Analysis of Rayleigh match data with psychometric functions

William H. Swanson
Retina Foundation of the Southwest, Suite 400, 9900 North Central Expressway, Dallas, Texas 75231, and
Department of Ophthalmology, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines
Boulevard, Dallas, Texas 75235

Received July 14, 1992; revised manuscript received January 21, 1993; accepted January 26, 1993

Color matches have been used for a variety of purposes, yet the psychometric properties of color-matching data have not been thoroughly investigated. A method is given for generating psychometric functions for the two ends of the color-matching range by use of a perceptual dimension for stimulus magnitude based on ratios of cone quantal catches. The analysis was applied to Rayleigh match data gathered from 250 naïve observers with an automated protocol. Slopes of the psychometric functions were significantly shallower for anomalous trichromats than for normal trichromats, consistent with the assumption that stimulus magnitude is based on ratios of cone quantal catches. These results indicate that the tester's criterion for response consistency can strongly affect Rayleigh match widths. The analysis may also be useful for other perceptual tasks, such as contrast matching and spatial alignment.

INTRODUCTION

Color-matching data have been used for more than a century to study abnormalities in color vision. In 1881, Lord Rayleigh reported matching data for red–green color mixtures compared with a yellow standard. He noted that there was significant variability among normals and that color defective observers accepted matches outside the normal range. Twenty-six years later Nagel designed a clinical instrument for color matching that he named the anomaloscope. Since then a variety of color-mixture equations and anomaloscope designs have been introduced.

Anomaloscopes have been used for a variety of tasks beyond identifying congenital color defects. Studies of Rayleigh match variation among normal trichromats have provided evidence for polymorphisms of human-cone photopigments, and comparisons of Rayleigh matches with genetic data have provided insight into relationships between gene sequences and perception. Studies of patients with retinal disease have shown that Rayleigh matches can be used to detect abnormalities in cone-photopigment regeneration and that abnormal Rayleigh matches can result from early age-related macular changes. Evaluation of Rayleigh matches of patients with a range of eye diseases has shown that optic-nerve diseases can produce widened matching ranges with normal luminosity and that retinal diseases can produce widened matching ranges with abnormal luminosity.

The usefulness of anomaloscopy in this range of endeavors is due in large part to the fact that reliable data can be gathered quickly from inexperienced observers, provided that the tester has been thoroughly trained and follows a standardized protocol. In a typical Rayleigh match two contiguous semicircles are presented: a standard field filled with a single monochromatic light (yellow) and a mixture field filled with two monochromatic lights (red and green). The tester's goal is to determine the range of red–green ratios for which the observer perceives the mixture field as matching the monochromatic field in color and brightness. The extremes of this range are the match end points, the middle of the range is the match midpoint, and the difference between the two end points is the match width. By systematically presenting different mixture ratios and noting the observer's responses, the tester can typically determine the two match end points in a few minutes.

Despite the widespread use of color matching, there has been little effort to investigate it as a psychophysical task. This may be due to the fact that in color matching the stimulus magnitude varies along a perceptual, rather than a physical, dimension. The stimulus is the chromatic difference between the mixture field and the standard field. Since the two fields are never physically identical, the match is perceptual and not physical. Whereas the use of a perceptual dimension for stimulus magnitude is not unique to color matching, it is quite different from the large majority of psychophysical tasks that employ stimuli that vary along a physical dimension of intensity (related tasks are mentioned below, in the Discussion section). In the latter case psychometric functions can be measured by determining probability of detection as a function of stimulus intensity; stimuli with a zero intensity can be used for blank trials, either to measure the false-alarm rate or to construct criteria-free methods (such as two-alternative forced choice). For instance, in measuring acuity the intensity is stimulus size, in measuring contrast thresholds the intensity is stimulus contrast, and in measuring spatial-frequency discrimination the intensity is difference in spatial frequency. For all these types of intensity, one end of the physical scale is zero intensity and the other end is maximum intensity.

The analysis of observer consistency in color matching is complicated by the fact that blank trials cannot be used. If the observer gives inconsistent responses (i.e., gives different responses to repeated presentations of the same
mixture or gives a response of “redder” to a mixture that has less red light than does a previous mixture to which the response was “greener”), then the trained tester must rely on decision criteria based on testing expertise to determine the matching range end points. Since the number of trials is small, it is not possible to quantify response consistency. These factors may reduce test–retest reliability and may introduce a source of test-site dependence for color-matching data.

Recently, automated protocols that use computer-controlled anomaloscopes have been introduced that offer the advantage of greater standardization and thus the potential for more widespread clinical use of anomaloscopy. Large numbers of trials and finer control over step size provide the potential for analysis of response consistency. The current study evaluated a large clinical database gathered with an automated protocol. A method was developed for analyzing color-matching data in terms of psychometric functions by use of a perceptual stimulus dimension based on ratios of cone quantal catches. Response consistency was analyzed in terms of the slopes of these psychometric functions, and matching range end points were determined as the 50% points of the functions.

METHODS

Observers
A total of 250 naïve observers participated in the study (data from laboratory personnel were excluded). Observers ranged in age from 4 to 79 years [mean ± 1 standard deviation (SD) = 26 ± 19 years]. There were 120 observers free of ocular disease (83 normal trichromats, 26 deuteranomalous trichromats, and 11 protanomalous trichromats) and 130 patients referred by ophthalmologists specializing in retinal diseases or glaucoma (patients with congenital color defects were excluded from this analysis). All observers were tested with the use of 2-deg fields; for a separate study of the effect of field size, 30 normal trichromats free of eye disease were also tested with a range of field sizes.

Apparatus
Rayleigh matches were measured with a custom-made computer-controlled briefcase anomaloscope, which uses a combination of light-emitting diodes (LED’s) and interference filters in Maxwellian view (with a 2-mm artificial pupil) to yield narrow-band primaries of 667, 588, and 551 nm. The left semicircle that is presented is filled with a red–green mixture (667 and 551 nm), and the right semicircle is filled with a yellow standard (588 nm).

The experiment was controlled by a Macintosh II computer equipped with a six-channel 12-bit digital-to-analog converter (DAC) board (National Instruments NA-AO-6) and a three-port digital input/output board (National Instruments NB-DIO-24). Each LED was driven by a linear current amplifier that in turn received input from one 12-bit DAC. For finer control over the red and green LED’s, their DAC’s had variable voltage ranges that were set by one additional DAC each.

Procedure
For each trial, the stimulus was presented continuously until the observer pushed one of three response buttons: either the left or the right button to indicate the left or the right semicircle as redder or else the middle button to indicate a perfect match. A bidirectional switch permitted control over the radiance of the yellow standard. For data collection on young children, the child described the colors seen on each trial and the tester pushed the appropriate button.

The strategy of asking the observer to decide which side was redder was adopted because of two problems experienced in early tests that used the more usual task of asking whether the mixture field was redder or greener than the standard. First, it was noted that some patients were not able to respond properly with color names of redder or greener yet were still able to define a range of mixture fields that matched the standard (e.g., some of these patients had severe tritan defects and referred to the color of the green LED as white). By asking the patients to determine which of the two sides was redder, we eliminated reference to apparent color, and the patient was required only to refer to color differences. Second, in the middle of the test some naïve normals would forget which of the buttons indicated which color and would begin pushing a button indicating a response opposite to the one that they intended. The strategy of asking which side was redder minimized this problem, since the buttons were aligned so that the left button meant left redder and the right button meant right redder.

The mean retinal illuminance of the red–green mixture field was held constant at 57 trolands (Td) (comparable with the 3 cd/m² of the Nagel model I anomaloscope), and if necessary the observer adjusted the radiance of the yellow standard. As with the Nagel anomaloscope, for all mixtures the radiances of the LED’s were such that, for deuteranopes, the brightness of the mixture matched the brightness of the standard. In this way R + G was held constant across all mixtures, and the stimuli are expressed in units of \( R/(R + G) \) ranging from 0.0 (pure green, with no red light) to 1.0 (pure red, with no green light). The initial brightness matches of the standard to the red and green primaries were used to determine whether an observer had deutan, protan, or scotopic spectral sensitivity. The anomaloscope was then run in deutan mode, proton mode, or scotopic mode. In deutan mode the retinal illuminance of the standard was always 57 Td. In protan and scotopic modes the radiance of the standard was varied according to protanopic or scotopic spectral sensitivity. These modes worked well for most observers, but some observers did require adjustments to the radiance of the standard on each trial; when this occurred it lengthened the duration of the session.

A staircase procedure was employed to determine the two ends of the matching range by use of two interleaved staircases, with one staircase controlling the mixture field for even-numbered trials and the other controlling the mixture field for odd-numbered trials. Each staircase began at either the red or the green end of the range, with a step size of half the range. At each reversal of the staircase the step size was halved, until a step size of 0.003 was obtained. The staircase continued until six reversals were obtained at this step size. Each staircase followed a “left redder” response by making the mixture field greener for the next presentation in the staircase and followed a “right greener” response by making the mixture
field redder for the next presentation in the staircase. The two staircases differed in how they followed a “match” response: the staircase for the even-numbered trials made the next presentation redder, whereas the staircase for the odd-numbered trials made it greener. Thus the staircases converged on the two match end points (computed as the means of the last six reversals for each staircase). The mean (±1 SD) number of trials for the 250 observers was 62 (±17); with normal adults this typically required 2 to 5 min, although more time was required for children and for some color-defective observers. An example is shown in Fig. 1.

Data Analysis
To analyze the data in greater detail, two yes–no psychometric functions were constructed for each data set by combining responses for both staircases: a G function for the question “Is the mixture greener than the standard?” and an R function for the question “Is the mixture redder than the standard?” The green (E_G) and red (E_R) ends of the matching range were determined from these functions. A response of “right redder” was coded as “yes” for the G function and “no” for the R function. A response of “left redder” was coded as “no” for the G function and “yes” for the R function. A response of “match” was coded as “no” for both functions. For each function the fraction of “yes” responses was computed for every mixture condition presented. Figure 2 shows the two psychometric functions generated from the staircases shown in Fig. 1.

Since the difference in color between the red–green mixture and the yellow standard is a perceptual rather than a physical dimension for stimulus magnitude, it is necessary to construct an appropriate stimulus metric. The metric used was based on the assumption that the match midpoint is the mixture ratio at which the quantic catch ratio for the long-wavelength-sensitive and middle-wavelength-sensitive cones equals the cone quantic catch ratio for the yellow standard. The magnitude of the chromatic difference between any given mixture field and the standard is therefore determined by the difference between the cone quantic catch ratios for the mixture and for the match midpoint.

Two constraints are placed on the stimulus metric. First, the metric must be referenced to the match end point, so that the end point is the 50% point of the corresponding psychometric function. For stimulus magnitude x, the probability P(x) of an observer's perceiving the color difference (i.e., perceiving the test field as either redder or greener than the standard, depending on which function was being evaluated) was computed by using

\[ P(x) = 1 - 2^{-x^\beta} \]  

where \( \beta \) is a parameter that determines the slope of the function. Therefore the metric must be constructed so that \( x = 1 \) when the mixture value equals the end point. Second, the metric must be such that the slope of the psychometric function is independent of the match midpoint, so that when the match midpoint is changed the function is simply moved horizontally along the stimulus axis without change in shape. These two constraints were satisfied by setting stimulus magnitude \( x \) equal to the antilogarithm of the difference between test mixture \( M \) and the corresponding end of the matching range:

\[ x = 10^{E_G - M} \quad \text{for the } G \text{ function,} \]  
\[ x = 10^{E_R - M} \quad \text{for the } R \text{ function.} \]

Usually the slope of a psychometric function is considered to reflect effects of internal noise. However, in this case the slope could also be affected by changes in cone photopigments. If the photopigments have altered cone quantic catches (as a result of either abnormal photopigments or reduced cone optical density), then the ratios of cone quantic catches for the test and standard hemifields may change more slowly with mixture ratio than for normal cone quantic catches. This would effectively minify the stimulus magnitude, reducing the slope by a constant factor.

During testing, some observers reported that they occasionally pushed the wrong button by mistake; others may have made such errors without being aware of it. Fitting such data with a function that asymptotes to 100% would result in a shallow slope. To avoid this potential source of shallow slopes, the parameter \( \gamma \) was introduced to denote the upper asymptote:

\[ \text{fraction of “yes” responses} = \gamma P(x). \]  

To fit the data, psychometric functions were constructed for a range of values of the following parameters: \( E_G \) and \( E_R \) from 0.0 to 1.0 in steps of 0.0025, \( \log(\beta) \) from 0.3 to 2.25 in steps of 0.15, and \( \gamma \) from 90% to 100% in steps of 2%. The likelihood of obtaining the data set was calculated for each parameter set, and the parameter set with the highest likelihood was used as the maximum-likelihood estimate. Details for this type of analysis are given elsewhere.

In Fig. 2 the R function shows considerable response inconsistency (fractions “yes” other than 0.0 or 1.0) when \( 0.55 < R/(R + G) < 0.58 \). This region is shown in the inset, along with the \( G \) and \( R \) functions with the highest
Fig. 2. R and G functions derived from staircase data shown in Fig. 1. Open circles connected by dashed lines give the fractions “yes” for the G function, and filled circles connected by solid lines give the fractions “yes” for the R function. Note that the R function is somewhat erratic just below $R/(R+G) = 0.6$. Since only a small number of trials (1–5) were presented for each mixture (there were a total of 52 trials across all mixtures), a few mistakes in button pushing could have produced this amount of noise in the data; the observer reported after the experiment had concluded that she had mistakenly pushed the wrong button on a few occasions. For $R/(R+G) = 0.578$, the observer made two responses of “same” early in the staircases and one response of “redder” later on, giving 33% for fraction “yes.” For $R/(R+G) = 0.569$, the observer responded “same” on one trial and “redder” on the other. Responses were “redder” for all of the remaining trials for test mixtures between 0.566 and 0.581. The inset shows data from the range $R/(R+G) = 0.54–0.59$, along with the best-fitting psychometric functions. Solid curves are the best-fit functions with all parameters permitted to vary; the dashed curve is the best fit R function when the upper asymptote $\gamma$ was fixed at 100%. These two R functions have similar values for the red end ($E_R$) of the matching range (differing by only 0.5 Nagel unit) but much different slopes (differing by 0.6 log unit).

Simulations

To investigate the reliability of the parameters obtained with the use of the data analysis, Monte Carlo simulations were performed by using model observers. Details of this type of simulation are given elsewhere. Each model observer was defined by a match midpoint and a range and by a slope and an upper asymptote for the two psychometric functions (the same slope and upper asymptote were used for each function).

For each model observer 100 simulated experiments were conducted. Each experiment used the same algorithm as was used in testing patients. Responses of the model observer to a test mixture were generated by selecting one random number (between 0 and 1) for each psychometric function; when the random number was higher than the value of the psychometric function at the test mixture, then this was coded as “higher”; otherwise it was coded as “lower.” If both of the psychometric functions yielded “lower,” then a response of “same” was returned. If only one of the psychometric functions yielded “higher,” then the corresponding response (“redder” or “greener”) was returned. If both functions yielded “higher,” then the response returned corresponded to the function that had the greater difference between random number and function value.

For each simulated experiment, results were analyzed by using the same maximum-likelihood procedure used for analyzing patient data. Results were summarized by storing the means and SD’s of the differences between the
parameters obtained by maximum-likelihood estimation and the parameters used to define the model observer. The mean differences provide estimates of the bias in the maximum-likelihood estimates, and the SD's provide estimates of the accuracy of the estimates.

Simulations were performed for a total of 36 different model observers, based on all possible combinations for three different midpoints, two different widths, three different slopes, and two different upper asymptotes. The three midpoints were selected as the mean values for the healthy normal trichromats \( [R/(R + G) = 0.56] \), the deuteranomalous trichromats \( [R/(R + G) = 0.26] \), and the protanomalous trichromats \( [R/(R + G) = 0.86] \). The three slopes were selected as the mean value for all 250 observers \( [\log(\beta) = 1.35] \) and as the values 2 SD's above and below this \( [\log(\beta) = 0.36 \text{ and } 2.34] \). The upper asymptotes were selected as high (100%) and low (90%), and the match widths were selected as narrow (0.01) and wide (0.20).

**RESULTS**

**Psychometric Functions**

The slopes of the psychometric functions showed considerable individual variability within each group, as indicated by the SD's for \( \log(\beta) \) listed in Table 1. There was also one significant between-group difference: congenital color-defective observers tended to have shallower slopes than did normal trichromats. Figure 3 compares values from healthy normal trichromats with values from healthy anomalous trichromats. The difference in mean values for \( \log(\beta) \) was highly significant \((t) = 7.96, p < 0.001\), although the difference in SD's was not \((F) = 1.29, p > 0.20\). Among normal trichromats there were no slope differences between adults with healthy eyes (21 to 66 years old), children with healthy eyes (4 to 10 years old), or adults with eye diseases (21 to 70 years old) (for comparisons among the three groups, in all cases \( t < 0.85, p > 0.50 \) and \( F < 1.67, p > 0.05 \)).

In addition to the slope differences across individuals, there were also slope differences within individuals. The mean values for \( \log(\beta) \) across all 250 data sets were not different for the \( G \) function and the \( R \) function \((t = 0.72, p > 0.50)\). However, for any given individual there were often considerable differences between the slopes of the \( G \) and \( R \) functions: the mean of the absolute value of the difference in \( \log(\beta) \) was \( 0.35 \pm 0.35 \) log unit. Slopes also tended to be shallower for the smallest fields. For the 30 normal trichromats tested with the use of several field sizes, mean \( \log(\beta) \) was \( 1.72 \pm 0.35 \) with a 4-deg field and \( 1.23 \pm 0.45 \) with a 0.5-deg field \((t = 6.53, p < 0.001)\).

For the majority of data sets the upper asymptotes were 100%: 142 (57%) of the data sets had \( \gamma = 100\% \) for both \( G \) and \( R \) functions, and only 5 (2%) of the data sets had \( \gamma < 100\% \) for both \( G \) and \( R \) functions. The mean value \( \pm 1 \) SD for \( \gamma \) was \( 98\% \pm 4\% \) for both the \( G \) and the \( R \) functions. For each of the subsets listed above, the mean value for \( \gamma \) was also \( 98\% \).

**Match Midpoints and Widths**

Overall, there was good agreement between matching ranges determined by the means of the final six reversals of the staircase and by maximum-likelihood estimation. For the following comparisons the \( R/(R + G) \) values were multiplied by 73, yielding units equivalent to those on the Nagel model I anomaloscope. Averaged across all 250 data sets, the difference between maximum-likelihood and mean-of-reversals estimates was \(-0.3 \pm 2.2 \) for \( E_g, 0.2 \pm 1.8 \) for \( E_R, 0.1 \pm 1.5 \) for the midpoint, and 0.5 \pm 2.7 for the width. The small mean differences indicate that there is little tendency for the maximum-likelihood procedure consistently to yield higher or lower estimates than the mean of reversals. On the other hand, the SD's indicate that the difference between the two estimates was large for some data sets. For 192 (77%) of the data sets the difference between the 2 estimates was not >2 Nagel units for both match midpoint and matching range, yet for 45 (18%) of the data sets the difference was >5 Nagel units for either match midpoint or matching range. Table 2 summarizes differences between these two groups of data sets. For the data sets in which the two estimates were in poor agreement, slopes were shallower, upper asymptotes were lower, and numbers of trials were greater (indicating that the staircases took longer to obtain the required number of reversals); for all parameters the differences were significant (in all cases \( t > 2.7, p < 0.01 \)).

Linear regression was used to search for effects of age on results from the 83 normal observers. There was no effect of age on either \( \log(\beta) \) (coefficient of correlation \((r) = 0.13, p > 0.05\)) or \( \gamma \) (\( r = 0.03, p > 0.50\)). Both match width (\( r = 0.34, p < 0.002\)) and match midpoint (\( r = 0.52, p < 0.001\)) tended to decrease with age, as shown in Fig. 4. More-detailed analyses were conducted by com-

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>( \log(\beta) )</th>
<th>Age (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No eye disease</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal trichromats</td>
<td>83</td>
<td>1.45 ± 0.40</td>
<td>19 ± 16</td>
</tr>
<tr>
<td>Anomalous trichromats</td>
<td>37</td>
<td>0.97 ± 0.46</td>
<td>15 ± 13</td>
</tr>
<tr>
<td>No congenital color defect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy adults</td>
<td>30</td>
<td>1.51 ± 0.41</td>
<td>37 ± 13</td>
</tr>
<tr>
<td>Adult patients</td>
<td>96</td>
<td>1.46 ± 0.49</td>
<td>42 ± 14</td>
</tr>
<tr>
<td>Healthy children</td>
<td>43</td>
<td>1.43 ± 0.38</td>
<td>7 ± 2</td>
</tr>
</tbody>
</table>

Fig. 3. Distributions for slopes of psychometric functions for normal and anomalous trichromats (both groups free of eye disease). The mean slope was significantly lower for anomalous trichromats, although the SD's were not different.
comparing the data for the 47 children of ages 4–13 years with the data for the 36 observers of ages 16–66 years. There was no effect of age on match width within either age group \((r < 0.15, p > 0.20)\), whereas the difference in match widths between the two age groups was highly significant \((t = 5.02, p < 0.001)\). In addition, the 3- to 13-year-old age group had much greater variability in match widths than did the 16- to 66-year-old age group \((F = 6.59, p < 0.001)\). There was no effect of age on match midpoint within either age group \((r < 0.25, p > 0.10)\), whereas the difference in match midpoints between the two age groups was highly significant \((t = 6.23, p < 0.001)\). There was no difference between the two age groups in the variability of the match midpoints \((F = 1.54, p > 0.10)\).

For the 30 normal observers tested with 5 different field diameters, as field diameter decreased the match width increased and the match midpoint decreased, as shown in Fig. 5. Linear regression across all 150 data points showed that these effects were significant (regression against log field diameter, \(r = 0.26\) for midpoint, \(r = 0.48\) for width; in both cases \(p < 0.001)\). The mean value for the midpoint decreased monotonically from 8 deg to 1 deg and then increased slightly for 0.5 deg. The SD for the midpoint was relatively constant for 1–8 deg (between 0.034 and 0.044) but was much larger for 0.5 deg (0.072). The mean value for the match width was approximately equal at 8 and 4 deg, increased slightly as diameter decreased down to 1 deg, and then increased dramatically at 0.5 deg. The SD for the match width also increased dramatically at 0.5 deg.

Simulations

To evaluate the effect of variations in the upper asymptote \(\gamma\) on estimations of slope \(\beta\), results were compared for two different analyses: one permitting all three parameters to vary to fit the data and one keeping \(\gamma\) fixed at 100%. These simulations were run for match midpoint = 0.56, range = 0.01, and all three slopes. When the upper asymptote for the simulated observer was 100%, both analyses yielded similar results for \(\log(\beta)\); for both bias and accuracy, maximum differences did not exceed 0.08 log unit and mean differences were 0.03 log unit. However, when the analysis in which the upper asymptote was fixed at 100% was compared with that in which the upper asymptote was permitted to vary, for the condition when the upper asymptote for the simulated observer was 90%, the analysis with upper asymptote fixed at 100% increased the bias of the slope by as much as 0.6 log unit (underestimating slope) and worsened the accuracy of the slope estimate by as much as 0.4 log unit. This establishes that fixing the upper asymptote at 100% can lead to underestimates of slope.

Once the necessity for permitting the upper asymptote to vary was established, simulations were performed for all 36 model observers (all combinations of 3 midpoints, 2 widths, 3 slopes, and 2 upper asymptotes). The primary factor affecting the bias and accuracy of the parameter estimates was the slope: as slope decreased, bias increased and accuracy worsened. Upper asymptote was a secondary factor: when the model observer had \(\gamma = 90\%), the accuracy of the width estimate was worse, the bias of the slope estimate was increased, and the accuracy of the asymptote estimate was worse. Match midpoint affected only the bias of the midpoint estimate (the bias was
larger for the two anomalous midpoints than for the normal midpoint). Width affected only the accuracy of the slope estimate (accuracy was better for the smaller width) and the bias of the width estimate (bias was smaller for the larger width).

To evaluate bias and accuracy, the means and SD's of the differences between actual and estimated parameters were computed in Nagel units (0.5 to 1 Nagel unit is the resolution limit of most manual anomaloscopes). For model observers with the midpoint of a normal trichromat, in all cases the bias of the midpoint estimate was <0.7 Nagel unit. However, for some of the model observers the accuracy of the midpoint estimate was poor: the SD's of the estimates were as large as 4.5 Nagel units for model observers with \( \log(3) = 0.36 \), dropping to 1.1 Nagel units for model observers with \( \log(3) > 0.36 \) and to 0.6 Nagel unit for model observers with \( \gamma = 100\% \) and \( \log(3) > 0.36 \). For model observers with anomalous midpoints, the bias averaged 1.7 times larger, whereas accuracy was unchanged.

For the width estimates, in many cases there was bias toward overestimating width (underestimates of width were never >0.5 Nagel unit). For model observers with the smaller width, for \( \log(3) = 0.36 \), bias was as large as 12 Nagel units and accuracy as poor as 6 Nagel units, whereas, for \( \log(3) > 0.36 \), bias was never >1.3 Nagel units and accuracy never worse than 1.3 Nagel units. For model observers with the wider width, the bias was usually half that for the model observers, with the smaller width and otherwise identical parameters.

Slope estimates tended to be good for model observers with \( \log(3) = 0.36 \), with bias never >0.1 log unit (overestimated) nor accuracy worse than 0.5 log unit. For model observers with \( \log(3) = 1.35 \), slope tended to be slightly underestimated, with bias never >0.2 log unit nor accuracy worse than 0.6 log unit. For model observers with \( \log(3) = 2.34 \), slope was more severely underestimated: bias was as large as 1 log unit and accuracy as poor as 0.8 log unit. Upper (lower) limits for the estimated slopes were computed as the mean ± 2 SD’s. For model observers with \( \log(3) = 0.36 \), the upper limit was 1.1 log units. For model observers with \( \log(3) = 2.34 \), with the smaller actual width the lower limit was always >1.2 log units, but for the larger width the lower limit decreased dramatically, to as low as 0.1. This means that a steep estimated slope rules out a shallow actual slope, but when the estimated width is high a shallow estimated slope does not rule out a steep actual slope.

Upper-asymptote estimates were fairly good when the actual asymptote was 100%, with the bias never >1.8% and the accuracy never worse than 3.5%. However, when the actual asymptote was 90% the bias was as large as 8% and the accuracy as poor as 5%. This means that if the mean estimated asymptote for a population is near 90% then the actual asymptote is below 100%, but a high mean estimated asymptote does not rule out a low actual asymptote.

**Discussion**

The results presented here demonstrate that color-matching data can be analyzed by means of psychometric functions that use a perceptual dimension based on ratios of cone quantal catches. This analysis interprets inconsistent responses to repeated presentation of a given stimulus value as indicating that this stimulus value lies under the ascending portion of one of the psychometric functions, whereas stimulus values yielding highly consistent responses are under the upper or lower asymptotes of both functions. This approach provides a method for evaluating and interpreting response consistency in color-matching data that not only should reduce test-site variability but also should permit more detailed investigation of the causes of reduced chromatic discrimination.

Analysis of data from 250 naïve observers shows that there is considerable within-group variability in slopes of the psychometric function and that for a given individual the slopes may vary across field sizes or between the two ends of the matching range. This indicates that the method used for evaluating response consistency can be important in interpreting Rayleigh match data.

**Validity of the Analysis**

The finding that anomalous trichromats have shallower slopes than do normal trichromats, despite similar SD's
for the two distributions of slopes, provides strong support for the assumption that perceptual stimulus magnitude is related to the difference in cone quantal catches for the mixture field and the yellow standard. For anomalous trichromats, the rate of change in cone quantal catch ratio versus color mixture is much lower than for normal trichromats.\textsuperscript{13} This means that a mixture that is a given distance in Nagel units from a match end point should yield a much lower perceptual magnitude for an anomalous trichromat than for a normal trichromat, shifting the distributions of slopes to lower values without affecting the SD. This is exactly what was found in the current study, as illustrated in Fig. 3. Pokorny and Smith\textsuperscript{13} give an estimate of ~0.4 log unit for the effect of deuteranomaly on the rate of change in cone quantal catch ratio versus color mixture; this is quite close to the 0.5-log-unit difference found in the current study for mean values for the slopes of anomalous versus normal trichromats.

The mean-of-reversals estimates were in general quite similar to the maximum-likelihood estimates, indicating that the two staircases tended to converge on the 50\% points of the two psychometric functions. This is similar to the situation with the more common experimental technique of using a staircase to manipulate physical stimulus intensity: a 1-up--1-down yes/no staircase estimates the intensity that is detected 50\% of the time. For the small fraction of data sets in which there was a large discrepancy between the staircase estimates and the maximum-likelihood estimates, the slopes tended to be shallower and the upper asymptotes lower; for experiments manipulating physical stimulus intensity, these factors have been shown to increase the difference in the two types of threshold estimate.\textsuperscript{12}

The match midpoints and widths obtained from normal trichromats free of ocular disease show systematic effects of age and field size, indicating that the technique is useful for measuring subtle effects on Rayleigh matches in populations of inexperienced observers.

Simulations
Results of simulations indicate that shallow slopes increase the bias of the width estimate (causing overestimates) but not of the midpoint estimate and that they decrease the accuracy of both the midpoint and the width estimates. Therefore it is useful to be able to identify data sets generated by shallow slopes. Permitting the upper asymptote to vary provides better estimates of the actual slope than does using a fixed upper asymptote. Simulation results indicate that data sets can be excluded as unreliable by setting a lower limit to acceptable slopes but that this will exclude some data sets for observers with shallow slopes and large match widths. As can be seen in Fig. 3, a lower limit of \(\log(\beta) = 1.0\) (suggested by the simulations) would exclude few normal trichromats but many anomalous trichromats.

Effect of Tester Criterion
The ability of the analysis to make match width relatively independent of the slopes of psychometric functions makes it possible to distinguish between reduced chromatic discrimination resulting from postreceptoral factors and reduced perceptual stimulus magnitude resulting from altered cone quantal catches. The 50\% points on the psychometric functions do not vary with changes in slope, so match widths measured at the 50\% points should not be affected by changes in cone spectral sensitivities. In manual anomaloscopy there is typically no method for determining what points on the psychometric functions are evaluated, since the number of trials is usually small and the rules for handling variable responses are not always explicit. If a value other than the 50\% point on the psychometric functions is determined, then match width should be dependent on the slopes of the psychometric functions.

The fact that the current study determined the 50\% points of the psychometric functions may explain why the matching ranges obtained for normal trichromats and anomalous trichromats were in general smaller than expected from the results of population studies that used manual anomaloscopy. The following are comparisons with results of Helve's large population study\textsuperscript{14} which are typical for the Nagel model I anomaloscope.\textsuperscript{4} For the current study 17 (46\%) of the anomalous trichromats had match widths <3 Nagel units, whereas Helve reported very few anomalous trichromats with match widths this small. For the 83 normal trichromats in the current study, match widths were also in general smaller than expected from Helve's data: for 25\%, match width was <1 Nagel unit, and for 48\%, match width was <2 Nagel units, compared with <5\% and <25\%, respectively, from Helve's data.

To model the possible performance of these observers under manual anomaloscopy with a stricter criterion for consistent responses, match widths for all observers tested with a 2-deg field were recomputed by use of the 80\% points on the psychometric functions. The match widths of the anomalous trichromats increased by an average of 8.9 ± 6.8 Nagel units, so that only 3 (8\%) had widths <3 Nagel units. In contrast, for normal trichromats the widths increased by only 2.8 ± 2.6 Nagel units (also bringing their data closer to those of Helve). This criterion change also caused small changes in the match midpoints: the mean of the absolute value of the change in the match midpoint was 0.4 ± 0.9 Nagel unit for the normal trichromats and 0.8 ± 1.1 Nagel units for the anomalous trichromats.

Limitations of the Analysis
The current study considers response variability in terms of psychometric functions based on cone quantal catch ratios. It does not account for the fact that an observer's response might not be "match" even when the mixture matches the standard. If the "left redder" and "right redder" responses were made randomly whenever the mixture matched the standard, the psychometric functions could be shallow and the width could be artificially narrowed. If the observer always responded "left redder" whenever the mixture matched the standard, then the psychometric functions could be quite steep, the midpoint shifted toward green, and the width artificially narrowed; this could be harder to detect than random guessing. These problems may also affect manual anomaloscopy but may not be obvious because of the
smaller number of trials. Further research is needed to determine how best to handle these problems.

For 5 (2%) of the data sets the value for \( E_0 \) was less than the value for \( \Delta E \), yielding a negative width; of these, the estimated slopes were shallow and the largest magnitude of the negative width was 2 Nagel units. This indicates that the estimates cannot be accurate, since match widths should never be less than zero. Of the 3,600 simulated experiments, there were also 5 (0.1%) for which the estimated width was less than zero; in all cases these were for the shallowest slope and the narrowest width. For the 5 simulated experiments with negative widths, the largest magnitude was also 2 Nagel units, the estimated slopes were shallow, and the largest error in estimate of the midpoint was as great as 12 Nagel units. This indicates that a negative estimated match width is evidence of an unreliable data set.

Advantages over Conventional Anomaloscopy
An advantage of the proposed analysis is its method for handling situations in which the observer never reports a color match. In conventional anomaloscopy, data cannot be obtained when patients never report a match. In the current study several patients (and some normal observers tested with a large field) remarked that the two sides never matched. The protocol for the briefcase anomaloscope estimates the ends of the matching range with the use of two staircases, even if a response of "same" is never used. The psychometric-function analysis used in the current study also produces match end points even if "same" is not used. The relatively large number of trials makes it possible to evaluate response consistency, and if consistency is high the match end points are meaningful even if the observer never reports a perceptual match: the match end points indicate transitions from the range of color mixtures over which the observer does not give consistent responses to the ranges of color mixtures over which the observer gives highly consistent responses. The two ranges over which the observer gives highly consistent responses define the color mixtures that the observer can reliably discriminate as redder than the standard (in one range) or greener than the standard (in the other range). This provides information about both the photopigments and the postreceptoral processing.

If the range of inconsistent responses lies entirely within the region of normal Rayleigh matches, then the observer cannot have abnormal photopigments. For instance, one of the normal observers who never accepted a match was highly consistent, responding "left redder" for all mixture ratios \( >0.561 \) and "right redder" for all mixture ratios \( <0.556 \). The region of inconsistent responses was only 0.4 Nagel unit wide; this observer appears to have normal photopigments and excellent discrimination, as verified on other color-vision tests. If the range of inconsistent responses lies entirely within the region of matches made by simple deuteranomalous trichromats and the region of reliable "mixture redder than standard" responses includes all the region of normal Rayleigh matches, then this is strong evidence that the observer is deuteranomalous. If the range of inconsistent responses is quite narrow, then the observer probably has good color discrimination and does not have a significant red-green defect.

Conventional anomaloscopy requires that the patient make a brightness adjustment on each trial. Nagel designed his anomaloscope to be in deutan mode, so that the color mixtures produced a constant quantal catch for an average deuteranope; these are also brightness matches for most normal trichromats and deuteranomalous trichromats. This means that, for deuteranopes, deuteranomalous trichromats, and normal trichromats, at most only small brightness mismatches will be present, and the observer never needs to make large adjustments of the brightness of the standard. In comparison, protanopes, protanomalous trichromats, and monochromats need to make large adjustments to the brightness. If brightness matches are not made, such observers can use brightness cues to make matches, artificially narrowing the matching range. In the current study several protanomalous trichromats were intentionally retested in deutan mode, for which their matching ranges were smaller than when tested in protan mode and for which their match midpoints were closer to normal.

In the current study the computer estimated the observer's spectral-luminosity function based on initial brightness matches (matching the standard to the green and red primaries) and selected one of three brightness conditions: deutan, protan, or scotopic. For the remaining presentations the radiance of the standard was adjusted to be a deutan, a protan, or a scotopic match to the test mixture, minimizing the need for observers to make brightness matches. This was particularly useful in testing young children for whom it was not possible to obtain reliable brightness matches because the appropriate brightness mode was determined from performance on a battery of color-vision tests (including flicker photometry). Since brightness matches that are significantly different from deutan, protan, or scotopic matches are of diagnostic utility for acquired defects, automated anomaloscopy could be improved by including a method for individualizing the brightness mode for each observer, continually estimating the observer's spectral sensitivity by analyzing brightness matches throughout the test.

Clinical Considerations
Clinically, anomaloscopy is used to diagnose congenital color defects and to evaluate acquired color-vision defects. For these purposes manual anomaloscopy by a skilled and trained tester may be preferable to automated anomaloscopy, since it permits personal interaction with an experienced tester. The analysis presented in the current study provides a formal method for evaluating and interpreting response consistency and can be applied to data from manual anomaloscopy if enough trials are presented and a careful record is kept of observer responses. For clinical research, the standardization provided by automated anomaloscopy may be desirable. Since in most cases the staircase and maximum-likelihood estimates were similar, the simpler staircase procedure should offer most of the benefits of the maximum-likelihood estimation.

The psychometric functions are defined in terms of questions regarding whether the mixture is redder or greener than the standard. It is important to note that these were not the questions put to the observer but are
rather ways of interpreting the observer’s responses. As mentioned above (in the Procedure subsection), the observer was actually asked which side was redder. In dealing with some patients, particularly those with tritan defects, it may be impossible for the patient to respond in terms of “mixture redder” and “mixture greener.”

Match width provides an index of red-green discrimination that can be useful in clinical studies, and use of a standardized method for evaluating response consistency should improve the accuracy of this index. For the 130 patients with eye disease in the current study, when the end points were recomputed for the 80% points of the psychometric functions the match width increased by 4.3 ± 4.9 Nagel units. This illustrates the importance of systematizing the calculation of match width when studying a clinical population.

The data from 83 normals tested with a 2-deg field showed effects of age on both match width and match midpoint. The decrease in match width with age appears to be due to the fact that many young children had large match widths. For the 4–13-year-old age group, the SD of the match width was much greater than for the 16–66-year-old age group. This suggests that children are much more variable than are adults in either their criteria for a match or their color-discrimination abilities. In comparison, SD’s in match midpoints were similar for both children and adults, indicating that the age effect was not due to greater variability than among the children. This suggests that children have either lower photopigment optical density or less dense preretinal filters than do adults. Whereas the sources of these age effects are not clear, it appears that in clinical use the norms for children should be established separately from the norms for adults.

For data from 30 normals tested with a range of field diameters, match width increased and match midpoint decreased as field size decreased. The increase in match width with decrease in field diameter is consistent with the known dependence of chromatic discrimination on field size. The decrease in match midpoint with decrease in field diameter is a phenomenon known as the color-match-area effect and has previously been used with manual anomaloscope to document photoreceptor changes induced by retinal disease. The results indicate that for field sizes of 1–8 deg the protocol is suitable for measuring the color-match-area effect in a naïve population. However, with the 0.5-deg field the variability in both match width and match midpoint was greater, mean match width was dramatically larger, and, instead of decreasing further, the mean match midpoint actually increased slightly. Estimated slopes were shallower for the 0.5-deg field, indicating less reliable parameter estimates. These results suggest that, for clinical studies of the color-match-area effect, field diameters smaller than 1 deg may not yield reliable data.

Comparisons with Psychometric Functions for Other Perceptual Tasks
Related analyses have previously been performed with perceptual tasks such as vernier acuity and hue judgments. In some cases a perceptual dimension of stimulus magnitude has been evaluated by use of a psychometric function based on a physical aspect of the stimulus. However, such metrics may not be appropriate. Whereas the current study is not directly applicable to tasks other than color matching, the basic principles and conclusions may be relevant.

Studies of spatial distortion in amblyopia have found that, for some amblyopes, test objects that are physically aligned may subjectively appear to be misaligned and some that are misaligned may appear to be aligned. Study of this phenomenon involves analysis of a perceptual stimulus offset in terms of physical stimulus offset. Rentschler and Hilz plotted fraction of “left” responses as a function of linear vernier offset and fitted the data with a psychometric function whose 50% point determined the positional offset and whose slope determined vernier acuity. Their analysis carries the implicit assumption that detectability of a vernier offset is a linear function of physical displacement. If the relation is actually more complex (such as logarithmic), then the shapes of the psychometric functions fitted in this way would become distorted for subjects with large differences between perceived alignment and physical alignment. As with the current study, if the observer adopted a strategy of giving a fixed response whenever the stimulus magnitude was nondetectable, the midpoint estimate could be biased. For vernier data, if the observer responded “left” whenever the stimuli seemed aligned, then the 50% point of the psychometric function would be at the left end of the range of physical locations that appeared subjectively aligned, and the SD could be quite small. Thus estimates of both offset and vernier acuity could be affected by observer strategy.

Unique-hue measurements require adjustment of a physical parameter (wavelength) to obtain a specific perception (unique hue). Schefrin and Werner constructed psychometric functions for unique-hue judgments. For unique green they asked the subject to respond either “blue” or “yellow” and plotted the proportion of “yellow” responses as a function of wavelength. They then defined unique green as the 50% point of this function. This analysis involves the implicit assumption that perceptual intensity is a linear function of wavelength. If this is not true, then the shapes of the psychometric functions could be distorted. Schefrin and Werner acknowledged this and performed control experiments to determine the effect of the use of a wavelength axis. However, the problem of observer strategy remains: if an observer always responded “yellow” when the hue appeared neither yellow nor green, then the unique-hue estimate could be affected.

A range of studies has measured perceived contrast, perceived spatial frequency, and perceived temporal frequency, often by using adaptive staircase procedures. When comparisons are made between eyes or between different retinal regions in the same eye, physically identical stimuli may be perceptually different, and the situation is similar to that in studies of spatial distortion. When different spatial or temporal frequencies are compared at the same retinal location (for instance, in contrast matching), the stimuli are never physically identical and usually are not perceptually identical, and a single perceptual attribute is compared. This is similar to color matching, except that in color matching it is often possible to produce perceptually identical stimuli. As with the color-matching data discussed in this paper, for these other
types of data it may be useful to analyze the staircases by use of psychometric functions if appropriate metrics can be found.

ACKNOWLEDGMENTS

This research was supported in part by National Eye Institute grant EY07716, by the Home Interiors and Gifts Fund of the Communities Foundation of Texas, Inc., and by the Delta Gamma Foundation of Dallas.

REFERENCES AND NOTES

10. Base 2 is used for convenience, so that the 50% point is at \( x = 1 \).
11. Complete consistency of responses would yield an infinite slope, and high levels of inconsistency could yield slopes near zero, so permitting extremely high and low values for \( \log(\beta) \) could skew the means and produce large SD’s. The limited range of 0.90–2.25 for \( \log(\beta) \) was used to avoid this problem. Across all 500 values obtained for \( \log(\beta) \), 17 (3%) were 0.30 and 37 (7%) were 2.25, indicating that the range was indeed limited, but not severely so. Since low values of \( \gamma \) could permit higher values for \( \log(\beta) \), if \( \gamma \) had been permitted to be quite low this could have skewed the values of \( \log(\beta) \) to higher values. To minimize this problem the lowest value that was allowed for \( \gamma = 0.90 \), since it seems unlikely that the wrong button was pushed more than 10% of the time. Out of the 500 values obtained for \( \gamma \), 94 (19%) were 0.90, indicating that such a lower limit was needed. Given the limited range of values of \( \log(\beta) \) and the use of the \( \gamma \) parameter, the SD for \( \log(\beta) \) can be considered a conservative estimate of individual differences in slopes of the psychometric functions.
16. Implementing this may be complex since, for patients with contrast defects, brightness matches can be highly variable and, at the extremes of the matching range, the color difference increases apparent brightness for normals, causing a nonlinear relation between brightness and luminance.