Quantifying Variations in Personal Color Spaces: Are There Sex Differences in Color Vision?

David L. Bimler,1* John Kirkland,1 Kimberly A. Jameson2

1 Health and Human Development, Massey University, Private Bag 11-222, Palmerston North, New Zealand
2 Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109

Received 18 September 2002; revised 17 April 2003; accepted 7 May 2003

Abstract: We report a search for group differences in color experience between male and female subjects, focusing on the relative prominence of the axes of color space. Dissimilarity data were collected in the form of triadic (odd-one-out) judgments, made with the caps of the D-15 color deficiency test, with lighting conditions controlled. Multidimensional scaling reduced these judgments to a small number of dimensional-weight parameters, describing each subject’s sensitivity to color axes, i.e., how much each axis contributes to the inter-color dissimilarities perceived by each subject. Normal trichromatic subjects in two age bands were examined, teenagers and university students, and in both cases males placed significantly less weight on a ‘red-green’ axis, and more on ‘lightness’. We consider the implications and possible explanations.

INTRODUCTION

Observers in color-perception research are typically screened for normal color discrimination, using pseudoisochromatic plates or a standardized hue test such as the Farnsworth 100-Hue, but separate analyses of males and females are not always considered or performed. This is despite indications that perceptually and cognitively, males and females may experience color appearance differently. Some of the research reviewed below supports the suggestion that these sex differences correspond to variations at the retinal level, but there remains uncertainty in the color vision literature concerning the connection between the two levels. Here we seek to shed light on these experiential differences by probing their correlates in other aspects of color perception.

A number of studies have compared male and female use of the color lexicon, and provided convergent evidence that women access a larger repertoire of words to describe standardized sets of color stimuli. The relative fluency with color terms among females extends to finding color samples to match them, and providing glosses or definitions. This trend emerges across a range of cultures and languages. It is already evident in childhood: girls are earlier than boys to learn the names of primary colors, though this may reflect an overall lead in language skills among girls. There may be a link to the tendency for females to be better than males at matching colors from memory. Cognitive and social phenomena could account for these differences: it may be that the divergent patterns of socialization for males and females instill a greater awareness of color among women.

However, a physiological explanation is also conceivable, given current knowledge about the biological basis of color perception. Advances in molecular biology in recent years have revealed the existence of females possessing more than three distinct classes of retinal photopigment, by virtue of X-chromosome inherited opsin gene dimorphisms (i.e., the opsins that determine each photopigment’s spectral sensitivity occur in multiple versions, tuned to different wavelengths of peak absorption). In a task requiring subjects to divide a spectrum into bands, female subjects on average perceived a greater number of distinct bands, but the difference was traceable to a specific subgroup of women heterozygous for opsin genes: homozygous women were no different from men.
Another study used reflected-light stimuli to examine unique-hue judgments from 22 female and 18 male subjects.\(^8\) Judgments of ‘unique red’ from the females formed a distribution with twice the variance as those made by males. The mean ‘y (unique yellow)’ was significantly longer in wavelength for females than for males, though the two distributions were equally broad.

These are not the only sex differences in color-perception behavior to be found in the literature. Women showed a distribution with twice the variance as those made by males. The mean ‘y (unique yellow)’ was significantly longer in wavelength for females than for males, though the two distributions were equally broad.

There are not the only sex differences in color-perception behavior to be found in the literature. Women showed greater responsiveness to the long-wavelength segment of the spectrum in a color-discrimination task.\(^9\) Men and women are also thought to differ in the physiological substrate of color vision: in one study, for instance, the average level of macular pigment (which selectively filters short-wavelength light) was 36% lower in women.\(^10\) However, such psychophysical variations are still in dispute; and if real, they are not necessarily manifested as gender differences at the experiential level, since there is evidence (reviewed in the Discussion) for compensatory mechanisms in color processing.

The range of these examples of sex differences in color experience points to a complex situation with many variables. The present study takes a complementary, cognitive approach and examines individual differences in the pattern of similarities among colors. We employ the geometrical paradigm of a ‘color space’, where the dissimilarity between any pair of chromatic stimuli is modeled by the distance between two points representing them in the space. Investigating the variations between subjects byeliciting the dissimilarities they perceive among stimuli is a prominent strand in color research.\(^11\)–\(^14\) Studies in this tradition are associated with analysis of such dissimilarities with multidimensional scaling (MDS).

In particular, individual-differences MDS algorithms can reduce the data to a small number of parameters for each subject, specifying how that subject’s personal perceptual color space departs from the consensus space.\(^15\) We assume that the personal spaces can be derived from the consensus space by compressing or elongating it along orthogonal axes, so individual variations are parameterized by the ‘weight’ or extent of elongation of each axis, i.e., its contribution to inter-point distances. Such variations would naturally be reflected as a distinctive pattern of perceived dissimilarities. Many forms of color-vision deficiency become amenable to quantification with MDS when they are considered as distortions (axial compressions) of color space.\(^14,16\) Note that the MDS analysis chooses a set of axes that best account for the variations among observers; specific axes need not be pre-determined.

One MDS study\(^17\) used 13 Munsell chips as stimuli, and a triadic procedure to elicit dissimilarity judgments: informants viewed three chips at a time, and selected the most similar pair in each triad. Women gave greater weight in their dissimilarity judgments to the third dimension, interpreted as a ‘yellow-green/yellow-red’ gradient. The second dimension was ‘cool/warm’ (or ‘blue/yellow’). Men gave greater weight to the first dimension, but its interpretation is hindered by the vagueness of the verbal glosses specifying the stimuli. Though the study identified it as ‘Hue/non-hue’, it could be a saturation gradient, or a distinction between median values of lightness and the extremes (black and white).

It is plausible that cognitive, culturally-influenced factors contribute to dimensional-weight variations of this form. Some dimensions of color space may be more accessible or noticeable to an observer having lower-than-average awareness of the subtleties of color. Although the color lexicon of English (and many other languages) emphasizes distinctions of hue, this is not universally the case: languages exist at an earlier stage of linguistic development, which focus on distinctions of brightness.\(^18\) There is a parallel with the way dichromats can learn to emphasize subjective lightness cues as an aid to distinguishing hues.\(^19\) In general, brightness or luminance differences matter more to color-deficient individuals than to normal trichromats.\(^12,20\)

With this in mind, the research question addressed here is whether sex differences in color perception appear in color similarity judgments. The present study follows Furbee et al.\(^17\) in employing a triadic procedure. Stimuli were 32 standardized caps used in a widely-available test for color vision. Analysis consisted of producing for each subject a configuration \(X_m\), that maximized the likelihood of generating his or her judgments, by finding optimal compressions and elongations along the axes of a spatial model of points (\(X_o\)) representing the stimuli. A form of constrained MDS\(^21\) was used, where the only degrees of freedom were the subject-specific distortion parameters.

**METHOD**

Stimuli were the 32 caps of the D15 and D15-DS color-vision panel tests (1.2 cm circles of pigment-coated paper in black plastic mounts, subtending an angle of 1.5° at normal viewing distance). In Munsell notation, the 16 caps of the D15 have Value 5 and Chroma 4 (with the exception of one cap with Chroma = 6).\(^22\) The D15-DS caps are lighter (Value 8) and less saturated (Chroma 2).

These were viewed resting on a sheet of gray card (N5), within a desktop booth of walls and baseboard painted matt gray. Two Philips 'TL'D/950 fluorescent tubes (color temperature 5300 K, Color Rendering Index > 95), suspended from struts extending above the booth’s walls, provided a luminance level that was measured by photometer and held constant at 500 lux by raising or lowering the two tubes.

Each subject began by arranging the 16 saturated caps into hue sequence, followed by the 16 desaturated caps (in each case starting with the 10B ‘pilot’ cap). This constitutes the standard panel tests for which the caps are designed. Next, stimuli were shuffled into five random triads with one left over at random each time. Subjects indicated the most-similar pair in each triad, by choosing the ‘odd-one-out’ stimulus (most dissimilar). This procedure was performed with the D15 and D15-DS sets separately; then repeated 12 times with a mixed set (in which caps 0, 3, 6, 9, 12 and 15 in the D15 sequence were replaced by their desaturated counterparts). In total, each subject made 70
odd-one-out decisions. No time limits were set on the responses. The entire process took 15–20 minutes. To protect the stimuli and avoid contrasting color stimulation, subject and researcher wore white gloves while making and recording these judgments.

A first group of informants consisted of New Zealand teenagers, ranging in age from 11 to 17. They were recruited as part of a twin study, and the majority were twins or triplets. Informants were excluded from the current analysis if their responses to the panel tests, or prior testing, identified them as color-vision deficient (CVD). This left 37 males and 30 females, with a median age of 14 in both cases.

A second group was recruited from students and staff at the universities of Massey (NZ) and Oakland (MI, US). The same sets of caps were used to collect data at both locations. As well as CVDs, smokers were excluded, since tobacco smoking may affect color vision. This left 11 males and 24 females, with most aged in the range 18–34, and median ages of 19 and 20 respectively.

Analysis. The multidimensional-scaling approach attempts to account for the data by representing the stimuli as points in a geometrical model, arranged so that the distances between each pair of points reflects the dissimilarity between the corresponding stimuli. The limitations of this approach are addressed in the Discussion below.

In a preliminary treatment of the data, one solution was obtained for the 37 boys in the first group of subjects, and another for the 30 girls (data from the second subject group were excluded here because possible differences between age groups, combined with the gender imbalance of the older subjects, could lead to confounding effects). We label these solutions as $X_b$ and $X_g$, each consisting of a matrix of coordinates for the points. Three dimensions were retained, since the stimuli vary in lightness (Value) as well as within the conventional two-dimensional color plane. To reduce the number of degrees of freedom, the points were confined to two parallel planes in space. The lighter (also desaturated) stimuli from the D15-DS were constrained to share a single value of the third dimension, while the darker (saturated) D15 stimuli shared a second value. Coordinates were adjusted iteratively to maximize their fit to subjects’ triadic comparison, using a maximum-likelihood algorithm implemented as a program MTRIAD. $X_b$ and $X_g$ are shown as Figs. 1 and 2.

We moved on to examine individual variations by using MTRIAD in its ‘weighted Euclidean’ mode. Here a separate solution $X_m$ is created for each subject by compressing or elongating a consensus solution $X_0$ along its axes. $X_0$, derived from the combined data of the first subject group, is essentially a compromise between $X_b$ and $X_g$. The axial compression parameters $w_m$ indicate the relative importance of the three dimensions to that subject. A limiting case such as $w_m = 0$ would correspond to an observer who perceives no difference between stimuli on one side of $X_0$, and their counterparts on the other (for instance, the points 10GY and 2.5YR in Fig. 1). This is represented in the geometric framework by compressing $X_m$ so that the points representing the indistinguishable stimuli coincide. In any triad involving such a pair of stimuli, the third stimulus would always be the odd-one-out.

The $w_m$ were found by maximizing the goodness-of-fit between $X_m$ and subject $m$’s data. Here ‘goodness’ is defined in terms of likelihood. The likelihood of $X_m$ giving rise to an observed triadic response is enhanced if the corresponding points form an elongated triangle, with the odd-one-out stimulus at its acute corner (naturally these triangles are differently elongated compared to their counterparts in $X_0$, perhaps in different directions). Defining $l_m$ to be the average likelihood over all 70 triad decisions from
that subject, the optimal values of \( w_m \) are a compromise that maximizes \( l_m \). To ensure that scale is constant across the separate solutions, the algorithm imposes the constraint that \( \sum_d w_{md}^2 = 3 \): the sum of squared distances between the points and the origin is thus the same in each \( X_m \).

**RESULTS**

Figures 1 and 2 show the first two dimensions of the male-only and female-only MDS solutions \( X_b \) and \( X_g \). They have much in common, with both giving a reasonable reproduction of the known arrangement of the D-15 caps in color space. Possibly there is more vertical and less horizontal dispersal of points in Fig. 1. It is more objective, however, to express the dispersal along each dimension (i.e., the variance) as a percentage of the total variance. For \( X_b \) the percentages are 53\%, 33\%, and 14\% for the Red-Green, Blue-Yellow and Lightness dimensions respectively. For \( X_g \) they are 65\%, 30\% and 5\%.

This suggestive contrast proved to be robust when checked with a split-half test. That is, two separate versions of \( X_b \) were generated, each using data from half of the teenage males, assigning them at random to two groups. The three dimensions accounted for 54\%, 32\% and 13\% of one solution, and 56\%, 31\%, and 13\% of the other. When two versions of \( X_g \) were generated, each from half of the teenage females, the dimensions accounted for 66\%, 30\% and 4\% of one, and 66\%, 29\% and 5\% of the other.

The split-half MDS solutions also show \( X_b \) and \( X_g \) to be robust. The two versions of \( X_b \) were closely related, with a correlation of \( r = 0.85 \) when the 32.3!2 = 496 inter-point distances in one version were compared with their counterparts in the other. The two versions of \( X_g \) were equally similar, with a correlation of \( r = 0.87 \).

The difference between the groups is also evident at the level of individual parameters summarizing the trend of each subject’s responses. To remove any possibility of age effects confounding this individual-difference comparison, the teenage and student subjects are considered separately. Subjects are represented as points in Figs. 3 and 4 where the ordinate and abscissa are \( w_{m1}^2 - w_{m2}^2 \), and \( w_{m3}^2 \). The central vertical axes show where \( w_{m1} = w_{m2} \), i.e., where red-green and blue-yellow differences contribute equally to perceived dissimilarity.

The individual goodness-of-fit index \( l_m \) can range up to 1 (if \( X_m \) predicts that subject’s responses exactly); a value of 0.5 would mean that the predictions from \( X_m \) were no better than chance. The values are quite high (see Table I), implying that suitably-compressed versions of the consensus solution \( X_0 \) are reproducing the subjects’ responses well.

The results of two-sided \( t \) tests on these parameters are listed in Table I. In both age-groups, the mean values of \( w_{m1}^2 \) were significantly lower for males than for females; mean \( w_{m2}^2 \) values did not differ significantly; mean \( w_{m3}^2 \) values were higher for males; and mean \( l_m \) values were lower for males. Table I also includes standardized coefficients comprising a canonical discriminant function for each age-group. Produced by canonical discriminant analysis, this function classifies subjects as male or female from a linear combination of the parameters, chosen to match actual sex as often as possible. The function for teenage subjects yields a correct classification in 48 of 67 cases. The function for students is correct in 32 of 37 cases.

Any reduction in \( w_{m1} \) entails increases in the other two weight parameters, because of their interdependence. The higher \( w_{m3} \) among males does not necessarily mean that lightness is more salient in absolute terms; only that lightness remains as a potential distinguishing feature on which they could base an odd-one-out decision despite decreased sensitivity to differences in the color plane defined by red-green and blue-yellow axes.
 functions chosen to distinguish the male and female populations.

DISCUSSION

Given the central role of multidimensional scaling in this study, a few caveats are in order. In broad terms, MDS is a tool for data-reduction, which extracts structure from a mass of data and displays it visually. The resulting spatial model may represent the raw data accurately but this does not guarantee any isomorphism with the actual structure underlying the phenomenon of interest—color perception, in this case. Studies using a large number of stimuli across multiple levels of lightness have yielded MDS solutions similar to theory-driven color-appearance models such as the Munsell system. This affords some confidence in the MDS approach, but its limitations should not be minimized. Proximity judgments are averaged across informants in such studies. Individual variations are thus obscured; they may not conform to the restricted form considered here, if indeed they fit the model of color-space distortions at all. Nor is it certain that a three-dimensional Euclidean space can adequately describe the global properties of color perception, even though three coordinates are enough to specify a color (in aperture mode).

With this reservation in mind, the MDS solutions used here (Figs. 1 and 2) are presented not as a model of color vision, but only as a heuristic device: a framework for summarizing the trend of each informant’s responses and indicating one way in which groups of informants differ.

Studies have shown the structure of semantic similarities (among chromatic terms) to resemble that of perceived similarities (among physical stimuli), i.e., the structures of cognition and perception are (nearly) isomorphic. However, this isomorphism applies to averaged judgments. It does not guarantee convergence of perception and cognition in single observers; for dichromats at least, the two diverge. Note also that the majority of relevant studies have used a small number of terms, too few to vary saturation or luminance systematically. Discrepancies from perceptual experience begin to emerge when the semantic items include achromatic terms (White, Black, Gray) and desaturated terms (e.g., Beige, Tan, Pink).

The semantic structure of the color lexicon is relevant even to studies involving physical stimuli, as here, because subjects can respond by verbally encoding stimuli and consulting the consensus ‘received opinion’ about the similarity of color terms. The role of a verbal-encoding strategy in judgments of color dissimilarity has been elegantly under-lined by disrupting it with verbal distraction. The attractiveness of this strategy (i.e., the importance of cognitive structure) is enhanced when the stimuli are saturated prototypical hues with high codability. Judging perceptual proximities for such stimuli is hindered by the large distances between them in color space, i.e., their incommensurate nature. “[. . .] two colors such as (5R 4/14, 5G 5/8) simply appear ‘entirely different’, and the perceptual difference in these pairs is not intuitively palpable as in more moderate color differences. There seems to be a limit within which impression of difference naturally takes place.” (Ref. 24, p 461).

One recent study examined both the proximities among eight saturated ‘focal hues’ and among the eight corresponding ‘Basic Color Terms’. Differences among the two MDS solutions were robust but not substantial. Differences of the same order were also found between male and female groups, between Chinese- and English-speakers, and between tasks for eliciting proximity judgments. Arguably these were of little consequence in the context of intra-group variation and noise. Male and female groups seemed to access consensus views about the color plane in a similar way, and the difference between them as spokespeople of these views was not substantial.

In contrast, the sex difference found in the present study is more relevant to color experience, because of the use of stimuli that were less suitable for verbal encoding (being unsaturated, they were poorer exemplars of color categories). Moreover, the stimuli were spaced at smaller intervals in color space, there being 32 of them rather than eight.

It is tempting to relate this difference—summed up by the contrast between Figs. 1 and 2, and best described in terms of the relative importance of the dimensions of color space—to sex differences in the color lexicon, and in unique hues. Thus we turn to the findings (reviewed in the Introduction) that women have readier access to the color lexicon. A preatory caveat is necessary: dimensional weighting is a higher-order feature of color space than the number of regions it is partitioned into (by the color lexicon) or the exact locations within it of category best-exemplars (unique hues). It is also a logically independent feature, that can differ between males and females without necessitating differences in the others.

Color-lexicon differences could be ascribed to differences in socialization, or to generally better verbal fluency.
However, in a task requiring subjects to divide a spectrum into bands of color, the gender difference could be narrowed down to a specific subgroup of female subjects: those who were heterozygous for variant forms of the M- or L-pho-
topigment genes.7 Homozygous women performed no dif-
differently from men. This result suggests that that the trend
towards more complex color vocabularies is traceable to rich color experience among some women, arising from
the details of retinal physiology.1 At the retinal level, fe-
males heterozygous in this way possess a possible source of
difference in the form of multiple populations of M- or
L-cone cells, with slightly different photopigment sensi-
tivities.

There is evidence that variations in the weight of color
dimensions are partly inheritable.23 Thus one can speculate
that these variations are also somehow rooted in genetic
polymorphism of the photopigments. This is only specula-
tion, however, with no evidence that our female subjects
included heterozygotes.

It may be relevant here that the Red-Green and Blue-
Yellow axes, conventionally associated with the color
plane, do not behave equivalently: two unique hues defining
the latter (°b and °y) are known to cancel out in practice,
whereas this is not true of °r and °g.30

However, not all the evidence is consistent with a close
linkage between retinal performance and color experience.
Color-naming, at least, seems to be decoupled or insulated
from other stages in visual processing. Webster et al.31
investigated observers’ unique-hue judgments. No correla-
tion was found with their sensitivity to the canonical axes of
cone color space; nor was there a significant association
between any pair of unique hues as they varied across
observers. In theory, °y should depend on an individual’s
ratio of cone numbers, but in practice it is relatively unre-
sponsive to variations in that ratio.32 Even with the extreme
L/M bias found in carriers of dichromacy, °y is within the
normal range.33

Further evidence can be gleaned from the constancy of
unique hues and of the subjective experience of color across
the life span.34 This is despite the yellowing of the lens in
later life (its absorption of shorter-wavelength light is en-
hanced by increasing pupillary constriction with age, mean-
ing that light cannot enter the eye through the optically-
thinner outer zones of the eye). Again, diabetic observers
may incur reduced S-cone sensitivity, but continue to name
unique hues within the normal range.35

CONCLUSION

In summary, males and females differed in their (average)
responsiveness to particular dimensions of color space. The
goodness-of-fit index Xm (the likelihood of correctly predict-
ing the m-th subject’s response to each triadic comparison)
was lower on average for males. This was the case for
independent subject samples at two age levels. As well as
replicating an earlier observation,17 this is consistent with
earlier observations that males have less access to the color
lexicon2–4 and are less attuned to consensus views on color
semantics.29 However, the difference is quite specific: males
placed less weight on inter-stimulus separation along the
Red-Green axis. In partial compensation, they placed more
weight on separation along the lightness axis. Thus it is
possible to identify subjects as male or female with greater-
than-chance accuracy—despite the overlap of the two distri-
butions—on the basis of a linear combination of their color-
axis weight parameters (i.e., a canonical discriminant
function).

These sex differences are more substantial than those
emerging from other research using MDS methodology.29
This can be partly attributed to the use here of individual-
differences MDS15 (specifically, the weighted-Euclidean
model of individual difference). The effect is to ignore any
individual variations that do not conform to this restrictive
model, and it may be that this narrowing of the scope has
increased the signal-to-noise ratio. Other procedural depar-
tures from Ref. 29 (the number and nature of stimuli) were
discussed above.

The framework used here for analysis, where a consensus
geometric model Xfn is compressed along a common set of
axes (red-green, blue-yellow, and luminance) has the ad-

dantages of acceptance in a long-standing strand of color
research; and of being conservative, in the sense that it is
hard to see how it could give rise to spurious or artifactual

group differences. Which is not to say the resulting indi-

cidual color-space models are the best way of accounting
for each subject’s judgements. The possibility remains that
better-fitting individual models could be generated by other
ways of distorting a consensus model (perhaps invoking
more subject-specific parameters).

A disadvantage of the incomplete triadic data collected
here is that they do not lend themselves to a sophisticated
partitioning of their variance (i.e., Cultural Consensus anal-
ysis) into contributions from different sources such as intra-
group and inter-group variance, and noise. This is compen-
sated by the simplicity and speed of the present procedure,
and the widespread availability of the stimuli.

Explanations for differences in color experience could be
sought at a number of levels, from retinal performance (e.g.,
photopigment heterozygosity in a subgroup of females) to
patterns of socialization. The former possibility could be
addressed by extending the MDS approach to subjects of
known photopigment genotype. In view of the possible
influence on color experience from absorption of short
wavelengths by macular pigment, it may be that variables to
control, as well as age and sex, should include the subjects’
levels of macular pigment.

We emphasize that differences between group means
such as those found here are superimposed on large intra-
group variability; they have little to say about individuals.
Nor can it be assumed that men and women are each
homogeneous groups. Between-group differences are com-
patible with the existence of subgroups.

2. Rich E. Sex-related differences in colour vocabulary. Lang Speech
234–247.
160–169.
20. Dr. Smith has noted that the *note at the bottom of page 360 in Volume 28, Issue 5, pages 360–365, has two errors; the article title and the email address. It should read

"*Extended version of the paper, Red or Read: The Built Environment is Coloured" presented at the 2001 AIC Congress, Rochester, NY. Email: d.smith@qut.edu.au."