A PSYCHOPHYSICAL STUDY OF THE VALIDITY OF ANOMALOSCOPIC ASSESSMENTS OF COLOUR VISION IN THE BLUE-GREEN RANGE

Jennifer Stillman
School of Psychology, Massey University Albany, Private Bag 102-904, North Shore MSC, Auckland, New Zealand. J.A.Stillman@massey.ac.nz

Abstract

The Roland IF-2 all-colour anomaloscope is used for assessment of genetic or acquired colour vision deficits. The device uses either the Rayleigh (red/green) or Moreland (green/blue) equations and requests forced-choice same-different judgements of the hue of light stimuli on either side of a 2-deg bipartite field. Validating the output of the anomaloscope presents some psychophysical challenges. These were met through the use of a modified same-different signal detection paradigm in which all stimuli within the reported matching range were assumed to be perceptually the same and distinct from all stimuli outside of the matching range. The diagnostic output, at least in the blue-green range with the procedure utilised by the anomaloscope, may not provide accurate assessment of individual differences in colour perception, because performance is likely to be confounded with response bias.

During an investigation of colour vision and contrast-sensitivity in dental professionals, questions were raised about the range of metameric matches reported by an anomaloscope utilising the Moreland equation. For clinical and research purposes it is necessary to be confident that reported differences among individuals with acquired colour vision deficits reflect actual differences in visual functioning. The reported anomalous quotient and matching range is obtained after relatively few trials, spread over a significant range of wavelength mixtures. Confirming the veracity of a range of subjective matches presents some psychophysical challenges. In order to meet these, physically different stimuli reported to be metameric matches were treated as “same” pairs within a same-different signal detection paradigm.

Method

Five volunteers, one man and three women, aged between 21 and 41 years of age, took part in the study; each had normal colour vision as assessed using the Farnsworth-Munsell 100-Hue test. Monocular assessments were performed using the anomaloscope in its automatic mode for acquired colour vision deficits. The stimulus was presented as a 2-deg circular bipartite field in which a constant stimulus on the right-hand side (desaturated blue, 480 nm and 580 nm) was presented for comparison with a variable stimulus on the left-hand side (green 490 nm and blue 436 nm). The amount of green in the comparison was varied between 1% and 99%. Alternate trials converge on 50% green from opposite ends of the presentation range. The observer’s task is to register whether the two halves of the bipartite field are of the same, or different, hues. Usually the device outputs a result after between 30 and 40 trials.

The calculated mid-point of the matching range is expected to be at around 50% green. Maaranen, Tuppurainen, and Mätyjärvi (2000) considered the normal matching range in the Moreland equation to be up to 11%, although the manufacturers of the Roland devise suggested the normal range should be no wider than 20% (J. Neukam, personal communication, September 27, 2006). Sommerhalder, Pelizzone, and Roth (1997) found a 7-deg field was required to obtain population variability in the matching range comparable to that for a Rayleigh equation obtained using a 2-deg field.

Following three automatic determinations conducted as described, the anomaloscope was used in its manual mode, with 30 trials at each of 17 green/blue comparison stimulus values presented for judgement in random order over a number of sessions. The percent green in the mixture was varied in fifteen 5% steps from 15% to 85% green. Additionally, to ensure an adequate number of trials in the matching range, 30 trials each of 48% and 52% green were included. Observers rated their certainty that the two halves of the bipartite field were of the same or different hues using the following scale: 1. certain different, 2. fairly certain different, 3. unsure, probably different, 4. unsure, probably same, 5. fairly certain same, 6. certain same. Comparison stimuli with a green percent reported to fall within the individual's matching range were assumed to form "same" pairs when presented for comparison with the constant stimulus, those outside of this range were assumed to form "different" pairs.

Results

Receiver operating characteristics (ROC) were fitted to the rating data from each observer, on the assumption that judgements were made using a differencing strategy (Macmillan and Creelman, 2005). For these analyses, comparison stimuli with green percentages which produced only category 1 (certain different) responses were omitted. The obtained sensitivity indices (d'), and chi-square (χ^2) goodness of fit statistics with their associated probabilities, are given in Table 1, along with each observer's matching range (average of 3 determinations). Although observers were presented with identical sets of 510 comparisons, the signal probability (where signal=same pairs) varies for each observer, because it is a function of the matching range, and of the number of trials included.

Table 1. Matching range, signal probability, number of trials, d', and χ^2 goodness of fit statistic with the associated probability, for each observer.

Subject	Matching Range (% green)	p(signal)	n	d'	χ^2_4	p
1	49 - 54	0.31	390	3.406	*0.910	.492
2	45 - 51	0.50	240	2.479	0.870	.648
3	37 - 55	0.54	330	3.957	0.868	.868
4	43 - 55	0.50	300	3.718	0.939	.405
5	45 - 54	0.38	390	3.485	10.592	.032

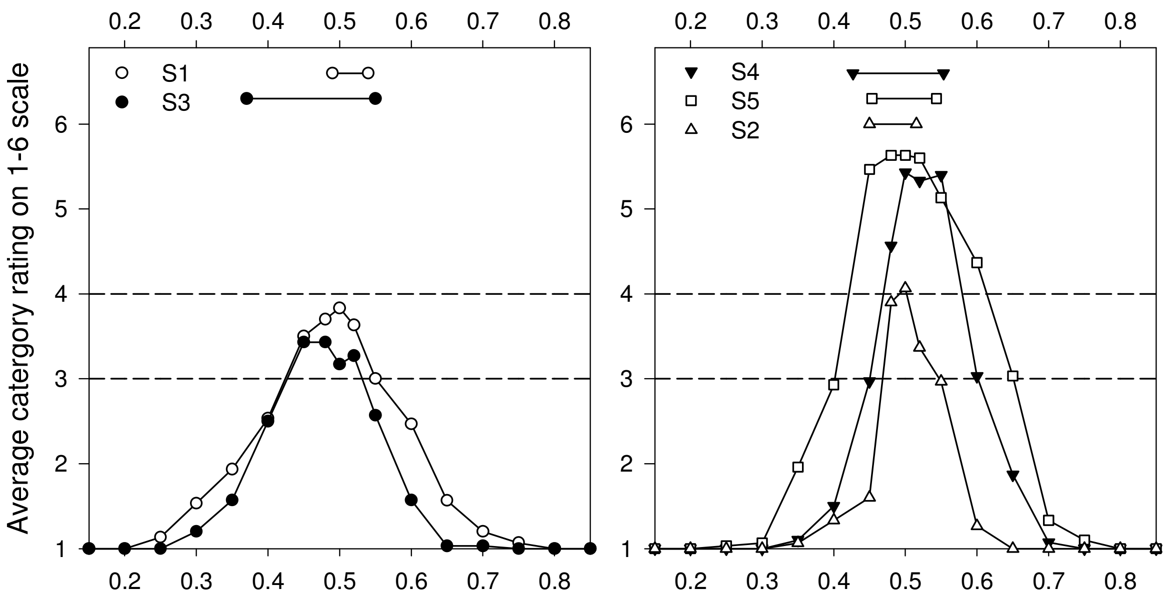
*=χ^2_3. Category 6 was never used by this subject, reducing the degrees of freedom.

d' was also calculated when each remaining stimulus between 35% and 65% green was presumed to be distinguishable from 50% green, the normative mid-point of the matching range. Comparisons of wavelength mixtures with < 35%, or > 65% green were not included because they invariably resulted in excellent discrimination and thus added little information. ROCs could not be fitted to some data close to 50% green for S1 and S5 because few response categories were used. In these cases, single-point d' estimates were obtained by treating ratings of 1, 2, and 3 as one response alternative ("different") and ratings of 4, 5, and 6 as the other ("same"). Results are given in Table 2.

Table 2. Discrimination, d', between a 50% green mixture and mixtures with other green percentages (n=60 trials). The χ^2 goodness of fit probabilities are given in brackets. Where rating categories were collapsed, the associated χ^2 degrees of freedom are indicated as follows: *=3df , **=2df, otherwise df=4. Single-point d' estimates are shown in square brackets.

% green	d' (p)				
	S1	S2	S3	S4	S5
35	5.00 (.59)*	5.00 (.49)	3.32 (.43)*	5.00 (.49)	3.72 (.18)
40	3.99 (.97)*	4.78 (.95)	2.29 (.75)*	5.00 (.98)	3.24 (.89)
45	[1.6]	1.94 (.44)	0.42 (.24)	3.54 (.52)	[0]
48	[0.94]	1.62 (.83)	0.90 (.48)*	1.98 (.98)*	[0]
52	[0.78]	0.53 (.07)*	0.45 (.15)*	0.58 (.91)**	0.38 (.51)
55	[2.22]	1.90 (.98)	2.20 (.83)*	0.51 (.08)**	1.18 (.37)
60	4.01 (.31)*	5.00 (.86)	5.00 (.95)*	3.88 (.99)	1.91 (.40)
65	5.00 (.98)*	5.00 (.78)	5.00 (.58)*	5.00 (.73)	3.04 (.29)

In order to examine individual response biases, the relationship between the observer's average category ratings and the matching range reported by the anomaloscope was explored. The data are displayed in Figure 1. Two panels are used for clarity. Observers are identified by individual symbols, and the matching range is shown by the horizontal bars at the top of each panel. The dashed lines delineate the equivocal categories "unsure probably different" (rating 3), and "unsure probably same" (rating 4).



Proportion of green in comparison field

Figure 1. Average category use as a function of the percentage of green in the comparison stimulus.

Discussion

The ROC analyses in Table 2, although based on small numbers of trials, suggest the reported matching ranges may lack precision, because there are instances of significant discrimination between stimuli within some observers' matching ranges. To illustrate this graphically, Figure 2 displays the best-filling ROC curves (refer Tables 1 and 2) for two observers for whom the anomaloscope reported a small (S2), and a large (S3), matching range.

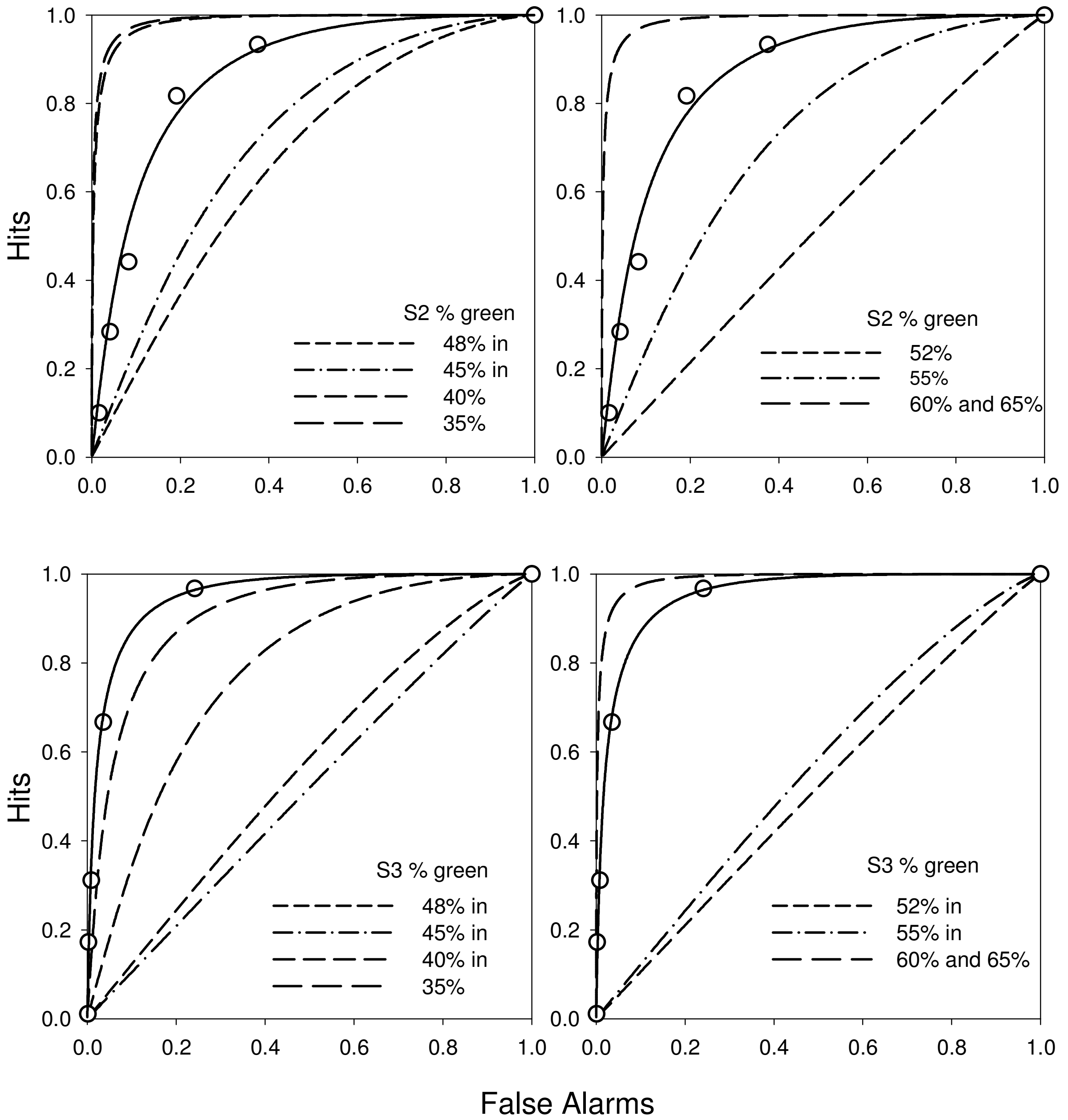


Figure 2. Best fitting ROCs to the data from two observers rating the hue of various wavelength mixtures as the "same" or "different" from 50% green. Also shown are the best fitting ROCs and data points (solid lines and open circles) from Table 1. Comparison stimuli within an individual's matching range are indicated in the figure legends.

In Figure 2 the solid lines and data points, repeated in both panels, show the overall performance of each observer as detailed in Table 1. The remaining curves are the ROCs listed in Table 2, when the "same" stimulus was taken to be 50% green, and the "different" stimulus comprised specified green percentages from 35% to 65%. For each observer, comparison stimuli with less than 50% green are represented in the left-hand panel, and those with more than 50% green are represented in the right hand panel. In each pair of panels ROCs with identical line styles represent comparison stimuli that were equally distant from a 50/50 mixture.

Some indication of the observers' perceptual experiences can be gained from the rating behaviours depicted in Figure 1. Rating responses are fairly symmetrical around a maximum which, apart from a minor exception for S3, occurs at 50% green, the expected mid-point of the matching range for normal observers.

On the assumption that category use is fairly consistent for an individual, changes in rating behaviour in response to various comparison stimuli should reflect differences in perceptual experiences. For example, note the differences for S3 and S5 between the average ratings for a 40% and 60% comparison stimulus (Figure 1), and between the associated d' values (Table 2). In general the rating responses and d' values do not correspond tightly to the reported width, symmetry, and location of the reported matching range.

With respect to bias, observers displayed different degrees of conservatism, but when uncertain, showed similar tendencies to choose 'uncertain probably different' over 'uncertain probably same.' This is contradictory to previous research in other modalities (Macmillan & Creelman, 2005; Stillman, Brown, & Troscianko, 2000). A possible reason for this may be that in the entire set of 510 trials, the probability of same pairs was relatively low. If the reported matching ranges are presumed to be accurate, the probability of same pairs would be in the range .23 to .35.

Taken together the outcomes of the study suggest that, notwithstanding the usefulness of anomaloscopic assessments for the diagnosis of colour blindness, there are some limitations with respect to investigations where, for example, it is desirable to track small progressive changes in colour vision over time. The psychophysical procedure currently employed by the instrument is likely to result in a reported matching range and anomalous quotient that is confounded by the response biases of individuals, particularly over the few trials in the range of interest. It should be possible for such instruments to employ more reliable psychophysical procedures to specify the borders of the matching range with greater precision. Although Table 2 suggests a marked change in perception with changes in wavelength composition, the observed rating behaviour suggests the transition from matching to non-matching is less than abrupt. If this is the case, it would be useful to define the borders of the matching range in terms of a chosen level of performance, evaluated over a greater number of trials. This would be most efficiently accomplished through the use of adaptive psychophysical procedures converging on a specified percentage of judgements of difference on both sides of the expected 50/50 mid point.

References

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. (2nd ed.). NJ: Lawrence Earlbaum.

Maaranen, T.H., Tuppurainen, K.T., and & Mäntyjärvi, M I. (2000). Color vision defects after central serous chorioretinopathy, *Retina, 20*, 633-637.

Sommerhalder, J., Pelizzone, M., and Roth, A. (1997). Stimulation parameters for automatic examination of color vision. Klin Monatsbl Augenheilkd, 210, 319-324. Abstract retrieved January 11, 2007 from PubMed database.

Stillman, J.A., Brown, G.M., & Troscianko, T. (2000). Influence of sensitivity on response bias in taste and audition. *Perception & Psychophysics, 62*, 1645-1654.

Acknowledgements

This research was support by a grant from the New Zealand Dental Research Foundation. Thanks and due to Michael Hautus, Department of Psychology, University of Auckland, New Zealand, who generously made available his computer program to fit the Receiver Operating Characteristic Curves.