Perception, 2002, volume 31, pages 445 – 461

DOI:10.1068/p3134

The effect of name category and discriminability on the search characteristics of colour sets

Steve Guest
Department of Experimental Psychology, Oxford University, South Parks Road, Oxford OX1 3UD, UK; e-mail: Steve.guest@psy.ox.ac.uk

Darren Van Laar
Department of Psychology, University of Portsmouth, King Henry I Street, Portsmouth PO1 2DY, UK; e-mail: Darren.van.laar@port.ac.uk
Received 18 September 2000, in revised form 14 September 2001

Abstract. Within (and between) cultures, people tend to agree on which parts of colour space are easiest to name and what the names for these regions are. Therefore it is likely that the manipulation of ease of naming (nameability) of colours should change performance in tasks where categorisation by colour name is important. More specifically, highly 'nameable' colour sets should lead to better performance than metrically equivalent but less categorically distinct sets, when the task requires categorisation. This hypothesis was investigated by testing observers on a name-based task, the naming and subsequent identification by name of colour sets with up to sixteen members. These sets were designed to be easy to name (nameable), maximally discriminable, or matched discriminable. The first were derived from previously generated data, the second by a standard algorithm to space colours widely in colour space, and the latter by closely matching their metric characteristics to those of an easy-to-name colour set. This final condition was metrically (but not categorically) equivalent to the nameable set. It was found that sets designed to be nameable did indeed lead to superior performance as measured by response times, confidence ratings, and response accuracy. Perceptual colour similarity, measured by a ΔE metric, did not predict errors. Nameability may thus be a valid, manipulable, aspect of sets of colours, and one which is not otherwise duplicated in the metric characteristics of such sets.

1 Introduction

Colour is well established as a method of enhancing information displays (eg Christ 1975). An enhancement often reported is in improving the search for information. Typically research in this area has concentrated on the colour-difference requirements for obtaining parallel search within colour sets, and the adequacy of standard colour-difference metrics (eg ΔE^*_{uv}) in providing useful measures to predict parallel search (eg Carter and Carter 1982; Luria et al 1986; Sanchez and Nagy 1989).

Less research has been directed towards the role of nameability (ease of naming of colours) in altering the search characteristics of colour sets. This is perhaps surprising since much research has determined within and between-culture agreement in what the fundamental names ('basic colour terms' or BCTs) are, and what sensations they denote (Berlin and Kay 1969; Rosch-Heider 1972; Kay et al 1997). Most research has determined a maximum of eleven such terms—white, black, red, green, blue, yellow, pink, purple, orange, brown, grey—and it has been proposed that these derive from a fundamental structure of the visual system shared between all humans (Ratliff 1976; Boynton 1997). Such universals suggest that any results from visual-search experiments involving colour naming could be stable, and widely usable.

Given the above data, the question arises as to why only limited effort has been directed towards the use of nameable colours in search (and other) tasks. In part this is because defining 'easy to name' is not itself easy, unlike defining perceived colour difference (eg ΔE^*_{uv}). There are no analogous systems to allow a ΔE equivalent for colour 'nameability'. Nor is there a naming framework which allows statements such as 'this green is a better green than that colour is a red', or 'these two greens are more likely to be confused together than those two reds', to be made.

Another reason why little research has been directed at the role of colour names in search tasks is because, for many search tasks, categorisation may be unnecessary. For instance, no colour-spacing algorithms include, or seem to require, categorical difference for optimal search performance (eg Carter and Carter 1982; De Corte 1990). However, describing a colour by name requires a categorical judgment, such as when directing an observer's attention to the 'purple object' on a computer screen, or when cueing a target group or class by a colour name. This latter task is, at least initially, a categorical one, although once the target subgroup has been identified from the name the subsequent search might no longer be categorical. It is these sorts of tasks where nameability could be important.

The primary research into ease of naming of colours (or colour 'nameability') as a variable in visual search has been that of Smallman and Boynton (1990, 1993). In the first study they investigated the usefulness of basic colour coding in a visual-search task, with basic colours defined via colour-naming data from Boynton and Olson (1987). In these studies, a desired target group was coded in a (basic) colour, and this group was subsequently cued by an example or a colour name. It was found that basic codes did indeed lead to excellent search performance. However, this performance was no better than when similarly metrically spaced nonbasic exemplars were used. The conclusion was that any search efficacy of basic colour sets was only because the basic exemplars were widely separated in colour space, though it must be noted that for the nonbasic parts of the experiment, no name-based cueing was used. The later Smallman and Boynton study used a more interactive design, with observers navigating colour space to set basic and 'personal' (ie nonbasic) exemplars for use in a subsequent search task. Results were essentially in line with the 1990 study, although there was a suggestion that initial responses to names given by others were relatively slow (when cued by name), an effect which disappeared rapidly with practice.

On the basis of previous knowledge of colour naming, and given the nature of tasks used, these results are perhaps unexpected. Despite previous findings which suggest basic foci as preeminent in (especially) memory tasks (Rosch-Heider 1972; Boynton et al 1989) and other measures of 'salience' (Whitfield 1981; Boynton and Olson 1990), these findings were not echoed in search paradigms. However, there are certain aspects of the name/search experiments which may have acted to disguise any, perhaps subtle, differences between responses to basic versus nonbasic exemplars.

First, the colour naming and search papers noted above used standard two-stage visual-search paradigms. That is, cueing to a target class, then identification of a target within that class. If the translation of the cue (ie identification of what was meant by the name) was relatively easy compared to the actual search task, then any effects of name type could be hidden within response variability in the search phase.

Second, and more importantly, the precise metric properties of the colour sets tested by Smallman and Boynton remain unclear. Neither the subjective bisection procedure of Smallman and Boynton (1990), nor the self-setting of colour palettes (Smallman and Boynton 1993) explicitly match metric characteristics between categorically distinctive and categorically indistinct sets. Furthermore, it is unclear how metrically different the personal and basic colour sets were in the 1993 study. That is, the two colour sets produced by subjects may have been very similar both in categorical and in metric terms. Or perhaps the personal set was more effectively generated and contained larger within-palette colour differences. The only certainty is that subjects were satisfied with their colour sets, regardless of the metric characteristics, which were not listed in the paper.

It is also true that even if the colour sets mentioned above were metrically equivalent, any precise role of colour difference (metrics) within categorical search remains uncertain. For instance, it has been established that for standard visual-search paradigms, the smallest pairwise colour difference present within the search set establishes its search efficacy (Carter and Carter 1982). For categorical tasks, such a statement cannot be made. It remains unknown which metrics—if any—are important when category-based tasks are undertaken. Moreover, if the findings of Smallman and Boynton were upheld with precise metric controls, this would verify that the metric properties alone are sufficient to determine the performance of a colour set in category-based tasks.

1.1 *Multiple category members*

Even if coding by basic names was in some sense optimal, it would allow the specification of only eight chromatic and three achromatic colours at most, ie the eleven basic colours. If one required (say) fifteen codes, then there are not enough BCTs to allow this.

Fortunately, the work of Smallman and Boynton suggests that 'optimally nonbasic' colours (colours spaced equally between basic foci) do have a place in the generation of nameable colour sets. These nonbasic exemplars almost certainly correspond to the specific nonbasic regions of colour space noted by Boynton and Olson (1987), Sturges and Whitfield (1995), and most recently Guest and Van Laar (2000). Such areas are portions of colour space where nonbasic terms tend to be used frequently by observers. This may be contrasted with the rest of colour space where basic naming predominates. Thus the use of colours from the nonbasic portions of space may still assist categorical tasks since the colours sampled do have some categorical distinctiveness even though the naming of these regions may be relatively slow, less confident, and inconsistent (Guest and Van Laar 2000). Indeed Smallman and Boynton suggested that this relative categorical distinctiveness could be why nonbasic examples were efficient search cues. Alternative colour sets without this property (eg inclusion of multiple and/or poor examples of the basic exemplars) might lead to degradation of performance as the categories begin to lose their distinctness. For example, a colour set including two (different) examples of green would facilitate more categorical confusions than a colour set with one example of green and one of yellow, even if there was an equivalent perceptual difference between the colours in the two cases.

An experiment to test rigorously the value of nameability and basic (and nonbasic) naming as devices to enhance search tasks needs careful design. In particular, ease of naming should be manipulated independently of other parameters of colour sets, such as perceptual colour difference. To succeed in this, we use ideas similar to those seen in the work of Smallman and Boynton, but with explicit control of the metric properties of colour sets.

It was hypothesised that performance in a task requiring identification of colour categories would be much improved when colours were designed to be nameable, versus when metrically similar sets not specifically chosen to be easy to categorise were used, or when compared with highly discriminable colour sets. It was further hypothesised that the types of errors made in the search task would be only weakly linked to the CIE colour-difference metric (ΔE^*_{uv}), being more strongly determined by name-category boundaries, which are poorly captured by standard metric measures.

2 Experiment 1

Stimulus displays were sets of colours displayed as rectangular matrices of coloured squares. There were three types of colour set—highly nameable, highly discriminable, and ΔE-matched discriminable. Each of these types were produced in twelve-member and sixteen-member versions, giving six different colour sets in total. These are described in more detail later. The experiment was divided into two phases. In the first phase, name generation, observers were asked to provide a colour name for every member of three colour-set types, for either twelve-member or sixteen-member sets. Each observer provided a set of names for each colour set they saw, ie each colour set had multiple name sets, each set generated by a different observer. In the second phase, name response, observers were required to indicate which member of a colour set was indicated by a displayed colour name. This name was either one generated by themselves, or by another participant, although the two possibilities were never mixed within a subsession. For practical reasons the naming phase always preceded the identification phase.

The design was a colour-set type (nameable, maximally discriminable, ΔE-matched discriminable) $\times$ set size (twelve or sixteen members) $\times$ name types (own or others' names) factorial. These factors will hereafter be termed Colour Set, Size, and Names. Colour Set and Names were within-subjects factors, whereas Size was a between-subjects factor. When considering Names, all observers responded to their own names and one set of names from a randomly chosen different observer.

Recorded response measures were response latencies and confidence ratings for each colour selection. Confidence ratings were made on a five-point scale by respondents. In the name-response phase the accuracy of each response was also recorded.

Since each colour within each colour set was seen and responded to twice by each subject (in the name-response phase), an additional factor was available corresponding to the first and second responses to each colour by the subject. This fourth independent variable (first or second response) is termed Trial. All other aspects of the design were randomised, including stimuli presentation orders.

2.1 *Stimuli*

All stimuli were presented on an NEC 6FGp 21-inch colour monitor, calibrated and driven by custom software running on an IBM-compatible PC. There was no ambient illumination.

The screen display consisted of a rectangular stimulus matrix against a dark screen background of luminance 2.7 cd m^{-2}. There was a 1 deg wide border around the edge of the screen, of luminance 63.0 cd m^{-2}. Both background and border had chromaticity coordinates of $u' = 0.198$, $v' = 0.468$ (ie simulated D_{65} white).

The stimulus matrix was centred on the display screen. Sixteen-colour matrices were displayed in a 4×4 array, those with twelve members in a 3×4 configuration. The matrix types and their production are described below. Note that the chosen Y_n (reference white luminance) for all colour-difference calculations followed the convention suggested by Post (1992), of being the on-screen reference white, which in this case was the white screen border.

2.1.1 *Nameable colour sets.* The members of these sets were derived from previously determined colour-naming data, as described in Guest and Van Laar (2000). In summary, naming data was available for 348 stimulus colours, split across three discrete stimulus luminance levels (7.6 cd m^{-2}, 16.6 cd m^{-2}, and 27.4 cd m^{-2}), presented against a background of luminance 2.7 cd m^{-2}. 'Nameability' scores were obtained via principal-components analysis of response-latency, confidence, and consistency data, derived through free naming of these colours. Colours were ranked by nameability, and the highest nameability occurrence of a particular modal colour name, for a given colour sample, was taken as the best example of that name. In other words, the name category was the criterion for selection, not the ΔE differences between colours. This procedure was repeated until the required number of colours (ie required number of different modal names) had been obtained. Therefore, the nameable twelve-colour set was a subset of the sixteen-member version. The technique is limited in the number of colours it can produce by the number of different modal names available, although it could be possible to extend the size of sets by considering second-most-popular names. Note that some of the categories within generated sets were nonbasic (eg 'khaki'), but these were still categorically distinct.

2.1.2 *Matched discriminable sets.* These sets were designed to have very similar metric characteristics to their equivalent-sized nameable set, but without the 'nameable' constraint (see table 1 and table 2 for comparison data). That is, colours were chosen purely on the basis of metric considerations. An algorithmic approach was taken to match the overall mean ΔE^*_{uv} for the colour set, the minimum ΔE^*_{uv}, and the mean of the colourwise minimum ΔE^*_{uv} [abbreviated C-$\overline{\min}(\Delta E^*_{uv})$]. C-$\overline{\min}(\Delta E^*_{uv})$ is a measure of the average smallest perceptual difference between colours in a set. It is calculated by considering each member of a colour set in turn, and finding the most perceptually similar colour by comparison with the other colour-set members, ie the smallest pairwise ΔE^*_{uv} difference involving the first colour-set member is found. This procedure is repeated for each colour-set member. These colour-difference values are then averaged to obtain an overall mean value for the colour set. It should be noted that data from the same colour pair within a colour set may be entered more than once in the calculation of the final value. For example, if colour 1 and colour 5 are the most similar during the comparison sequence for colour 1, then it is possible (but not certain) for the same pair of colours to produce the minimum ΔE^*_{uv} for the comparison sequence for colour 5. A variant ofC-$\overline{\min}(\Delta E^*_{uv})$ which disallows such replication can be generated, although such a statistic is not reported here. The minimum ΔE^*_{uv} in the set, and the C-$\overline{\min}(\Delta E^*_{uv})$, were matched very precisely, but the overall average ΔE^*_{uv} was matched slightly less well. Perfect matching was not feasible within the constraints of the (linear optimisation) algorithm used. Since the minimum colour-difference present within a set has been shown to be the statistic which best determines search efficacy (Carter and Carter 1982), the close matching of this—and the conceptually related C-$\overline{\min}(\Delta E^*_{uv})$—was favoured over precisely matching the overall mean ΔE^*_{uv}.

2.1.3 *Maximally discriminable sets.* These sets were chosen such that the overall minimum pairwise ΔE^*_{uv} within the set exceeded that generally required for parallel visual search, ie a ΔE^*_{uv} of at least 40. The algorithm used to generate these sets was that of Carter and Carter (1982). The overall mean ΔE^*_{uv} and C-$\overline{\min}(\Delta E^*_{uv})$ were not explicitly manipulated. Values of the mean ΔE^*_{uv} were comparable with those for the other set types, whilst values of C-$\overline{\min}(\Delta E^*_{uv})$ exceeded those in the other sets. The minimum ΔE^*_{uv} value in the maximally discriminable sets was around twice the minimum seen in the nameable and matched sets, for both set sizes. These statistics illustrate that in a traditional visual-search task, one would expect this colour set to be superior to both the nameable and matched sets.

For matched and maximally discriminable sets, the two required set sizes were generated independently of each other. Both matched and maximal sets were restricted in lightness to a 30 to 75 range, comparable with the nameable sets (which were restricted to discrete lightnesses of 41.4, 58.3, or 71.8 since naming data were available for only these values).

Summaries of ΔE^*_{uv}-difference data for the three palette types are presented in table 1 and table 2, while the colorimetric coordinates for all colour sets are shown in table 3 and table 4. Table 4 also includes a list of the expected colour names for the nameable stimuli, which is based on the prior work noted above.

2.2 *Participants*

A total of twenty-two participants were obtained after screening for colour-vision defects with the Ishihara test. Two groups of eleven participants were selected, corresponding to the two experimental size conditions. The first participant in each of the

Table 1. Summary colorimetric difference data for the twelve-member colour sets.

	Colour Set type		
	nameable	matched	discriminable
Minimum ΔE^*_{uv}	29.5	30.0	49.3
Overall mean ΔE^*_{uv}	118.9	105.0	103.9
C-$\overline{\text{min}}(\Delta E^*_{uv})$	41.8	41.1	55.5

Note: C-$\overline{\text{min}}(\Delta E^*_{uv})$ is a measure of the mean minimum pairwise colour difference for a colour set (see main text for details).

Table 2. Summary colorimetric difference data for the sixteen-member colour sets.

	Colour Set type		
	nameable	matched	discriminable
Minimum ΔE^*_{uv}	20.3	21.0	49.4
Overall mean ΔE^*_{uv}	118.5	112.4	114.5
C-$\overline{\text{min}}(\Delta E^*_{uv})$	41.7	46.1	55.9

Note: C-$\overline{\text{min}}(\Delta E^*_{uv})$ is a measure of the mean minimum pairwise colour difference for a colour set.

Table 3. Colorimetric coordinates (u', v', L^*) for the twelve-member colour sets.

Member	Colour Set type			
	matched	maximal	nameable	
1	(0.247, 0.435, 49.0)	(0.202, 0.391, 80.1)	blue:	(0.160, 0.260, 58.3)
2	(0.345, 0.492, 58.2)	(0.140, 0.409, 56.9)	green:	(0.140, 0.540, 58.3)
3	(0.343, 0.518, 39.7)	(0.208, 0.275, 59.9)	pink:	(0.340, 0.420, 58.3)
4	(0.168, 0.225, 33.0)	(0.240, 0.303, 30.5)	orange:	(0.340, 0.520, 58.3)
5	(0.203, 0.259, 56.9)	(0.312, 0.506, 69.1)	red:	(0.380, 0.500, 41.4)
6	(0.269, 0.482, 52.1)	(0.350, 0.500, 33.7)	grey:	(0.200, 0.440, 58.3)
7	(0.221, 0.498, 53.3)	(0.145, 0.355, 74.9)	purple:	(0.220, 0.260, 58.3)
8	(0.272, 0.332, 53.3)	(0.259, 0.321, 56.9)	brown:	(0.260, 0.500, 71.8)
9	(0.124, 0.536, 51.2)	(0.171, 0.199, 32.6)	yellow:	(0.240, 0.540, 71.8)
10	(0.161, 0.491, 57.5)	(0.216, 0.445, 30.5)	turquoise:	(0.140, 0.440, 58.3)
11	(0.195, 0.383, 58.1)	(0.306, 0.396, 46.3)	khaki:	(0.220, 0.540, 41.4)
12	(0.287, 0.532, 58.0)	(0.130, 0.498, 50.6)	peach:	(0.260, 0.500, 41.4)

Note: For the nameable colours, names expected on the basis of previous work are shown.

two size conditions only took part in the first (name-generation) phase, simply because there were no names given by others available until after they had been generated by the first person.

The groups consisted of seven females, four males (twelve-colour group), and four females, seven males for the sixteen-colour group. Neither sex nor age (range 17 to 40 years) were considered for analysis. All participants had normal or corrected-to-normal visual acuity.

2.3 *Procedure*

The experiment was divided into two distinct, separately run sessions; a name-generation session, and a name-response session. The former always preceded the latter.

The participant was seated in front of the experimental display and lights were extinguished. All instructions to participants were given in the dark cubicle, illuminated

Table 4. Colorimetric coordinates (u', v', L^*) for the sixteen-member colour sets.

Member	Colour Set type			
	matched	maximal	nameable	
1	(0.200, 0.460, 58.3)	(0.268, 0.308, 64.5)	blue:	(0.160, 0.260, 58.3)
2	(0.330, 0.520, 58.3)	(0.424, 0.508, 53.0)	green:	(0.140, 0.540, 58.3)
3	(0.190, 0.387, 71.8)	(0.284, 0.344, 30.5)	pink:	(0.340, 0.420, 58.3)
4	(0.252, 0.303, 66.2)	(0.268, 0.499, 31.6)	orange:	(0.340, 0.520, 58.3)
5	(0.150, 0.545, 71.8)	(0.158, 0.407, 31.2)	red:	(0.380, 0.500, 41.4)
6	(0.250, 0.530, 71.8)	(0.183, 0.489, 64.6)	grey:	(0.200, 0.440, 58.3)
7	(0.270, 0.520, 42.6)	(0.225, 0.537, 67.0)	purple:	(0.200, 0.260, 58.3)
8	(0.150, 0.550, 42.6)	(0.157, 0.267, 30.8)	brown:	(0.260, 0.500, 71.8)
9	(0.295, 0.435, 67.3)	(0.184, 0.361, 74.4)	yellow:	(0.240, 0.540, 71.8)
10	(0.152, 0.466, 71.8)	(0.298, 0.441, 58.0)	turquoise:	(0.140, 0.440, 58.3)
11	(0.219, 0.544, 58.3)	(0.134, 0.514, 71.3)	khaki:	(0.220, 0.540, 41.4)
12	(0.170, 0.295, 58.3)	(0.311, 0.512, 57.2)	peach:	(0.260, 0.500, 41.4)
13	(0.250, 0.400, 64.0)	(0.218, 0.410, 79.8)	cerise:	(0.300, 0.380, 35.0)
14	(0.170, 0.420, 71.8)	(0.141, 0.395, 68.0)	mauve:	(0.240, 0.340, 71.8)
15	(0.393, 0.507, 41.4)	(0.157, 0.235, 49.7)	light blue:	(0.160, 0.400, 71.8)
16	(0.182, 0.366, 71.8)	(0.288, 0.354, 55.8)	light green:	(0.160, 0.540, 71.8)

Note: For the nameable colours, names expected on the basis of previous work are shown.

only by the experimental display. This allowed some three to five minutes for adaptation to the viewing conditions. Some adaptation time was also available during practice trials, which were provided for both experimental phases.

2.3.1 *Name-generation phase.* Initially each participant was required to generate a set of personal names for nameable, matched, and maximally discriminable sets at the size condition they had been assigned to. The order of presentation for naming of the three colour-set types was randomised. The generation sequence is summarised below.

The sequence began with the display clear of the stimulus colours. Subsequently, a flashing block cursor with accompanying high-pitched tone directed the observer's attention to the location on the screen where the colour to be named would appear.

Once the matrix of colours appeared on screen, the observer's task was to quickly assign a name to the primed colour. The naming was unconstrained, and this was stressed to each participant. It was emphasised that naming should be fast, but without sacrificing perceived 'quality' of colour name. Following naming, a confidence rating on a 1 (low confidence) to 5 (high confidence) scale was provided by the participant. This indicated the confidence that the observer would successfully pick out that colour from that palette, when just given the name assigned. Participants were warned that duplicate names for colours within a colour set were illegal—since such a situation would clearly lead to ambiguity in the response phase—although names could be duplicated between colour sets. As soon as a response was given, the experimenter tapped the space bar—which stopped the internal timer—then entered the colour name on the computer keyboard.

Prior to making recorded responses, subjects took part in an initial practice session consisting of making eight name responses to a metrically matched colour set containing twelve members. This set was not otherwise seen in the experiment.

2.3.2 *Response phase.* During this phase the observer performed essentially the reverse task to that of the generation phase. That is, the participant attempted to select screen-presented colour samples which matched his or her own and one (randomly chosen) other's pregenerated colour names.

The appropriate colour set appeared with the stimuli randomised in position in the stimulus matrix. Approximately 1.5 s later, after a short warning tone, a colour name was displayed at the bottom of the screen and in the centre of the stimulus matrix, a mouse cursor. At this point the internal timer in the computer started.

The observer was required to select, using the mouse, which of the coloured stimuli present best exemplified the displayed name. After this selection was made, a confidence choice box, containing the numbers 1 to 5, appeared to the right of the stimulus array. The observer then made his or her confidence choice, again with the mouse. After this choice the screen cleared, and after a brief intertrial pause the response cycle continued.

Prior to the actual experimental trials, a dummy stimulus set was run (ten responses), to familiarise the subject with the procedure. These responses were not recorded. Sessions for generation and response phases took from 25 to 35 min each.

2.4 *Results*

The analysis and presentation of results have been split into generation-phase and response-phase sections.

2.4.1 *Generation phase.* During the generation of names for colour sets, very many name terms were produced; the agreement in naming any given sample within any palette was relatively low between subjects. In fact, only two colours out of the seventy-two used were given the same name by all subjects. These were the nominal 'orange' in the twelve-member nameable set and colour 1 (see table 4) in the sixteen-member matched set which was named 'grey' by all. The mean number of different names used for each colour in each set was as follows: nameable twelve-member—5.25, sixteen-member—5.94; maximal twelve-member—6.67, sixteen-member—7.13; matched twelve-member—7.42, sixteen-member—6.13.

The response time and confidence data were entered into a Colour Set (nameable, maximal, or matched) × Size (twelve or sixteen members) factorial MANOVA. The former factor was within participants, the latter was between participants.

Given the different sizes of stimulus sets, responses to twelve or sixteen colours were available. This could be considered a nested factor (ie colour types—individual stimuli—nested within the colour-set factor). However, to simplify the analysis, all responses within a design cell by a subject were averaged. Thus for the initial analyses each subject provided a single mean confidence rating and response time for each colour set.

The MANOVA revealed a significant overall effect of Colour Set on the combined response time and confidence data (Wilks' $\lambda = 0.354$, $F_{4,17} = 7.75$, $p < 0.001$). Subsequent univariate F tests revealed that both confidence ratings and response times were influenced by Colour Set (response times: $F_{2,40} = 15.17$, $p < 0.001$; confidence ratings: $F_{2,40} = 11.93$, $p < 0.001$). Scheffé tests on pairwise differences disclosed that the generation of names for nameable sets led to significantly more confident ratings ($p < 0.05$), and faster responses ($p < 0.01$) then for matched or maximal palettes. Matched and maximal sets did not significantly differ from each other on either response measure ($p > 0.05$). Neither Size (Wilks' $\lambda = 0.891$, $F_{2,19} = 1.16$, ns), nor the Colour Set × Size interaction (Wilks' $\lambda = 0.928$, $F_{4,17} = 0.331$, ns) were significant. Observers did not name colours any slower or less confidently for larger set sizes. Figure 1 shows the mean response times and confidence ratings for the three colour-set types collapsed across the two size conditions for the (a) generation and (b) response phases.

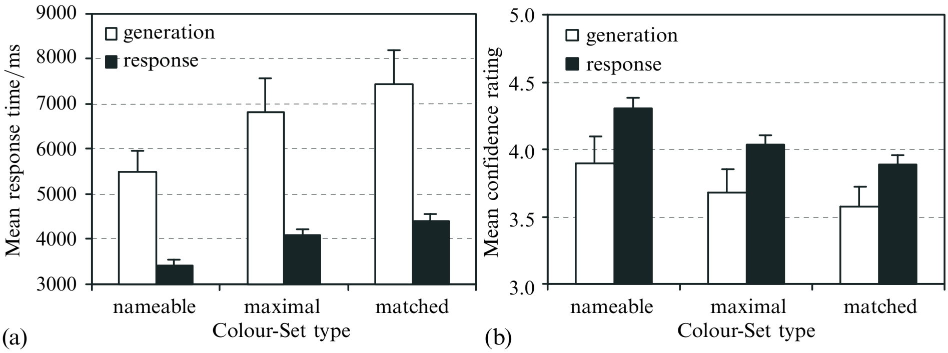

Figure 1. Mean response times (a) and mean confidence ratings (b) for the three colour-set types (nameable, maximally discriminable, matched discriminable) for the generation-phase data (white bars) and response-phase (black bars), collapsed across the two sizes of colour set (twelve or sixteen members). Error bars show +1 SE.

2.4.2 *Response phase.* Response-phase data consisted of response times, confidence ratings, and response accuracy (ie whether the observer correctly identified the colour sample from the colour name). Owing to the different nature of the accuracy data versus the response-time and confidence data, the analysis was split, with the last two measures analysed with ANOVA approaches, and the first measure analysed with log-linear analysis.

Both response-time and confidence data were separately analysed in Colour Set (3) × Names (2) × Trial (2) × Size (2) factorial ANOVAs. The first three factors were within subjects, the fourth factor was between subjects. In similar fashion to the analysis for the generation phase, data were averaged so that each subject provided a single mean confidence rating and response time for each design cell. The analysis was then divided into two subanalyses, one for response-time data, one for confidence data.

For both the confidence and response-time data, three of the four main effects were significant: Colour Set (confidence: $F_{2,36} = 25.71$, $p < 0.001$; response time: $F_{2,36} = 14.38$, $p < 0.001$), Names (confidence: $F_{1,18} = 38.03$, $p < 0.001$; response time: $F_{1,18} = 15.42$, $p < 0.001$), and Trial (confidence: $F_{1,18} = 20.46$, $p < 0.001$; response time: $F_{1,18} = 93.58$, $p < 0.001$). Mean confidence ratings and response times for the Colour Set data are shown in figure 1 (Scheffé tests revealed that the three colour-set types differed significantly); means for other factors are shown in table 5.

Table 5. Mean confidence ratings (CF, on a scale of 1 to 5) and response times (RT, ms) for the response phase, broken down by trial and name type. SEs are given in parentheses.

	Trial		Name type	
	first	second	own	other
Mean CF	4.03 (0.058)	4.13 (0.056)	4.28 (0.051)	3.87 (0.055)
Mean RT	4264 (129)	3658 (120)	3762 (126)	4160 (122)

Of the significant interactions, two were common both to response times and to confidence ratings, the Colour Set × Size interaction (confidence: $F_{2,36} = 9.54$, $p < 0.001$; response time: $F_{2,36} = 5.14$, $p < 0.05$) and that of Names × Trial (confidence: $F_{1,18} = 5.46$, $p < 0.05$; response time $F_{1,18} = 4.69$, $p < 0.05$). The former is plotted for confidence ratings in figure 2, where it appears that the interaction is due to

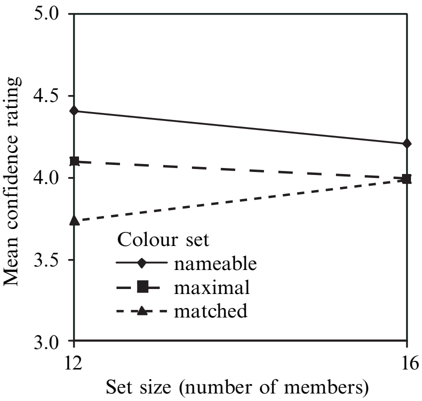

Figure 2. Interaction plot of overall mean response-phase confidence ratings for the three colour-set types (nameable, maximal, matched) against the two sizes (twelve and sixteen members).

the matched set actually having higher confidence ratings for the larger palette size (as compared with the smaller one), a pattern unique to this colour set. This was echoed in the response-time data which are not graphed here. These are unexpected effects, possibly caused by a fortuitously effective larger matched set.

The interaction between Names and Trial was due to a relative increase in confidence or decrease in response times when responding to others' names over the two trial levels. At the first trial respondents had an average confidence rating 0.51 units higher and average response time 507 ms lower for their own names (versus those chosen by others). This difference decreased to 0.36 units in confidence and 288 ms for response times by the second trial.

For response times, the Trial × Size × Names interaction ($F_{1,17} = 4.57$, $p < 0.05$), though statistically significant, was both small in magnitude (as determined by the mean squares for this effect) and difficult to interpret. Therefore no plot has been provided for this aspect of the data.

2.4.3 *Error (confusion) analysis.* The number of errors made during the response phase were calculated for the Colour Set (3) × Names (2) × Size (2) × Trial (2) design. An error was when the incorrect colour for the supplied colour name was chosen. A total of 615 errors were made out of the 3360 responses in total, an error rate of 18.30%.

These data were subjected to a hierarchical log-linear analysis, with simple deletion of effects. The tests of possible effects within the model are given in table 6. The model generated thus contained no interactions. Tests of partial association for all possible model components are presented in table 7. Note that only effect levels reported as significant from table 6 are included. A significant interaction, that of Names × Size was revealed in the analysis of all possible effects; it is nevertheless dropped since the overall test of two-way interactions revealed no significant effects (Tabachnick and Fidell 1989), as may be seen in table 6.

The model selected (Colour Set + Size + Names) had no cells as outliers, as indicated by the standard residuals for each cell; the largest standard residual had a value

Table 6. Log-linear analysis on error data, to test possible effects in generated model.

Effect level	Degrees of freedom	Likelihood ratio χ^2	p
1	5	191.93	<0.001
2	9	12.37	0.230
3	7	1.56	0.980
4	2	0.03	0.980

Table 7. Summary of partial association effects for hierarchical log-linear analysis of overall distribution of errors.

Effect	Degrees of freedom	Partial association χ^2	p
Names	1	30.78	<0.001
Colour set	2	87.87	<0.001
Size	1	71.05	<0.001
Trial	1	2.23	0.14

of −1.36. The overall likelihood ratio was $\chi^2 = 15.12$, $p = 0.71$. Note that this statistic tests the significance of the deviation between the observed and expected (model) frequencies, and thus nonsignificant χ^2 indicates a good fit between observed and expected frequencies. Error frequencies collapsed across each of the model factors were as follows: Colour Set (nameable 104, maximal 233, matched 278), Size (twelve-member 204, sixteen-member 411), and Names (own 239, other 376).

To evaluate the effect of perceptual colour difference (ΔE^*_{uv}) on which (pairs of) colours were confused, all data were pooled and the correlation between error frequency and ΔE between the colours so confused was calculated. Since the very large number of unconfused data pairs skewed the data markedly, these were omitted from the analysis. The resulting data consisted of 130 confused colour pairings, with $r = -0.194$, $p < 0.05$. This low correlation value—accounting for less than 4% of the variability in error frequency—was consistent with the hypothesis that colour difference is not a good predictor of ease of naming. The direction of the correlation indicated a significant but weak effect of more-perceptually-different colours being confused less frequently. The scatterplot, figure 3, shows this relationship.

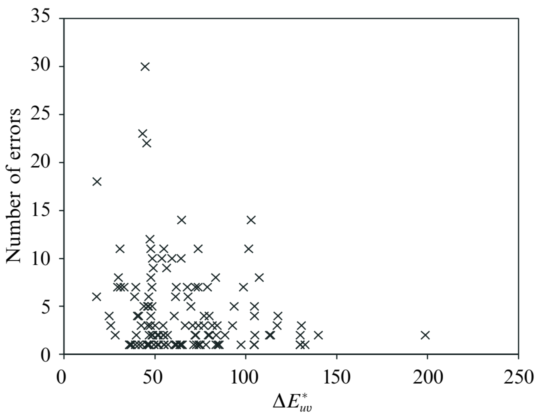

Figure 3. Scatter-graph of perceptual difference between pairs of colours (ΔE^*_{uv}) and the number of errors in naming occurring between pairs of colours.

In a subsequent exploratory approach correlations between ΔE and error frequency for all six of the data sets (3 colour sets × 2 set sizes) were calculated, again removing all zero-error cases. None of these correlations were better than $r = -0.32$, and none were statistically significant.

2.4.4 *Nameability and search performance.* For each of the six colour sets, the overall mean-nameability value was obtained and the correlations with mean response times and confidence ratings for the set were calculated. For matched and discriminable sets, nameability values were taken from the most similar colour for which an empirical nameability value was available. For nameability versus response times, $r = 0.54$ ($p > 0.05$); for nameability versus confidence ratings, $r = 0.90$ ($p < 0.05$). Data for the latter, statistically significant effect, are shown in figure 4.

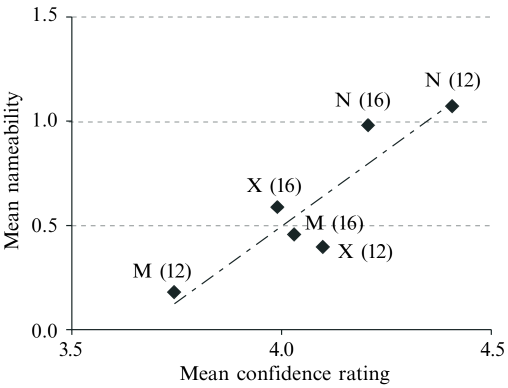

Figure 4. Scatter-graph of mean nameability versus mean confidence rating for data averaged for all combinations of colour-set type and size. Key: N—nameable; M—matched; X—discriminable. Set size is shown in parentheses and the line of best fit is also shown.

Mean response times, confidence ratings, and errors made for all eighty-four individual colours across all colour sets were also correlated with nameability, response times ($r = -0.42$, $p < 0.001$), confidence ratings ($r = 0.47$, $p < 0.001$), and errors ($r = -0.29$, $p < 0.01$). Figure 5 shows the response time versus nameability plot which is broadly representative of all the correlations calculated here. The lower values for these correlations versus those calculated for colour sets overall are expected because whether a colour was present within a nameable, matched, or maximal set is ignored. This latter fact is important because a set containing, say, twelve excellent examples of 'green' will certainly not be an excellent set for name-based search tasks despite each of the set members having a high nameability value.

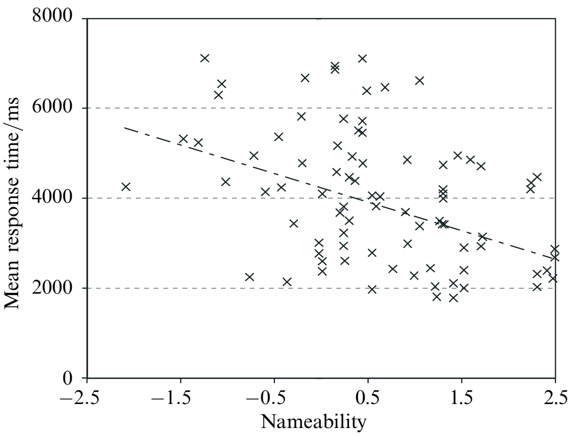

Figure 5. Scatter-graph of nameability versus mean response time for all individual colours responded to by subjects. The line of best fit is also shown.

2.5 *Discussion*

The primary finding from the experiment was that nameable sets did lead to superior performance on a naming task, as compared with metrically matched or maximally discriminable sets of colours. For the generation phase, assignation of names to colour sets was found to be easier (both faster and more confident) when the palette was designed as nameable. The lack of effect of the number of members in the palette on generation data may imply that observers were making absolute rather than relative naming assignations. That is, if observers were naming individual colour-set members in effective isolation, this would occur independently of the numbers items within the set.

The generation-phase results were echoed in the response phase, where responses to nameable palettes were faster, more confident, and more accurate when identifying the set members, as compared both with matched and with maximal palettes.

A change in response times and confidence ratings occurred across the two trials, with a relatively greater change in responses (more confident, and faster) to names given by others versus own names. However, there was no increase in accuracy (measured by a reduction in numbers of errors made) over the two trials. Thus observers may become faster and more confident, but they do not seem to become any more accurate, or at least the rate of any improvement is considerably slower than the rate of increase of confidence and speed.

As the palette size was increased, the distinction between the colour-set types faded somewhat. Notably, matched and maximal sets were indistinguishable on response time and confidence measures at the larger set size. Nameable palettes still produced faster and more confident responses, but the size of the difference reduced. This suggests a limit in the extent to which nameable colours may strongly influence behaviour in category-dependent tasks. Extrapolating, performance on the task reported here may become asymptotic across colour sets with similar metric characteristics when approximately twenty members are present in the sets. This will certainly vary with different tasks, different time scales between naming and identification, and doubtless many other factors. For instance, Derefeldt and Swartling (1995) have achieved successful name recognition of up to thirty colours. In their method stimuli were first named by subjects, and the test task was the recall of the correct (personal) name on the subsequent presentation of each of the stimuli. This is more or less a reversal of the task we used, and may suggest asymmetry in the linkage between colour name and colour example.

2.5.1 *Errors.* When identifying colours from names, colour pairs with a small perceptual difference (ΔE) are no more likely to be confused than colour pairs with large ΔE values, ie perceptually very different colours. In other words, the colour difference between samples does not allow one to specify which pairs will be confused in a naming task.

Since ΔE does not predict name (category) errors, then what does? The nameability of colours provides a somewhat better predictor of error likelihood than ΔE, but is still only weakly linked. This is unsurprising because nameability says nothing about colour categories which must be critical in allowing effective naming to occur. The combination of ease of naming and minimisation of the number of category duplications within sets would seem to be a likely candidate for true optimisation.

When considering categories, it is possible that confusions arise because observers have different ideas of what a (polylexemic or nonbasic) colour term means. For instance, observer one may have a prototypical image of 'lilac' as being (perceptually) a desaturated blue – purple, whereas the prototype for observer two may be perhaps bluer and more saturated. Confusions would then arise when observer one's name prototype is closer to an incorrect prototype for observer two. These types of error should be greatly reduced when basic exemplars are viewed (and hence basic colour terms used), since within-culture prototypes should be similar. This would lead to lower errors for nameable sets and, of course, lower errors for responses to one's own names—both true in the data set. However, there are no metrics for category membership and thus no easy way of quantifying this type of confusion.

The above reasoning suggests a personal, stable colour-name space. An alternative view might be that observers do not have an especially stable colour-naming space; instead they simply have the advantage of knowing what names they themselves generated in the current experiment, while having no knowledge of the names generated by others. If observers had prior knowledge of the others' names they were tested with (or if observers lost the knowledge of what their own names were), it is possible that the effect of own versus names given by others would reduce.

This is an important consideration. If observers have very stable personal colour naming, then the facilitation of colour communication between observers via nameable colours may be hindered. If, on the other hand, the advantage of personal names is purely through immediate knowledge of such names, then the advantage of own names versus those given by others should disappear either if observers return after a delay, by which time they will have forgotten their originally generated names, or if observers are provided with a list of the names given by the other. The former approach is used in the next experiment.

3 Experiment 2

To disentangle whether the performance advantage for own names is due to a personal, stable colour vocabulary, the response phase of the experiment was reprised after a delay of six to eight weeks. It was hypothesised that in doing this observers should lose any immediate familiarity with their own names that they had gained through the generation task. Thus if the superior performance when responding to one's own names was due to the effect of immediate familiarity, then this effect should disappear on the repeat trials. This design attempts to pull observer responses down to a baseline level of performance (ie a no-familiarity level), rather than attempting to push observers towards a performance ceiling, as presupplying lists of names given by others might do. Finding a baseline is probably a more fundamental aim, representing the way untrained (naïve) observers might behave when trying to communicate colours to each other.

3.1 *Method*

The general methodology and equipment remains as described earlier. Only changes in methodology are noted here.

3.2 *Participants*

After a delay of six to eight weeks, seven participants (one male, six females) out of the ten who originally completed the twelve-colour, response-phase, condition agreed to return to complete the study again. The sixteen-colour condition was not reconsidered since insufficient numbers of subjects were available the second time.

3.3 *Procedure and apparatus*

The general procedure remained as previously described, except that the initial phase (name generation) was not required, as the participants had already generated sets of names when they completed the experiment the first time. This shortened the duration of the experimental session by approximately 10 min.

3.4 *Results*

The available response latencies and confidence ratings were analysed in a similar fashion to previously, except that a new factor named Time was introduced. This refers to the two experimental sequences, the second undertaken after the delay of six to eight weeks.

Both confidence and response-time data were analysed separately by a Colour Set (3) × Names (2) × Trial (2) × Time (2) factorial ANOVA, with all factors being within subjects. Notice that in this phase of the experiment there was no Size factor since only the twelve-colour condition was tested. Each participant provided 288 data points for each of response times, confidence ratings, and accuracies. The seven available participants thus provided 2016 responses for the three response measures.

The only statistically significant effects for both response times and confidence ratings were those of Colour Set (confidence: $F_{2,12} = 32.16$, $p < 0.01$; response times: $F_{2,12} = 4.70$, $p < 0.05$), Names (confidence: $F_{1,6} = 21.88$, $p < 0.01$; response times: $F_{1,6} = 7.23$, $p < 0.05$), and Trial (confidence: $F_{1,6} = 7.72$, $p < 0.05$; response times: $F_{1,6} = 26.92$,

$p < 0.01$). Therefore there was no effect whatsoever of Time upon confidence ratings or response latencies.

A Colour × Set × Trial interaction was found for response times ($F_{2,12} = 4.31$, $p < 0.05$) which was not present in the initial experiment. The interaction seemed to indicate some deviations between matrices in amount of response-time improvement over the two levels of Trial. Such effects were very small in magnitude.

3.5 *Discussion*

The results obtained from the returning participants indicated that introducing a time delay had no discernible overall effect on responses. Observers did not become equivalent in performance on their own names and those given by others. Responding to one's own names remained faster and more confident. Note also that when they created their names, participants did not know that they would be tested again.

4 General discussion

Overall, observers perform better when they are producing names for, or identifying samples by name from, nameable sets of colours. Although this is what one would expect given the general body of colour-categorisation work, prior research using search tasks has not suggested such a distinction. The work of Smallman and Boynton (1990, 1993) found that coding with basic exemplars or coding with idiosyncratic (personal) exemplars led to virtually no difference in subsequent search performance. This is contrary to the results from the current experiment. The possible reasons for this discrepancy will be addressed now.

One possible reason for the discrepancy relates to design differences between Smallman and Boynton's (1993) study and the current study. Unlike in Smallman and Boynton's study, the design used in the experiment reported here was a purely example-cued identification task—there was no visual-search part of the experiment as such. If the time to identify a target class from a given name is a great deal lower and less variable than the time taken to actually search for the target class, then any effect of name type (ie basic versus idiosyncratic) could be lost in the overall data when 'traditional' searches are investigated. The current experiment was free of such additional search aspects which could overshadow effects of name type. One might argue, of course, that if effects of category are so easily lost within genuine visual-search data, then category-specified searches are no more useful than any other search-specification method.

A second important point is that the study reported here specifically matched the colour-difference characteristics of the different palettes, at least within certain limits. This should have eliminated any fortuitously 'good' or 'bad' colour sets. Smallman and Boynton could, in theory, have produced colour sets with rather different metric properties (especially in the 1993 study where participants defined their own colour sets). However, this does seem unlikely given that participants were satisfied with the discriminability of all sets they produced. One should also note that, although different palette types within the current experiment were metrically manipulated, perfect matches between metrics of nameable and matched sets were not obtained. One possible way to achieve a near-perfect match would be to derive colour sets which are rotations of each other in colour space, to maintain precisely the relationship between all colours within each set. It would then be possible to consider whether category per se or the complex metric relationship between the colours is the critical factor. Whether it is practically feasible to produce multiple rotated colour sets in colour space is another matter, with the gamut of available colours presenting rather rigid constraints, especially for CRT work.

One interesting (and uncontrolled) difference between the colour sets was the variability in ΔE values within each set. Though the data are not explicitly shown in this report, nameable colour sets always contained the most variation. That is, nameable sets tended to be less homogenous in their ΔE relationships than the other sets. Whether this sampling difference is critical in determining the name performance of sets is unclear. It is possible that nonhomogenous distributions reduce possible category confusions to very specific colour pairs within the set, and that the small colour differences for such colours actually cross category boundaries for the nameable sets. One could test this in further experiments by matching the within-set variability independently of category membership in addition to the other metric parameters controlled in this study. Once more this assumes sufficient degrees of freedom to produce colour sets algorithmically with such constraints.

Notwithstanding the limits in the colorimetric matching, we found no correlation between the colour difference of a colour pair and the chance of that pair being confused in the identification task. If this result is robust, it has important implications for any colour-set-generation algorithms which aim to optimise naming and search performance. It seems insufficient simply to space colours very widely in colour space in order to obtain perceptually distinctive and categorically distinct colours. Instead, if one wishes to have nameable and maximally discriminable colours within the same algorithm, the principle suggested by Van Laar and Flavell (1994) seems more realistic. In this method colours are maximally spaced (by simple linear optimisation) with the additional constraint of limiting the movement of colours to prevent movement away from predetermined name centres in colour space. The quantification of the loss in nameability as colours are moved around colour space may be possible with the data of (for example) Boynton and Olson (1987), Sturges and Whitfield (1995), or Guest and Van Laar (2000). A further worthwhile aim would be to combine such nameability data into a unified nameability model, perhaps by using the approach of Lammens (1994) which could allow more pragmatic generation of nameable discriminable colour sets. We should note here that optimising nameability alone will be insufficient to produce excellent sets since this optimisation in itself says nothing of actual category membership. It is critical to consider categories and nameability together.

The issue of personal colour language is an important one. We found that responding to the colour language of others was significantly more difficult than responding to one's own, even when immediate familiarity of own names was reduced. This suggests that personal colour vocabulary may be fairly stable, and with some idiosyncrasy. However, observers improve rapidly when responding to the vocabulary of others—the gap between performance on own and on others' names reduces. Thus mapping the unfamiliar vocabulary to one's own equivalent (or directly to the colour example), is both possible and rapid. So it seems likely that the idiosyncratic names can become clear in the colours they denote between observers. This is somewhat analogous to the distinctively nonbasic regions of colour space identified by, for example, Boynton and Olson (1987), Sturges and Whitfield (1995), and Guest and Van Laar (2000); ie difficult-to-name regions of colour space can still be categorically distinctive. It is also likely that providing a list of unfamiliar names to observers would further reduce the difference between performance on own and on others' name sets and so aid the efficiency of the colour communication.

These issues link directly to the question of an optimally discriminating colour language. This would be a set of carefully chosen colour terms and modifiers designed to facilitate the accurate identification of colours from name alone. The current work indicates that manipulation of name characteristics of colour sets is feasible, and that basic colour terms might form the cornerstone of this discriminating language. One might consider the ISCC-NBS colour-naming scheme as a basis for a more complex, but controlled, vocabulary. A practical test of the use of the system in communicating colour to naïve and trained observers would be a worthwhile aim. This would indicate the efficacy of the naming system in communicating colour information to observers, and would test whether the system is in some sense optimal for real communication, especially compared with the untrained naming considered here.

References

Berlin B, Kay P, 1969 *Basic Color Terms: Their Universality and Evolution* (Berkeley and Los Angeles, CA: University of California Press)

Boynton R M, 1997 "Insights gained from naming the OSA colors", in *Color Categories in Thought and Language* Eds C L Hardin, L Maffi (Cambridge: Cambridge University Press) pp 135 – 150

Boynton R M, Fargo L, Olson C X, Smallman H S, 1989 "Category effects in color memory" *Color Research and Application* **14** 229 – 234

Boynton R M, Olson C X, 1987 "Locating basic colors in the OSA space" *Color Research and Application* **12** 94 – 105

Boynton R M, Olson C X, 1990 "Salience of chromatic basic color terms confirmed by three measures" *Vision Research* **30** 1311 – 1317

Carter R C, Carter E C, 1982 "High-contrast sets of colours" *Applied Optics* **21** 2936 – 2939

Christ R E, 1975 "Review and analysis of colour coding research for visual displays" *Human Factors* **17** 542 – 570

De Corte W, 1990 "Recent developments in the computation of ergonomically optimal contrast sets of colours" *Displays* **4** 123 – 128

Guest S, Van Laar D L, 2000 "The structure of colour naming space" *Vision Research* **40** 723 – 734

Derefeldt G, Swartling T, 1995 "Color concept retrieval by free color naming—identification of up to 30 colours without training" *Displays* **16** 69 – 77

Kay P, Berlin B, Maffi L, Merrifield W, 1997 "Color naming across languages", in *Color Categories in Thought and Language* Eds C L Hardin, L Maffi (Cambridge: Cambridge University Press) pp 21 – 56

Lammens J M G, 1994 *A Computational Model of Color Perception and Color Naming* unpublished doctoral dissertation, University of New York, Buffalo, NY

Luria S M, Neri D F, Jacobsen A R, 1986 "The effects of set size on color matching using CRT displays" *Human Factors* **28** 49 – 61

Post D L, 1992 "Colorimetric measurement, calibration, and characterization of self-luminous displays", in *Color in Electronic Displays* Eds W Widdel, D L Post (New York: Plenum Press) pp 299 – 312

Ratliff F, 1976 "On the psychophysiological basis of universal color terms" *Proceedings of the American Philosophical Society* **120** 311 – 330

Rosch-Heider E R, 1972 "Universals in color naming and memory" *Journal of Experimental Psychology* **93** 10 – 20

Sanchez R R, Nagy A L, 1989 "Interaction of color and luminance differences for optimal visual search" *SID 89 Digest* **20** 296 – 299

Smallman H S, Boynton R M, 1990 "Segregation of basic colours in an information display" *Journal of the Optical Society of America A* **7** 1985 – 1994

Smallman H S, Boynton R M, 1993 "On the usefulness of basic colour coding in an information display" *Displays* **14** 158 – 165

Sturges J, Whitfield T W A, 1995 "Locating basic colours in the Munsell space" *Color Research and Application* **20** 364 – 376

Tabachnick B G, Fidell L S, 1989 *Using Multivariate Statistics* (New York: HarperCollins)

Van Laar D L, Flavell R, 1994 "A name based algorithm for generating maximally discriminable colour sets" *Journal of Photographic Science* **42** 83 – 84

Whitfield T W A, 1981 "Salient features of colour space" *Perception & Psychophysics* **29** 87 – 90