77, p < .03$, Scale $\times$ Surface, $F(12, 228) = 4.49, MSE = 4.59, p < .0001$, and Surface $\times$ Display, $F(12, 228) = 7.10, MSE = 6.49, p < .0001$.

The main effect for surface was also significant $F(4, 76) = 5.62, MSE = 4.64, p < .0005$. Participants required the least amount of time for Surface 2 (3.63 s), Surface 5 (3.66 s), and Surface 1 (3.82 s), which did not differ statistically from each other but, in turn, differed from Surface 3 (4.00 s) and Surface 4 (4.08 s; Student Newman-Keuls, $p < .05$). There were three other significant higher order interactions. Each was of the same order of magnitude as the error variance and, in total, accounted for a very small proportion of variance.

Goodness of region selection. The hue-only scale produced a lower percentile score (92%) than the other three scales (H+S+B = 95%, Bipolar = 95%, Brightness = 94%), which did not differ statistically from each other, $F(3, 57) =$
5.83, \( MSE = 172.30, p < .002 \), Student Newman-Keuls, \( p < .05 \) (see Figure 6). It is of interest to compare the four scales in terms of the proportion of poor responses made by participants; generally, participants performed very well on the task, and the vast majority of their choices produced percentile scores in excess of 95%. We arbitrarily divided the participants’ percentile scores into two groups: those responses above the 90th percentile and those below. If a response lay below the 90th percentile, then the participant’s choice of region was noticeably less than optimal. The proportions of these poor responses with each of the scales were as follows: .11 for brightness, .12 for bipolar, .11 for H+S+B, and .19 for hue only. The hue-only scale thus yielded almost twice as many poor responses, Pearson \( \chi^2(3, N = 3,200) = 29.6, p < .0001 \), than the other scales, which did not differ statistically from each other.

The type of surface had an effect on the participants’ ability to select high or low regions, \( F(4, 76) = 3.61, MSE = 137.85, p < .01 \), but although performance varied between approximately 93% and 95%, distinct groupings were not found using the Student Newman-Keuls procedure at the .05 level. The type of display had a large effect, \( F(3, 57) = 111.58, MSE = 215.01, p < .00001 \); participants were most accurate with the contour display (99%) followed by the mosaic display (97%), and these two displays differed from the random-contour display (94%) and random-mosaic display (87%), Student Newman-Keuls, \( p < .05 \). The effect of display on goodness of region selection was similar to its effect on latency; displays requiring the longest response times (the random contour and random mosaic) also produced the lowest goodness of selection percentile scores.

Several interaction effects were significant, but most of these accounted for very small proportions of variance and exhibited negligible nonadditivity—the large number of associated degrees of freedom conferring considerable power. Of these interactions, the following were most important: Scale \( \times \) Display, \( F(9, 171) = 8.25, MSE = 159.70, p < .0001 \), Scale \( \times \) Height, \( F(3, 57) = 8.12, MSE = 357.58, p < .0004 \), and Height \( \times \) Surface, \( F(4, 76) = 5.25, MSE = 140.91, p < .001 \). The Scale \( \times \) Display interaction was significant largely because of the extremely poor performance of the hue-only scale with the random mosaic display (80%), whereas performance with the other scales on the random displays was around 90% and above 98% for the other (nonrandom) displays. The Scale \( \times \) Height interaction was similarly due to the performance of the hue-only scale being poorer than expected when participants had to find the lowest region. Examination of the Height \( \times \) Surface interaction means showed that there was more variation in perfor-
mance across the different surfaces when participants were trying to find the highest region than when trying to find the lowest region.

**Discussion**

The hue-only scale was the worst for the identification of the highest and the lowest regions of the display. There were also interactions of height and display with scale in which performance with the hue-only scale was even worse than would be expected on the basis of an additive model. This idiosyncratic performance of the hue-only scale probably reflects confusions made by the participants between the relatively similar colors (blue and a bluish purple) at the opposite ends of the scales. In terms of latency, the H+S+B scale led to significantly faster response times than the brightness, bipolar, and hue-only scales. Thus, the scales that varied in hue were both the best and the worst—the crucial differentiating aspect was whether or not brightness and saturation were covaried with hue. Hence, the results are partially consistent with the principle of perceptual linearity, but hue plays a more important role than it does with simple tasks.

Unlike in Experiment 1, participants in Experiment 2 made use of the structure of the display. Random displays produced much poorer and slower performance. The task is more difficult with random displays in which any clustering is an accident of the random process used to create the display. However, the same general pattern of performance resulted with both the random and function displays. Although a well-defined structure enhances the speed and accuracy of cluster detection, the relative effectiveness of the scales is similar to that observed when the display contains no regular or well-defined spatial structure. This gives us greater confidence that participants have not somehow used whatever partial knowledge of structure they may have been able to extract to help identify extreme regions and that the effects observed are due to coding scheme alone.

According to Herrmann and Pickle (1994), perception of the highest and the lowest regions requires an impression of the dominant rate on the display. They predicted that it would be easier to distinguish adjacent scale categories for monochromatic gray than multiple-hue scales because of the ease of identifying adjacent categories with monochromatic scales. Lewandowsky et al. (1993) also found an advantage for a monochromatic scale compared with multiple-hue and bipolar scales for cluster identification. In contrast, our study suggests that the highest and the lowest areas are identified more rapidly with a particular type of multiple-hue scale (the H+S+B scale) than with brightness, bipolar, or hue-only scales. The H+S+B scale is also most accurate, although not significantly more so than the brightness and bipolar scales. The advantage of the H+S+B scale for cluster identification is probably due to the fact that this scale varied not only in brightness and saturation but also in hue. Two dimensions (brightness and saturation) establish a natural order (perceptual linearity), and the third (hue) assists in segregation; the use of hue is highly effective for segregation and classification. Variation in hue in the H+S+B scale is the key element that confers an advantage over the brightness scale when the task requires segregation as well as the determination of magnitude. Covarying saturation and brightness has the added benefit of maximizing the differences among hues and helps minimize simultaneous and chromatic contrast effects.

**General Discussion**

Both experiments confirmed our hypothesis that performance varies with perceptual linearity. As predicted, brightness and H+S+B scales supported better performance than the bipolar scale and hue-only scales. Furthermore, the linear scales were superior to the nonlinear scales on both simple and complex tasks. However, the bipolar scale, which had a fairly high index of perceptual linearity, performed much better than the hue-only scale.¹

Although we prefer to account for the good performance of three of the scales largely in terms of the principle of perceptual linearity, our findings are compatible with other explanations. The excellent performance of the brightness scale is in good accord with some previous research (e.g., Antes & Chang, 1990; Levkowitz & Herman, 1992; Lewandowsky et al., 1993). In our study, we found that the brightness and the H+S+B scale led to approximately equal performance, with a small advantage to the latter in a complex task. Because the brightness scale is

¹ Although unlikely, it is possible that our findings are peculiar to the particular colors selected rather than the general properties of the scales.
fundamentally linear (it does not require the observer to combine and weight dimensions), we expected it would show optimal performance. The H+S+B scale, which is perceptually linear, may be a better overall choice because the incorporation of variation in hue gives it an edge when cluster detection is required. However, if the observer does combine and weight dimensions, as we have suggested, there may be a small speed penalty. Ware's (1988) experimental sequence is similar to our H+S+B scale, but without monotonic variation in saturation; Ware found that his experimental scale produced fewer errors than brightness, saturation, and red–green scales. Our findings are somewhat at variance with Hastie et al. (1995), who found that spectrum-based scales produced better results when compared with a monochrome scale; however, their rainbow scales and our hue-only scale are not comparable because they did not hold brightness and saturation constant.

There is general agreement on the beneficial role of hue in more complex tasks. This is almost certainly due to the fact that the visual system preattentively forms color categories, despite the fact that the visible portion of the electromagnetic spectrum varies smoothly, without discontinuity. Our ability to differentiate regions of different wavelength effortlessly allows us to make absolute judgments of color rapidly and with high reliability (Berlin & Kay, 1969; Boynton, 1989; Sturges & Whitfield, 1995). These abilities are undoubtedly useful when clusters must be detected in spatial displays.

The bipolar scale, which has been proposed as an alternative to monochromatic and rainbow scales (especially when deviations from the average are of interest), was not as effective as in some previous work (Carswell et al., 1995; Lewandowsky & Behrens, 1995; Petchenik, 1983). These studies may have used bipolar scales that were closer to perceptual linearity—perhaps varying only saturation, not both as in our experiments. Alternatively, the difference could be partly due to the fact that we compared only one bipolar scale with three different types of scales, whereas Carswell et al. and Lewandowsky and Behrens compared them with only one or two. Carswell et al. compared three bipolar scales to one monochromatic gray scale, and Lewandowsky and Behrens did not include a monochromatic scale in their study. However, because Lewandowsky et al. (1993) and Lewandowsky and Behrens (1995) did not specify the color coordinates used in their studies, it is difficult to make a direct comparison.

As predicted, the nonlinear hue-only scale led to the worst performance on both tasks. Use of this type of scale should be avoided. Instead, scales with approximately equal discrimination steps within a single hue or among hues covarying in brightness or saturation should be used. An added benefit is that even those with congenital or acquired color vision anomalies will be able to detect the changes in level and do so in the correct order.

If hue is chosen to represent quantity, the steps between hues should be combined with changes in brightness and saturation. If a single hue, two hues, or multiple hues are chosen to code numbers on an equal interval scale, brightness and saturation should be varied in such a manner that a perceptually linear and equally spaced sequence is created. Linear variation in brightness and saturation facilitates simple tasks such as magnitude estimation or paired comparisons, and the addition of hue enhances performance with more complex cognitive tasks.

References


Boynton, R. M. (1989). Eleven colors that are almost


Received January 20, 1998
Revision received February 12, 1999
Accepted February 22, 1999