Reprinted from C. Blakemore (1990) 'Vision: Coding and Efficiency', Cambridge University Press

11

The two subsystems of colour vision and their rôles in wavelength discrimination

J. D. Mollon, O. Estévez and C. R. Cavonius

Introduction

Horace Barlow makes only occasional forays into the field of colour vision (Barlow, 1958, 1982), but when he does, he always leaves us with much to think about. In his 1982 paper 'What causes trichromacy?', he gave us a novel way of considering the information content of a coloured spectrum: he expressed the detailed structure of the colour spectrum in terms of its Fourier components and he treated the three photopigments (Fig. 11.1) as low-pass filters that would differentially attenuate the different Fourier components. Owing to the broad bandwidth of the filters, the visual system is insensitive to the fine structure of the colour spectrum; that is to say, if the amplitude of a stimulus varies periodically with wavelength and if the period of this modulation is small, then the response of the visual system will show little variation as the phase of the modulation is changed (Barlow, 1982).

In considering his main question – that of why our colour vision is three-dimensional – Barlow was led also to ask several secondary questions: 'Why do the photopigments have such broad bandwidths?', 'Are broad bandwidths deleterious to hue discrimination?' and 'Why are the peak sensitivities of the pigments so asymmetrically placed in the spectrum?' We hope that the present paper may say something in answer to these secondary questions. We first put forward a general view of the early stages of colour vision, the view that it consists of two subsystems, one recently overlaid on a much earlier one; and then we review some experimental work on wavelength discrimination, work that bears on the two subsystems of colour vision.

The emerging view of colour vision is one that has long been suggested by the distribution of colour discrimination among the mammals (Jacobs, 1982); by the relative incidences of different forms of colour blindness in man (Ladd-Franklin, 1892); and by a number of asymmetric features of our colour vision, such as the relative paucity of the short-wave receptors (compared with the middle- and long-wave receptors) and indeed the asymmetric arrangement of the absorbance curves of the photopigments (Gouras, 1984; Mollon, 1986; Mollon & Jordan, in press). It is a view consistent with Derrington, Krauskopf & Lennie's electrophysiological analysis of the parvocellular layers of the lateral geniculate nucleus (Derrington *et al.*, 1984). But, above all, it is a view prompted by the molecular biology of the visual pigments published by Nathans and his collaborators (Nathans *et al.*, 1986a,b). The most salient findings of Nathans *et al.* are: first, that the genes for the long- and middle-wave pigments lie very close together on the q-arm of the X-chromosome, and second, that the amino-acid sequences for the two pigments are 96% identical. It appears that visual pigments all consist of seven helices, which span the membrane and form a palisade surrounding the retinal, the prosthetic group that gives the pigment its spectral sensitivity (Fig. 11.2). The solid circles in the diagram to the bottom right of Fig. 11.2 show the small number of amino acids that Nathans *et al.* identify as different between the long- and middle-wave pigments. The strong implication is that these two pigments were differentiated only very recently,[1] by duplication of an ancestral gene. On the

[1] 'Very recently' would here mean 'within the last 30 to 40 million years' (see Nathans *et al.*, 1986a). That the duplication event occurred after the divergence of platyrrhine and catyrrhine monkeys is also suggested by the fact that New-World primates appear to have only one X-chromosome locus for a visual pigment (Mollon, Bowmaker & Jacobs, 1984; Bowmaker, Jacobs & Mollon, 1987)

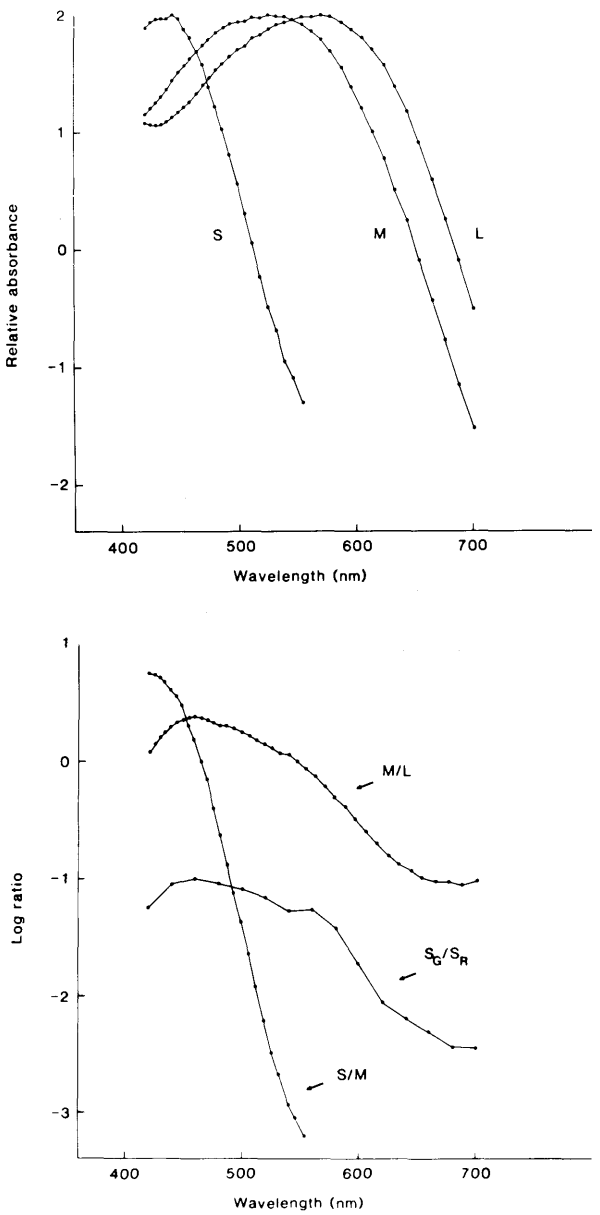

Fig. 11.1. The upper panel shows the spectral sensitivities of the short-wave (*S*), middle-wave (*M*) and long-wave (*L*) cones. These 'König fundamentals' are those derived by Estévez from the Stiles–Burch 2-deg colour-matching data. In the lower panel are plotted the log ratio of middle-wave to long-wave cone sensitivity (*M/L*) and the log ratio of short-wave to middle-wave sensitivity (*S/M*). Also plotted in the lower panel is the log ratio of middle-wave to long-wave cone sensitivity (S_G/S_R), as obtained by electrophysiological recording from individual cones of *Macaca fascicularis* by Nunn, Schnapf & Baylor (1985) (this function is arbitrarily placed on the ordinate). In order to allow direct comparison with the electrophysiological results, which were obtained by transverse illumination of receptors, the sensitivities of the 'König fundamentals' are given as absorptances for a pigment solution of low concentration. There is only a very small shift in the position of the maximum of *M/L* when allowance is made for self-screening of the pigment *in vivo*; and pre-receptoral absorption cannot change the position of the maximum.

other hand, the gene for the short-wave pigment is located on chromosome 7 and the amino-acid sequence for that pigment differs as much from the sequences for the long- and middle-wave pigments as it does from that for the rod pigment, rhodopsin: the implication is that the short-wave pigment has long enjoyed an independent existence.

The two subsystems of colour vision

These considerations suggest the nature of the two subsystems that underlie human colour vision.

Widespread among mammals is a primordial, dichromatic, system of colour vision that compares the rates at which photons are absorbed in the short-wave cones, on the one hand, and, on the other, in a second class of cone with a peak sensitivity that varies across species but lies always in the green to yellow region of the spectrum (Jacobs, 1982). This ratio is extracted by a minority class of ganglion cells that exhibit little sensitivity to spatial contrast: such cells (and their counterparts in the parvocellular laminae of the lateral geniculate nucleus) behave as if they draw antagonistic inputs from coextensive or nearly coextensive regions of the receptor array, and thus this subsystem comes close to being a pure colour system (Gouras, 1984; Derrington & Lennie, 1984). There is some evidence that a morphologically distinct channel subserves the extraction of this primordial colour information: Mariani (1984) has described an uncommon type of primate bipolar cell that resembles the usual midget invaginating bipolar, except that the cell body gives rise to several dendrites and may make contact with two, well-separated, cone pedicles (which Mariani takes to be short-wave cones). The short-wave cones, and the higher-order cells that carry their signals, can be sparsely distributed, because they are given little part to play in the analysis of spatial detail (Tansley & Boynton, 1976; Thoma & Scheibner, 1980); and this in turn is likely to be because the short-wave component of the retinal image is chronically degraded, since non-

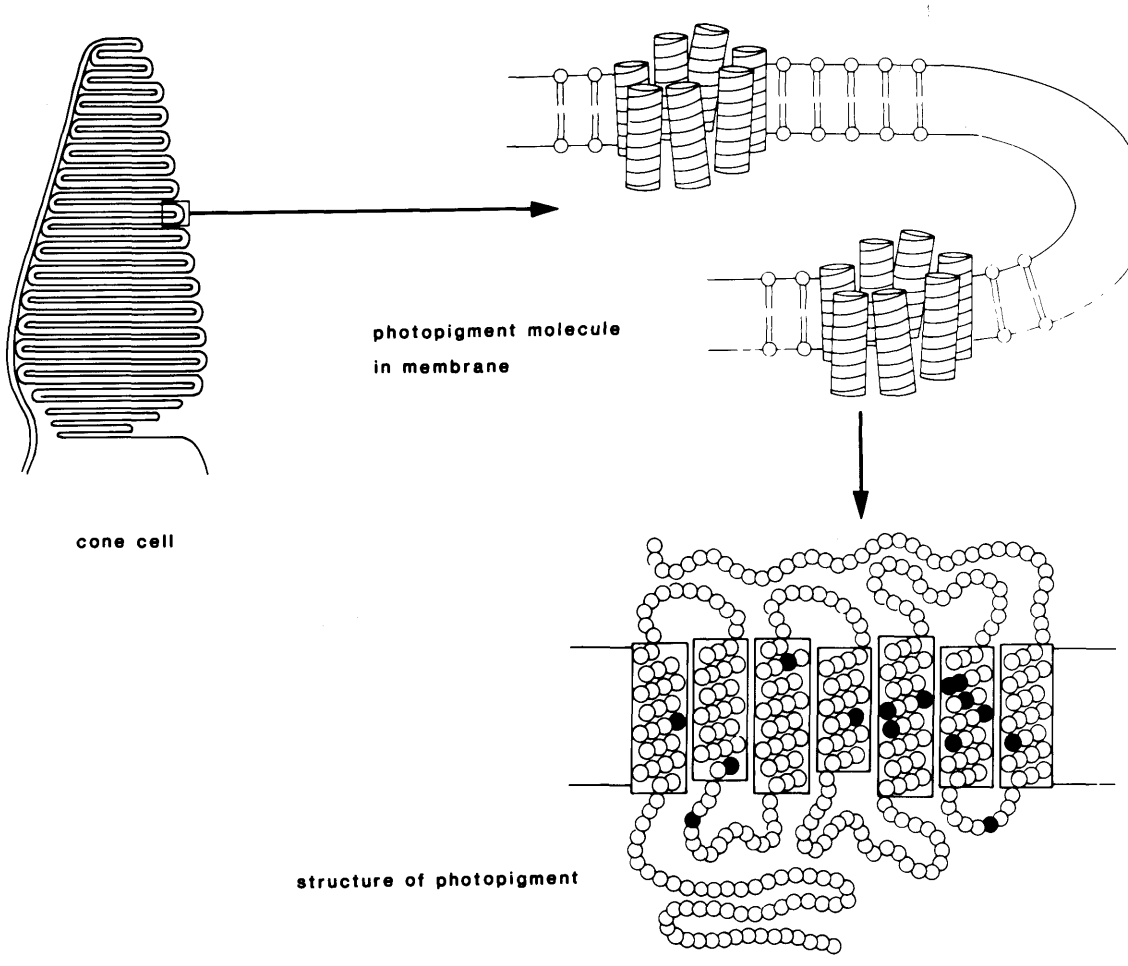

Fig. 11.2. The structure of visual pigments. The upper panels illustrate the arrangement of the pigment molecules in the infolded membranes of the outer segment of a cone: each molecule consists of seven transmembrane helices, which cluster around the chromophore. At the bottom right is represented (after Nathans) the sequence of amino acids in the protein part of the photopigment: the filled circles indicate those amino acids that differ between the long-and middle-wave pigments of the human retina.

directional skylight dilutes the shadows of the natural world and since the eye itself is subject to chromatic aberration.

In the Old World monkeys and in man, the ancestral subsystem is overlaid by a second colour vision system that compares the rates at which quanta are caught in the long- and middle-wave cones. This recently evolved system is parasitic, we suggest, on a channel that still has as its chief purpose the analysis of spatial detail. The substrates for this channel are the midget bipolars, the midget (or 'Pβ') retinal ganglion cells, and the predominant type of parvocellular unit in the lateral geniculate nucleus. Such cells draw their antagonistic inputs not only from distinct classes of cone but also from distinct regions of the field; and thus they respond both to colour and to spatial contrast.[2]

In summary then, the three cone pigments of

[2] It is not uncontroversial to claim that this channel still has for its main business the analysis of spatial detail. On the basis of the higher contrast sensitivity and higher gain of the magnocellular system, Shapley & Perry (1986) have suggested that the *principal* function of the parvocellular laminae is colour vision. And there have been several reports that maximal spatial resolution is similar for units in the magnocellular and parvocellular streams of the visual system (e.g. Blakemore & Vital-Durand, 1986; Crook, Lange-Malecki, Lee & Valberg, 1988). For a detailed discussion of the position adopted here, see Mollon & Jordan (in press).

normal human vision have different origins and different evolutionary rôles. The ancestral middle-wave pigment, maximally sensitive near the peak of the solar spectrum, subserved the primary purposes of vision, the analyses of motion, form, and depth. A long time ago, the short-wave cones were added, sparingly, for the sole purpose of colour vision. Most recently, the ancestral middle-wave pigment differentiated to give a second dimension of colour vision; but in this case, both the daughter pigments developed within the constraints of the original function, namely, spatial vision.

In the sections that follow, we consider the respective rôles in wavelength discrimination of the two subsystems of colour vision. In treating separately the contributions of the two subsystems, our analysis resembles that of Walraven and Bouman, who distinguished 'deuteranopic' and 'tritanopic' components of hue discrimination (Walraven & Bouman, 1966; Bouman & Walraven, 1972), or that of Judd & Yonemura (1970), who constructed the normal wavelength-discrimination curve from 'protanopic' and 'tritanopic' components.

Wavelength discrimination: methodological considerations

In one respect, the modern study of wavelength discrimination has been curiously backward. In psychoacoustics, it has been for twenty years almost unacceptable to use any method but two-alternative temporal forced-choice, which minimizes the effect of variation in the observer's criterion; those who work on spatial frequency have nearly as honourable a record; and even in colour vision it is commonplace to use such performance measures when chromaticity is modulated on displays with fixed phosphors. But, to the shame of our sub-branch of visual science, the discrimination of wavelength is still commonly measured by a time-honoured method of adjustment (Wright & Pitt, 1935): the subject increases the difference in wavelength until the standard and variable fields look different and then sees if he can eliminate the perceptual differences by adjusting the luminance of the variable. If he can, he increases the wavelength difference and repeats the process until the two half-fields look different at all luminance settings. It is easy to imagine that the criterion used by the observer might differ according to the spectral region being examined, and in particular, according to which subsystem of colour vision (i.e. which ratio of cone signals) was mediating discrimination.

Two obstacles may excuse the psychophysical backwardness that has characterised studies of wavelength discrimination. The first is the mechanical one, that it has been less easy to manipulate wavelength in real time than to manipulate acoustic frequency or spatial frequency. This difficulty has passed with the introduction of monochromators with integral stepping motors: the experiments we describe were all done with monochromators that allow the centre wavelength of the passband to be adjusted in steps of 0.05 nm (Type M300E, Bentham Instruments, Reading, UK).

The second difficulty is that in some parts of the spectrum a change of wavelength can be detected by a change in luminosity more readily than by a change in hue (Laurens & Hamilton, 1923). If we adopt a performance measure and eschew the observer's subjective judgements, then we need to ensure that we are testing colour discrimination and not luminosity discrimination. We have adopted two solutions to this second problem:

Method of Average Error

The most venerable of performance measures is the Method of Average Error, the method used by König & Dieterici to measure wavelength discrimination in their great paper of 1884. In our version of this procedure, the observer views a 2-deg bipartite field, one half of which is fixed in luminance and wavelength. The computer repeatedly offsets the wavelength and luminance of the other half field by amounts that are random in size and direction; the subject, manipulating a joystick in two dimensions, must restore a complete match; and the performance measure is the standard deviation of 50 settings (Mollon & Estévez, 1988). The subject cannot use luminosity as a guide to wavelength: he must himself make a match on both dimensions.

Forced-choice

We have allowed the psychoacousticians to teach us how to measure frequency discrimination by two-alternative forced-choice. When modern measurements of acoustic frequency discrimination were first attempted, the resolution at high frequencies appeared much better than in the classical study of Stücker (1908), who, using Galton whistles, tested Mahler and other musicians of the Vienna Court Opera. Henning (1966), however, introduced small random variations of amplitude into the tones to be discriminated and found that his results agreed with those of Stücker. The implication is that Stücker was not able to blow his Galton whistles absolutely steadily and so was able to secure true measurements of

frequency discrimination. We have translated Henning's paradigm to vision (Mollon & Cavonius, 1987). On each trial, there are two presentations of a bipartite field. The lower, standard, half-field is fixed in wavelength and luminance. On one of the two presentations the wavelength of the upper half-field is the same as that of the standard; and on the other presentation the upper field has a wavelength that is greater by a value $\triangle\lambda$. To prevent the use of luminosity cues, small random variations in the luminance of the upper half-field are introduced within and between trials, the range of variation being ± 0.1 log unit. The subject must indicate on which of two presentations the fields differ in wavelength, and $\triangle\lambda$ is adjusted by a staircase procedure to track the 71% correct point. The value of $\triangle\lambda$ is initially 3 nm; this value is increased by a factor 1.25 after each incorrect response and decreased by a factor 0.8 after two correct responses (Moore, Glasberg & Shailer, 1984). Feedback is given by tone signals after each trial.

The reader may wonder why we did not use a simpler form of the forced-choice paradigm, in which the observer was asked to say whether the longer wavelength was in the upper or lower half of a single bipartite field. A spatial forced choice of this kind was used in the strikingly modern study of Laurens & Hamilton (1923). For our present purposes, a specific difficulty is that 'longer' corresponds to different appearances in different spectral regions; and in particular, when the stimuli are bright, there is a wavelength near 460 nm where excursions in longer and shorter directions give a similar change in appearance (Mollon & Estévez, 1988). The particular forced-choice method we adopted (see above) requires the subject to make only the simplest discrimination: on which of the two presentations is a physical difference in wavelength more likely to have been present. He is not asked to report the direction of the difference.

The short-wave pessimum

Figure 11.3 offers an overall view of wavelength discrimination as examined by the two-alternative temporal forced-choice method (see above). The troland value of the lower, standard, half-field was always 50, with Judd's correction (Wyszecki & Stiles, 1982, Table 5.2.2) being applied below 460 nm; and the duration of each presentation was 800 ms (these turn out to be significant parameters). The bandwidths of the monochromators were set at 1 nm, and broad-band gelatin blocking filters, appropriate to the standard wavelength, were placed in the final common beam. Thresholds were estimated from the last 16 of 20 reversals of the staircase. Wavelengths were sampled every ten nanometers in two different randomised sequences; the plotted thresholds represent the means of the two separate estimates for each wavelength.

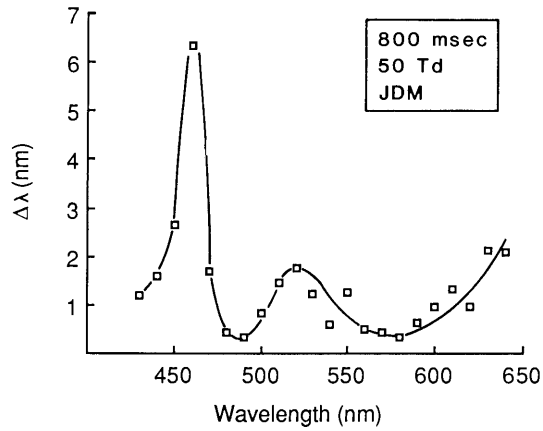

Fig. 11.3. Wavelength discrimination measured by two-alternative temporal forced-choice. For details of methods, see text. Observer: JDM.

As in the classical data (König & Dieterici, 1884; Wright & Pitt, 1935), there are two clear minima in the function of Fig. 11.3, near 490 nm and 580 nm. Owing to the forced-choice method, the absolute values of the thresholds in these regions are lower than those found with adjustment methods (Wright & Pitt, 1935), dropping to as little as 0.25 nm. It is noteworthy that the visual system achieves this impressive resolution with photodetectors that have a bandwidth of the order of 100 nm (Fig. 11.1). But perhaps the most salient feature of the present results is the size of the peak at 460 nm, where the threshold rises to 7 nm. We shall refer to this peak as the 'short-wave pessimum'. The useful term 'pessimum' was introduced by Robert Weale (Weale, 1951) and it avoids the ambiguity of speaking of a maximum in the wavelength-discrimination function. In the rest of this paper, we concentrate on an analysis of the short-wave region of the spectrum, since it serves well to illustrate the rôle in wavelength discrimination of the two subsystems of human colour vision and the factors that limit their performance.

Figure 11.4 gives a magnified view of discrimination in the short-wave region. Measurements were made by the method of average error (Mollon & Estévez, 1988) and standard wavelengths were sampled every 5 nm). The thresholds are typically lower than those obtained by forced-choice, since here they

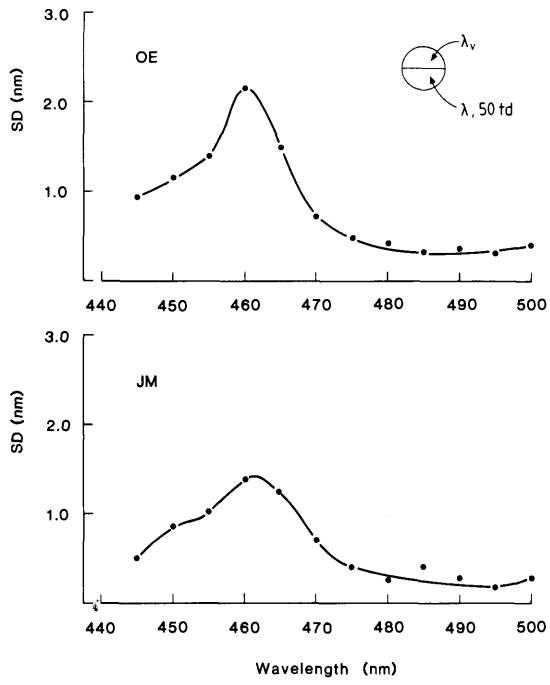

Fig. 11.4. Wavelength discrimination in the short-wave region, measured by the method of average error. The ordinate represents the precision (standard deviation) of settings of the variable wavelength (λ_v) when matching standard wavelengths (λ) in the range 445 to 500 nm. Each data point represents the r.m.s. value for two independent sets of 25 matches. Results are shown separately for two of the present authors. Note the well-defined peak ... the function at 460 nm, the 'short-wave pessimum'.

represent the standard deviation of matches, but again a sharply defined peak occurs at 460 nm.

So, at this point in the spectrum, neither of the two subsystems of colour vision seems to be doing us much good. But the reasons are different, we suggest, for the two subsystems. Consider first the subsystem that compares the quantum catches in the middle- and long-wave cones. The most obvious factor that must limit wavelength discrimination (and the factor considered in traditional analyses) is the rate of change with wavelength of the ratio of quantum catches. Near 460 nm the ratio of middle-wave to long-wave sensitivity (*M*:*L*) goes through a shallow maximum (see Fig. 11.1, lower panel) and the rate of change must slow down to zero. That this is so is implied by the colour-matching functions for tritanopes, who lack short-wave cones, and by the corresponding functions for normal observers viewing under tritan conditions (Fischer, Bouman & Ten Doesschate, 1952; Wright,

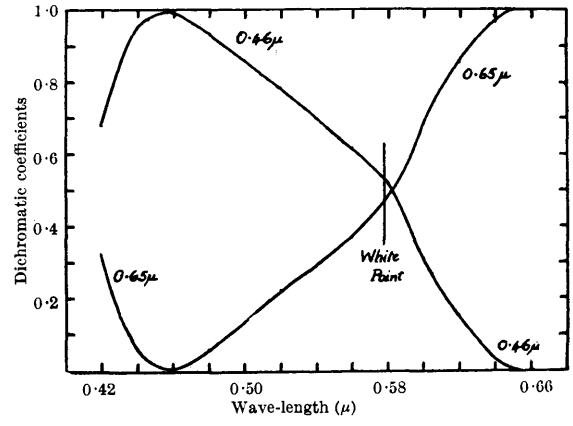

Fig. 11.5. Dichromatic coefficient curves for a centrally fixated colour-matching field that subtended 20′ of arc (from Willmer & Wright, 1945). The primaries were 460 and 650 nm; the units of these two stimuli were chosen to be equal in a match to a yellow of 582.5 nm, and the values for the 460-nm (*b*) and 650-nm (*r*) coefficients in any match were adjusted to bring $b + r = 1$. All wavelengths between 420 and 650 nm could be matched by positive mixtures of the two primaries, and this is the basis for our argument that the ratio of middle- to long-wave cone sensitivity must peak near 460 nm (see text). Reproduced by permission of Professors E. N. Willmer and W. D. Wright.

1952; Willmer & Wright, 1945; Alpern, Kitahara & Krantz, 1983). Figure 11.5 shows colour matching results for W. D. Wright when the bipartite stimulus field was confined to the central 20′ of the foveola, conditions under which the eye is dichromatic and matches are thought to depend only on the long- and middle-wave cones (Willmer & Wright, 1945). The ordinate of Fig. 11.5 shows the proportion of each primary in the match and the important result is that all wavelengths between 420 and 650 nm can be matched by positive mixtures of the 460-nm and 650-nm primaries. Provided that we allow the minimal assumption that the ratio *M*:*L* is higher at 460 nm than at 650 nm, these tritanopic colour-matching results oblige us to conclude that the ratio *M*:*L* is maximal at 460 nm. For suppose the maximum occurred at some other wavelength λ_x, then *ex hypothesi* monochromatic light of 460 nm would give a smaller ratio of middle- to long-wave signals than did λ_x. Adding the 650-nm primary could only reduce the ratio further. There would be no positive mixture of the 460-nm and 650-nm primaries that gave the high ratio of cone signals produced by λ_x, and thus a colour match would not be possible. So, there cannot be another

wavelength, λ_x, that represents the maximum.[3] Of course, it is tritan colour-matching data, along with protan and deutan data, that constrain the König fundamentals to be what they are.

We should like to labour this point a little. The understanding of colour vision was historically held back by the application of colour names to the receptors. König was among the first securely to grasp that a red light is not a light that maximally stimulates the long-wave cones but a light that produces a high ratio of long- to middle-wave cone excitation. 'Red', 'green' and 'blue' cones have today been expelled from almost all but ophthalmological textbooks; yet respectable visual scientists still use the terms 'red–green' and 'blue–yellow' for the second-order channels of the visual system. This is not an innocuous practice: it misleads students and their betters alike. In fact, the so-called 'red–green' channel is maximally polarised by red light on the one hand, and on the other by a wavelength (460 nm) close to unique blue; while the so-called 'blue–yellow' channel is maximally polarised by violet light and red or yellow light. It remains an open question whether there exist, at a more central stage, red–green and yellow–blue processes that correspond to those postulated by Opponent Colours Theory (Mollon & Cavonius, 1987).

That the ratio *M*:*L* passes through a shallow maximum in the blue spectral region is independently suggested by results obtained from a modification of Stiles' two-colour increment-threshold method (Estévez & Cavonius, 1977). And if further confirmation be needed, it is given by the *M* and *L* cone sensitivities recorded electrophysiologically from macaque cones by Nunn *et al.* (1985): the log ratio of these sensitivities is shown in the lower panel of Fig. 11.1.

So, this is the explanation – a traditional one – why the *M* vs. *L* subsystem is of little help to wavelength discrimination at 460 nm: the rate of change of the ratio slows down to zero. But the ratio of short-wave to middle- or long-wave sensitivity (*S*:*M* or *S*:*L*), extracted by the more ancient subsystem, is seen to be changing rapidly in the region of 460 nm (Fig. 11.1, lower panel). Why cannot our ancient subsystem help us here? The answer is suggested by three experimental operations that paradoxically improve wavelength discrimination at 460 nm.

[3] An objection occasionally put to us, from those who know something of colour theory, is that it is well known that colour-matching results are compatible with an infinite number of sets of fundamental sensitivities; and that one cannot therefore deduce the position of the maximal value of M:L from tritan colour mixing results. But an objection of this kind cannot counter the specific conclusion being drawn here from dichromatic data.

Three paradoxes of wavelength discrimination

Reduction of intensity: the König–Dieterici Anomaly

In the region of 460 nm, wavelength discrimination can be paradoxically improved by reducing the luminance of the stimuli. Such an effect is visible in the classical curves of König & Dieterici (1884), and Mollon & Estévez (1988) have proposed that it should be known as the 'König–Dieterici Anomaly'. The anomaly was rediscovered by McCree (1960). It is also apparent in the wavelength-discrimination curves given for normal observers by Haase (1934), and those for a deuteranope given by Walraven & Bouman (1966). The latter data (obtained at 1 and 10 trolands) are especially instructive, since the deuteranope's discrimination may be thought to depend only on the comparison of short- and long-wave cone signals; and the improvement near 460 nm is seen to be a consequence of a leftward shift of the deuteranope's U-shaped function as luminance is reduced (see Fig. 11.9).

Figure 11.6 shows direct measurements of the König–Dieterici Anomaly at 460 nm, using the method of average error. Each data point represents the root-mean-square value for three independent sets

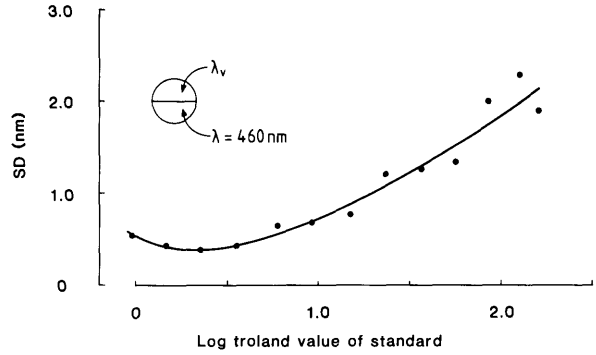

Fig. 11.6. The König–Dieterici Anomaly, measured by the method of average error. The figure shows the precision (standard deviation) of wavelength matches for monochromatic standard lights of 460 nm plotted as a function of their luminance. Each data point represents the r.m.s. value for three independent sets of 25 matches. Note that discrimination is best at low luminance.

of 25 matches. In this experiment the standard deviation of wavelength matches falls from around 2.0 at 100 td to well under half a nanometer at 5 td.

Reduction of duration

The second operation that paradoxically improves wavelength discrimination at 460 nm is a shortening of stimulus duration. Figure 11.7 shows results obtained by two-alternative temporal forced-choice for durations between 8 and 800 ms. Different durations were sampled in random order in two separate sequences. The standard and variable fields formed the lower and upper halves of a 2-deg bipartite field, the standard field had a troland value of 50, and the bandwidth of the monochromators was 1 nm. Other details of the procedure were as described for the experiment of Fig. 11.3.

It can be seen from Fig. 11.7 that discrimination at 460 nm improves systematically as duration is reduced below 50 ms. Under the conditions of the present experiments, reduction of duration and reduction of intensity thus produce effects in the same direction; but it is interesting to note that we have here an instance – perhaps unique – where reducing duration has an effect that is opposite in direction from the effect of reducing *area*: a reduction of field-size usually *impairs* wavelength discrimination in the region of 460 nm (Willmer & Wright, 1945), whereas we find here that the threshold falls when duration is reduced.[4]

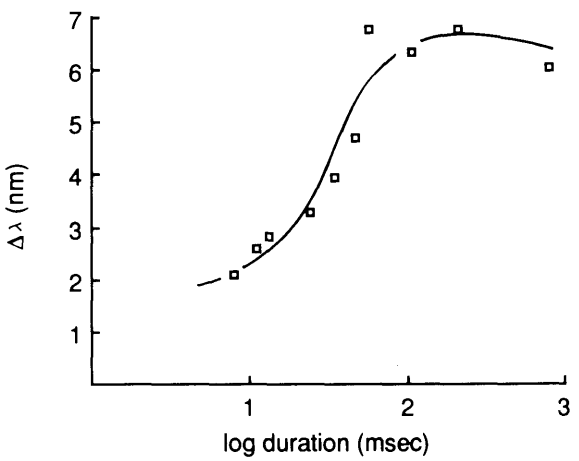

Fig. 11.7. Wavelength discrimination at 460 nm as a function of the duration of the flash. Measurements were made by a two-alternative temporal forced-choice procedure and the observer was JDM. Notice that performance improves as stimulus energy is reduced.

The König–Dieterici Anomaly and the effect of reducing stimulus duration both point to a saturating signal as the explanation why the older subsystem cannot sustain good discrimination for 50-td lights near 460 nm: when the quantum flux presented to the short-wave cones is reduced, by lowering the stimulus luminance or by shortening duration, we suppose that there is a shift to shorter wavelengths of the spectral region where the older subsystem begins to be saturated (and that there is an concomitant shift of the region of optimum discrimination, since the middle wavelengths now give too small a quantum catch in the short-wave cones[5]). But what is the site of the saturation? Do the short-wavelength cones themselves saturate, as in the model of Vos & Walraven (1972), or does the saturation occur in a post-receptoral colour-differencing system? There is also the (less plausible) possibility that rods, initially saturated, come to play a rôle in the discrimination of 2-deg centrally-fixated fields when stimulus energy is reduced.

[4] Farnsworth (1958) suggested that discrimination was relatively better for the tritan direction of colour space at short durations. It is not clear whether his result is related to the one shown in Fig. 11.7, since (a) the range of durations over which his improvement occurred (2000 vs 200 ms) is of a different order from the critical range found here and (b) the improvement was only relative to discrimination on a red–green dimension and the absolute values of the thresholds were higher at 200 ms. Farnsworth does not give the luminance of his stimuli.

[5] A hypothesis of this kind will readily explain the results of Foster, Scase & Taylor (1987), who found that wavelength discrimination near 500 nm was severely impaired when the duration of a 100-td stimulus was reduced from 1 s to 3 ms. The short-wave cone system has a particularly large critical duration: for a 100-td 500-nm stimulus the value is of the order of 130 ms (Krauskopf & Mollon, 1971, Figure 3). Thus, for the ancient, 'deutan', subsystem of colour vision, a change in duration from 1000 to 3 ms represents a reduction of, say, 1.65 log units in the effective stimulus energy. This attenuation of the stimulus will shift the minimum of the deutan discrimination to shorter wavelengths (Walraven & Bouman, 1966). Wavelength discrimination at 500 nm will deteriorate, owing to an inadequate quantum catch in the short-wave cones. The sharp cusp near 500 nm in the 3-ms data of Foster *et al.* can be interpreted as the intersection of separate 'deutan' and 'tritan' functions (see Fig. 11.9).

Cancellation by added fields

There is a third experimental operation that counter-intuitively lowers wavelength-discrimination thresholds at 460 nm; and this effect shows that at least some of the saturation occurs in a post-receptoral channel which draws inputs from the short-wave cones on the one hand and from the long- or middle-wave cones on the other. In this experiment (Mollon & Estévez, 1988), we held constant at 40 td the luminance of the standard half-field. To both halves of the bipartite field we added yellow-green light (560 nm) of increasing intensity, the added light forming a uniform, steady, circular field. In this case then, the quantum catch of the short-wave cones themselves cannot fall; it effectively remains constant as the yellow-green field is increased. If the short-wave cone signal is already saturated, then adding yellow-green light will not make it less so. Thresholds were measured by the method of average error.

The abscissa of Fig. 11.8 shows the troland value of the added yellow-green field (note that the 460-nm field is fixed). For each observer the standard deviation of the matches is halved at a point between 2.0 and 3.0 log units of added yellow-green light. In this case, the recovery cannot be explained by a release from saturation of the signals of the short-wave cones themselves (nor of rod signals); rather we may suppose that the saturation occurs in a post-receptoral colour-differencing channel that draws signals of opposite sign from the short-wave cones, on the one hand, and some combination of the middle- and long-wave cones, on the other. 460-nm light of 50 td places this channel in a saturating region of its response characteristic; and adding an opposing middle-wave field brings the channel back to a sensitive part of its operating range. The 560-nm light is a desaturant in both the phenomenological and the mechanistic sense. By this account, the present result is essentially the same as the 'combinative euchromatopsia' (the facilitation of increment sensitivity) observed for violet targets when a long-wave auxiliary field is added to a primary blue field (Mollon & Polden, 1977; Polden & Mollon, 1980).

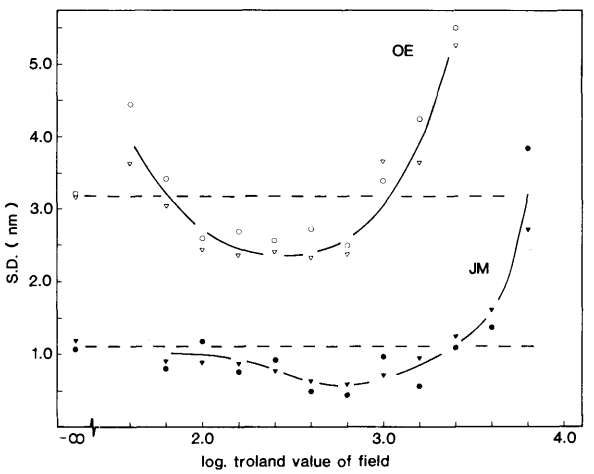

Fig. 11.8. The precision (standard deviation) of wavelength matches at 460 nm as a function of the log troland value of a 560-nm 'desaturating' field. The 460-nm standard field was fixed at 40 td. Data are shown for two independent runs for two of the authors. The results for OE have been displaced vertically by 2 nm. The horizontal broken line indicates the precision of matches when no desaturant was used.

In conclusion, then, an analysis of wavelength discrimination at short wavelengths suggests that our hue discrimination is limited by two factors, first, by the rate of change with wavelength of the ratios of quantum catches and, second, by the limited dynamic range of post-receptoral channels. In the case of the younger subsystem (the middle-wave/long-wave system), of course, both factors may operate at 460 nm, since the ratio of quantum catches there reaches its extreme value.

The rôles of the two subsystems

So, if a full answer is one day to be given to the questions raised by Horace Barlow in his 1982 paper (see Introduction, above), we think it will be inappropriate to treat human colour vision as a single system, designed all at once and for the sole purpose of hue discrimination. Rather, it will be necessary to consider the evolutionary history of colour vision, and the constraints imposed by the eye's other functions.

Initially, we suppose, there was a single cone pigment in the mid-spectral region. It was designed to subserve the chief tasks of vision, the discrimination of movement, flicker and form. To ensure high sensitivity, its bandwidth was broad and its peak wavelength was close to the peak of the daylight spectrum. Later, a second class of cones were (very frugally) added. These cones, with their peak sensitivity at short wavelengths, far removed from that of the middle-wave cones, served only for colour vision. By comparing their quantum catch with that of the middle-wave cones, the visual system achieved the most basic form of colour discrimination, the discrimi-

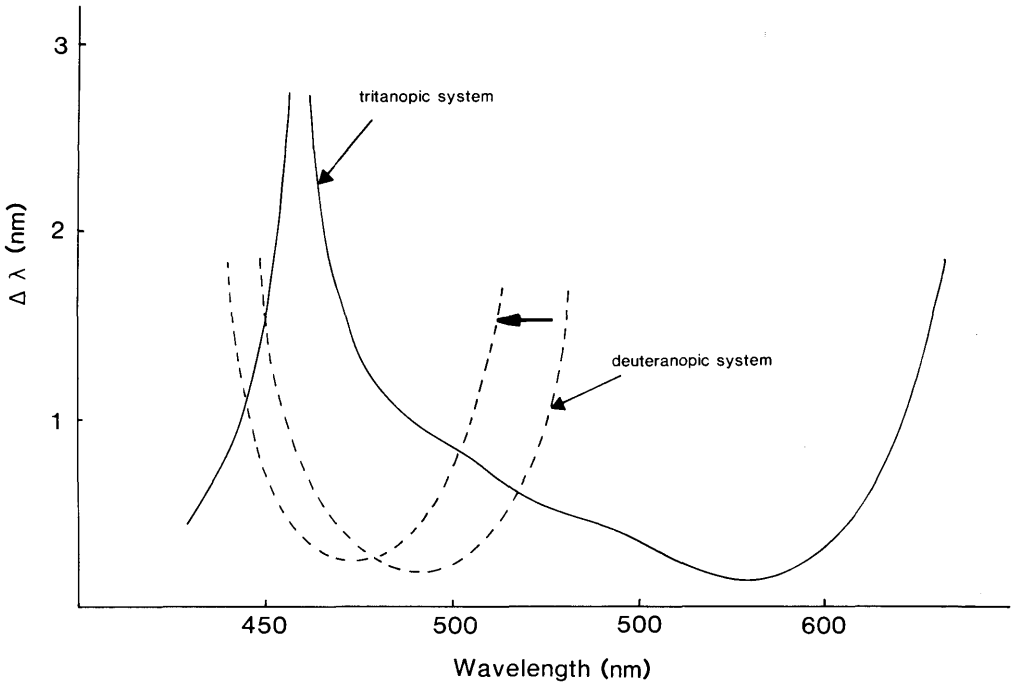

Fig. 11.9. The 'deuteranopic' and 'tritanopic' component functions that underlie normal hue discrimination. The rightmost of the broken lines represents the deuteranopic component when the troland value of the field is of the order of 100; when the troland value is reduced the deuteranopic function shifts leftwards as shown. The functions shown are schematic, but the deuteranopic component is based on the wavelength-discrimination curves given for deuteranopes by Pitt (1944) and by Walraven & Bouman (1966); and the tritanopic component is based on tritanope 'A' in the study of Wright (1952) and the functions obtained for normal observers by Cavonius & Estévez (1978) under tritan conditions of viewing.

nation of the sign and the gradient with which an object's reflectance changed from one end of the spectrum to the other. Subjectively, this ancient discrimination is the discrimination of warm colours from cool, the discrimination of reds, browns, yellows and olives from blues and blue-greens – and both of these from neutral greys and whites. If we adopt Barlow's Fourier approach to colour mechanisms, but consider the *subsystem* (rather than the individual photopigment) as a channel, then the job of the ancient subsystem is to extract the lowest Fourier component of colour information, i.e. the phase and amplitude of the variation between one end of the visible spectrum and the other. At the level of 50 td, the ancient subsystem does not support the discrimination of monochromatic or near-monochromatic lights, except in a small spectral interval near 500 nm; this is the interval where monochromatic lights yield ratios of short- to middle-wave excitation that resemble the ratios produced by daylight. The interval is limited on the short-wave side by the Scylla of saturation, and on the long-wave side by the Charybdis of tritanopia: at shorter wavelengths, the ratio *S*:*M* (or *S*:*L*) is too large, and at longer wavelengths the quantum catch of the short-wave cones is inadequate. As stimulus energy is reduced (by reducing luminance or duration), the operating interval of good spectral discrimination shifts to shorter wavelengths (see Fig. 11.9).

It is noteworthy that the second subsystem of colour vision, added more recently by the duplication of a gene on the X-chromosome, extends our wavelength discrimination not only at long wavelengths, but also at short (Fig. 11.9). It has always been of interest that tritanopes (who depend on the second subsystem alone) enjoy good hue discrimination throughout the spectrum, except in the vicinity of 460 nm (Wright, 1952; Fischer *et al.*, 1952; see Fig. 11.9); and in normal vision, at least for stimuli of 50 td or more, it is the ratio *L*:*M* that sustains the discrimination of wavelengths below 460 nm. This may be one

consequence of the substantial overlap of the middle- and long-wave cone sensitivities:[6] there is no wavelength that produces a very large or very small value of the ratio *M*:*L*. If the peak sensitivities of the middle- and long-wave cones were more separated (or of their bandwidths were narrow), we should expect the spectral discrimination of the second subsystem to be limited in the same way as the first: as soon as we left the narrow spectral window where the antagonistic cone signals were in balance, one of the two inputs would quickly become very large while the other would be limited by a failing quantum catch.[7]

So, the second subsystem extends discrimination in the wavelength domain. But there is another way of looking at its job: perhaps its true function is to extend discrimination in the Fourier domain. This is the domain to which we were introduced by Barlow (1982), although here it is the subsystem, rather than the photopigment, that is taken as the filter. Because the absorbance spectra of the long- and middle-wave pigments overlap so substantially, the second subsystem will be insensitive to the low Fourier components that well stimulate the ancient subsystem; its maximal response will be to intermediate components, although, as Barlow (1982) showed us, it must always be limited by the bandwidths of its two underlying photopigments.

It is often claimed that higher Fourier components are rare in the reflectances of the natural world (e.g. Lennie & D'Zmura, 1988), but this is the case only when spectroradiometric samples are taken from large surfaces. If the measurement is confined to part of an individual leaf or individual fruit, then fine detail is readily apparent in the spectra of the world of plants (Goodwin, 1965). Perhaps it was for discrimination among the carotenoids, the chlorophyls, and the flavins that the second subsystem of colour vision was given to us.

Acknowledgements

We are grateful to L. Winn for instrument making, to S. Astell and G. Jordan for experimental assistance, and to H. B. Barlow and S. Shevell for comments on the text. The experiments described above were supported by MRC Grant G8417519N.

References

Alpern, M., Kitahara, K. & Krantz, D. H. (1983) Classical tritanopia, *J. Physiol. (Lond.)*, **335**, 655–81.

Barlow, H. B. (1958) Intrinsic noises of cones. In *Visual Problems of Colour*, National Physical Laboratory Symposium No. 8, vol. 2. London: HMSO.

Barlow, H. B. (1982) What causes trichromacy? A theoretical analysis using comb-filtered spectra. *Vision Res.*, **22**, 635–43.

Blakemore, C. & Vital-Durand, F. (1986) Organization and post-natal development of the monkey's lateral geniculate nucleus. *J. Physiol. (Lond.)*, **380**, 453–91.

Bouman, M. A. & Walraven, P. L. (1972) Color discrimination data. In *Handbook of Sensory Physiology*, vol. VII/4, ed. D. Jameson & L. M. Hurvich, pp. 484–516. Berlin: Springer.

Bowmaker, J. K., Jacobs, G. H. & Mollon, J. D. (1987) Polymorphism of photopigments in the squirrel monkey: a sixth phenotype. *Proc. Roy. Soc. Lond.*, B**231**, 383–90.

Cavonius, C. R. & Estévez, O. (1978) π Mechanisms and cone fundamentals. In *Visual Psychophysics and Physiology*, ed. J. C. Armington, J. Krauskopf & B. R. Wooten, pp. 221–31. New York: Academic Press.

Crook, J. M., Lange-Malecki, B., Lee, B. B. & Valberg, A. (1988) Visual resolution of macaque retinal ganglion cells. *J. Physiol. (Lond.)*, **396**, 205–24.

Derrington, A. M. & Lennie, P. (1984) Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *J. Physiol. (Lond.)*, **357**, 219–40.

Two other possible explanations can be distinguished. (a) Since the eye exhibits chromatic aberration and since the long- and middle-wave cones both continue to be used for form vision, spatial resolution would be impaired if the peak sensitivities became too separated in the spectrum (Barlow, 1982). (b) There may be no rapid evolutionary route from the ancestral middle-wave opsin to an opsin with an amino-acid sequence that gives a peak sensitivity lying below 530 nm. These explanations, and the one given in the text, are not, of course, exclusive.

Despite the relatively limited range of values of the ratio M:L, it may still be the case that no individual Pβ ganglion cell accommodates the full range. Whereas the retinal ganglion cells that draw inputs from short-wave cones form a very homogeneous group, a wide variation has been reported in the relative weightings of long- and middle-wave cone inputs to those colour-opponent units that collectively comprise the second subsystem (Zrenner & Gouras, 1983). This variation is conventionally expressed as a variation in spectral neutral points. The response function of any individual cell may exhibit maximal sensitivity over only a limited range of values of M:L, while the subsystem as a whole, comprising a population of cells, accommodates the full range.

Derrington, A. M., Krauskopf, J. & Lennie, P. (1984) Chromatic mechanisms in lateral geniculate nucleus of macaque. *J. Physiol. (Lond.)*, **357**, 241–65.

Estévez, O. & Cavonius, C. R. (1977) Human color perception and Stiles' π mechanisms. *Vision Res.*, **17**, 417–22.

Farnsworth, D. (1958) A temporal factor in colour discrimination. In *Visual problems of colour*, National Physical Laboratory Symposium No. 8, London: HMSO.

Fischer, F. P., Bouman, M. A. & Ten Doesschate, J. (1952) A case of tritanopy, *Documenta Ophthal.*, **5**, 55–87.

Foster, D. H., Scase, M. O. & Taylor, S. P. (1987) Anomalous loss in blue–green hue discrimination in very brief monochromatic stimuli presented to the normal human eye. *J. Physiol. (Lond.)*, **381**, 64P.

Goodwin, T. W. (1965) *Chemistry of Plant Pigments*. New York: Academic Press.

Gouras, P. (1984) Color vision. In *Progress in Retinal Research*, vol. 3, ed. N. N. Osbourne & G. J. Chader. New York: Pergamon.

Haase, G. von (1934) Bestimmung der Farbtonempfindlichkeit des menschlichen Auges bei verschiedenen Helligkeiten und Sättingungen. Bau eines empfindlichen Farbpyrometers. *Annalen der Physik*, **20**, 75–105.

Henning, G. B. (1966) Frequency discrimination of random-amplitude tones. *J. Acoust. Soc. Amer.*, **39**, 336–9.

Jacobs, G. H. (1982) *Comparative Color Vision*. New York: Academic Press.

Judd, D. B. & Yonemura, G. T. (1970) CIE 1960 UCS diagram and the Müller theory of color vision. *Proceedings of the International Color Association, Stockholm, Sweden, 1969*, pp. 266–74. Göttingen: Munsterschmidt.

König, A. & Dieterici, C. (1884) Über die Empfindlichkeit des normalen Auges für Wellenlängenunterschiede des Lichtes. *Annalen der Physik*, **22**, 579–89.

Krauskopf, J. & Mollon, J. D. (1971) The independence of the temporal integration properties of individual chromatic mechanisms in the human eye. *J. Physiol. (Lond.)*, **219**, 611–23.

Ladd-Franklin, C. (1892) A new theory of light sensation. In *International Congress of Psychology, 2nd Congress*. London. (Kraus reprint, 1974).

Laurens, H. & Hamilton, W. F. (1923) The sensibility of the eye to differences in wave-length. *Amer. J. Physiol.*, **65**, 547–68.

Lennie, P. & D'Zmura, M. (1988) Mechanisms of color vision. *CRC Critical Reviews*. **3**, 333–400.

Mariani, A. P. (1984) Bipolar cells in monkey retina selective for the cones likely to be blue-sensitive. *Nature*, **308**, 184–6.

McCree, K. J. (1960) Small-field tritanopia and the effects of voluntary fixation, *Optica Acta*, **7**, 317–23.

Mollon, J. D. (1986) Molecular genetics: understanding colour vision. *Nature*, **321**, 12–13.

Mollon, J. D., Bowmaker, J. K. & Jacobs, G. H. (1984) Variations of colour vision in a New World primate can be explained by polymorphism of retinal photopigments. *Proc. Roy. Soc., Lond.*, **B222**, 373–99.

Mollon, J. D. & Cavonius, C. R. (1987) The chromatic antagonisms of opponent process theory are not the same as those revealed in studies of detection and discrimination. In *Colour Deficiencies VIII*, ed. G. Verriest (Documenta Ophthalmologica Proceedings Series 46). The Hague: Martinus Nijhoff.

Mollon, J. D. & Estévez (1988) Tyndall's paradox of hue discrimination. *J. Opt. Soc. Am. A*, **5**, 151–9.

Mollon, J. D. & Jordan, G. (in press) Eine evolutionäre Interpretation des menschlichen Farbensehens *Die Farbe*.

Mollon, J. D. & Polden, P. G. (1977) Further anomalies of the blue mechanism. *Invest. Ophthal. and Vis. Sci. (Suppl.)*, **1b**, 140.

Moore, B. C. J., Glasberg, B. R. & Shailer, M. J. (1984) Frequency and intensity difference limens for harmonics within complex tones. *J. Acoust. Soc. Amer.*, **75**, 550–61.

Nathans, J., Thomas, D. & Hogness, D. S. (1986a) Molecular genetics of human color vision: the genes encoding blue, green and red pigments. *Science*, **232**, 192–202.

Nathans, J., Piantanida, T. P., Eddy, R. L., Shows, T. B. & Hogness, D. S. (1986b) Molecular genetics of inherited variation in human color vision. *Science*, **232**, 203–10.

Nunn, B. J., Schnapf, J. L. & Baylor, D. A. (1985) The action spectra of rods and red- and green-sensitive cones of the monkey *Macaca fascicularis*. In *Central and Peripheral Mechanisms of Colour Vision*, ed. D. Ottoson & S. Zeki. London: Macmillan.

Pitt, F. H. G. (1944) The nature of normal trichromatic and dichromatic vision. *Proc. Roy. Soc. Lond.*, **B132**, 101–17.

Polden, P. G. & Mollon, J. D. (1980) Reversed effect of adapting stimuli on visual sensitivity. *Proc. Roy. Soc. Lond.*, **B210**, 235–72.

Shapley, R. & Perry, V. H. (1986) Cat and monkey retinal ganglion cells and their visual functional roles. *Trends Neurosci.*, **9**, 229–35.

Stücker, N. (1908) Über die Unterschiedsempfindlichkeit für Tonhöher in verschiedenen Tonregionen. *Z. f. Sinnesphysiologie*, **42**, 392–408.

Tansley, B. W. & Boynton, R. M. (1976) A line, not a space, represents visual distinctness of borders formed by different colors. *Science*, **191**, 954–7.

Thoma, W. & Scheibner, H. (1980) Die spektrale tritanopische Sättigungsfunktion beschreibt die spektrale Distinktibilität. *Farbe und Design*, **17**, 49–52.

Vos, J. J. & Walraven, P. L. (1972) An analytical description of the line element in the zone-fluctuation model of colour vision – I. Basic concepts. *Vision Res.*, **12**, 1327–43.

Walraven, P. L. & Bouman, M. A. (1966) Fluctuation theory of colour discrimination of normal trichromats. *Vision Res.*, **6**, 567–86.

Weale, R. A. (1951) Hue discrimination in para-central parts of the human retina measured at different luminance levels. *J. Physiol. (Lond.)*, **113**, 115–22.

Willmer, E. N. & Wright, W. D. (1945) Colour sensitivity of the fovea centralis. *Nature*, **156**, 119.

Wright, W. D. (1952) The characteristics of tritanopia. *J. Opt. Soc. Am.*, **42**, 509.

Wright, W. D. & Pitt, F. H. G. (1935) Hue-discrimination in normal colour-vision. *Proc. Physical. Soc.*, **46**, 459–73.
Wyszecki, G. & Stiles, W. S. (1982) *Color Science*. New York: Wiley.
Zrenner, E. & Gouras, P. (1983) Cone opponency in tonic ganglion cells and its variation with eccentricity in rhesus monkey retina. In *Color Vision: Physiology and Psychophysics*, ed. J. D. Mollon & L. T. Sharpe, pp. 211–23. London: Academic Press.