A Model for the Observer on the Farnsworth-Munsell 100-Hue Test

B. J. Craven

Purpose. To use a theoretical model of the observer on the Farnsworth-Munsell 100-hue test to estimate the magnitude of random variation in 100-Hue test error scores.

Methods. The model was based upon classical signal detection theory. Results from the model were obtained by computer simulation.

Results. There is a fairly regular relationship between mean test scores over many tests and the standard deviation of those scores. This relationship is for practical purposes unaffected by polarity in the observer’s hue discrimination and by changes in the detailed assumptions of the model.

Conclusion. The model provides a flexible tool for further theoretical research into the 100-hue test. Invest Ophthalmol Vis Sci. 1993;34:507-511.
open to some question. Victor’s procedure chose pairs of caps to be swapped entirely at random, arguing that successive swaps are independent events. However, consider as an example the following cap order: 1 2 3 6 5 4 7 8 9. If successive swaps are independent, the next swap is just as likely to involve caps 4 and 7 as it is to involve caps 1 and 2. However, the difference in hue between caps 4 and 7 is greater than that between caps 1 and 2 (by a factor of 3, if we assume hue differences are equal and additive), and a subject whose hue discrimination was uniform around the hue circle would be more likely to swap caps 1 and 2 (resulting in an error score of 12) than to swap caps 4 and 7 (resulting in an error score of 16). Thus, we can expect the errors produced by human subjects to be more evenly distributed than those obtained by independent swappings of caps, with a resulting effect on the scores obtained.

Three cautions must be added. First, the very fact that cap 4 has been moved up the order (giving us evidence that the subject has misassessed the hue of cap 4) makes us more confident that the subject should want to swap it with cap 7 than we would have been had cap 4 been in its rightful position (that swaps only of nearest neighbors are allowed does not affect the argument). In other words, the difference in the probabilities of swapping caps 1 and 2 and caps 4 and 7 may not be as great as suggested above. The second caution is that modeling incorrect cap arrangements by repeated swapping of caps should not be taken as implying that a real subject does it that way. However, the example given does suggest that a real subject is likely to distribute errors more uniformly than does Victor’s cap-swapping process. The final caution is that the 100-hue test is such a complicated system that calculating the effects (on score variability) of differential discrimination performance along other perceptual dimensions. Many different theoretical models have been used to try to account for the pattern of subjects’ behavior in standard psychophysical tasks. One of the most widespread and influential models is that of signal detection theory (SDT). The popularity of SDT can be partly accounted for by the ease and elegance with which it provides models for a very wide range of different psychophysical tasks, using essentially identical assumptions in each case. To the author’s knowledge, no previous attempt has been made to apply the SDT approach to the 100-hue test.

Outline of Signal Detection Theory Approach

Whatever the particular psychophysical task, signal detection theory makes the fundamental assumption that somewhere within the subject, the relevant attributes of each of the stimuli is quantitatively represented as the value of a variable (internal to the subject), and that the subject’s decision on a given trial is based solely upon the values of these variables. The theory assumes further that this quantitative representation is subject to random errors—that is, a given stimulus may elicit a different internal value each time it is presented. In other words, the encoding of the relevant stimulus dimensions is subject to noise. Thus, one stimulus that is physically greater than another stimulus may on some occasions elicit a smaller internal value and thus be perceived as being smaller. It is this noise that limits performance in discrimination tasks and sets what is commonly called the threshold for the discrimination in question. In all cases, the subject is assumed to apply a fixed decision rule to determine from the values of the internal variables the response that is most likely to be correct.

The same fundamental assumptions can be applied without modification to model a subject performing the 100-hue test. We assume that the hue of each cap is represented by the value of an internal variable, and that the subject sorts the caps according to these internal representations. Because the internal representation of hue is subject to random error, the ordering produced by the subject will not necessarily be the correct one, and the greater the magnitude of the random error, the more haphazard the cap order produced by the subject. An inevitable consequence of this process is that subjects’ scores on the test, being determined in part by random processes, will show...
random variation. Indeed, if a subject consistently ordered the caps in the same incorrect order, we would be forced to conclude that although the subject’s color matching performance was abnormal, his or her hue discrimination was very good. Imperfect hue discrimination necessarily implies variation in test scores from test to test. The amount of random variation (we will consider the standard deviation of a subject’s test scores) is likely to depend upon the absolute level of performance. The following model was constructed with a goal of determining the relationship between the mean score over many tests and the standard deviation of the scores.

METHODS

We assume that the color of each cap $i$ (where the cap number $i$ takes values from 1 to 85) is encoded by the value of some internal 1-dimensional variable $X_i$ (although color space is three-dimensional, the cap sequence of the FM 100-hue test is one-dimensional). If the subject encoded color perfectly, without random error, the value of each $X_i$ would be $i$ (this choice is merely for convenience; any regular spacing would do). With the introduction of random errors, each $X_i$ is perturbed by the addition of a random variable $r$ drawn from a Gaussian distribution of zero mean and standard deviation $\sigma$. Values of $r$ are drawn independently for each cap. The standard deviation $\sigma$ is a free parameter; by varying $\sigma$ we can manipulate the absolute level of performance. It will be shown later that the choice of a Gaussian distribution of $r$ is not crucial.

On each test, each cap $i$ is assigned a value $X_i$, as just described, and the caps within each box are sorted into increasing order of $X_i$. Because of the random component of the $X_i$’s, the sorted order may not be the correct order. The sorted caps are then scored according to the Farnsworth convention, exactly as if a human subject had ordered the caps. By repeating this procedure many times, adequately precise values of the mean score and the standard deviation of scores can be obtained.

On a computer, the above procedure was performed for many values of the internal standard deviation $\sigma$ to obtain means and standard deviations of scores corresponding to a wide range of performance levels.

RESULTS

Figure 1 shows the relationship between mean score (on the x axis) and standard deviation of scores (on the y axis). Both axes are logarithmic. Note that neither axis represents an independent variable: The true independent variable is the internal standard deviation $\sigma$. The different points on the graph were obtained using different values of $\sigma$. Each point is the result of 10,000 simulated runs of the 100-hue test. The data points fall on a slight curve. For scores of 100 or less, the relationship could be conveniently summarized by saying that the standard deviation of scores on the 100-hue test for a single consistent subject is close to twice the square root of the mean score for that subject.

Effect of the Shape of the Internal Noise Distribution

So far, it has been assumed that the random errors with which the cap hues are encoded are normally distributed, but this need not be the case. To see whether the assumption of Gaussian noise is crucial, the procedure was repeated, but with the Gaussian noise distributions replaced by distributions of different shapes. The shapes used were a rectangular distribution, and the highly asymmetric distribution of the product of pairs of numbers drawn from a rectangular distribution, which rises abruptly to a sharp peak on the low side and declines gradually on the high side. This choice of shapes was atheoretical and arbitrary, but the intention was to use distributions highly unlike the original Gaussian distribution. Based on Figure 2, in which results for all three distribution shapes are plotted on the same axes, the shape of the internal noise distribution is not critical in determining the relationship between the mean and standard deviation of 100-hue test scores.

Effect of Polarity in the 100-hue Plot

It is common for color-deficient subjects to distribute their errors nonuniformly among the 85 caps in the test. This nonuniformity was modeled by varying the standard deviation $\sigma$ of the internal noise distribution...
FIGURE 2. As for Figure 1, but with data plotted for three different shapes of the subject's internal error distribution. The relationship between mean and SD is not greatly affected by the different assumptions.

sinusoidally as a function of cap number. For a given maximum value of \( \sigma \), which we will call \( \sigma_{\text{max}} \), the function is

\[
\sigma = \sigma_{\text{max}}(1 + \sin(2\pi iN/85)),
\]

where \( i \) denotes the cap number and \( N \) the number of complete cycles of the sinusoid in 85 caps. Thus, the minimum value of \( \sigma \) was always zero. Values of \( N \) of 1, 2, and 4 cycles were used, and \( \sigma_{\text{max}} \) was varied to vary the level of performance.

In Figure 3, the results for the three values of \( N \) are plotted with data for constant \( \sigma \) (as in Fig. 1). The three curves for the sinusoidally varying errors lie closely on top of each other and differ only slightly from the curve for uniform errors. The relationship between the mean and standard deviation of test scores is not affected to any significant extent by polarity in the subject's hue discrimination.

Consistency With Empirical Data

The results in Figures 1, 2, and 3 show there is a regular relationship between the mean score on the FM 100-hue test and the standard deviation of the scores. The relationship is robust with respect to the only true free parameter of the model—the shape of the internal noise distribution—and with respect to nonuniformities in the subject's performance. But we do not yet know whether the model provides a good prediction of any data we might collect from human subjects. There are arguments both that we should expect greater variability from human subjects and that we should expect lesser variability from human subjects. These arguments are covered in the remainder of this section, followed by a brief comparison with data from human subjects.

The model of the subject assumes there are no sources of random fluctuation of scores apart from the inevitable random perturbations of the internal encodings of the colors of the caps. In practice, this assumption may be untenable. The subject's performance may be affected by motivational, physiological, diurnal, or even (if the conditions are not well controlled) photometric variables. It would be very diffi-
cul to quantitatively assess the effects of these factors. All that can be said is that there are reasons to believe the standard deviations of test scores obtained in practice may be higher than those suggested by theory.

However, it is possible that as well as suffering from random errors of encoding, the subject also makes systematic errors. For example, there may be two caps that the subject finds essentially indistinguishable on the basis of hue and hence should order randomly from test to test. If, however, there is a detectable brightness difference between the caps, the subject may order them on the basis of brightness, without realizing he or she is doing so. The order chosen by the subject for the two caps may be consistently wrong. In such cases, the variation in test scores will be less than that expected on the basis of theory.

Given that the theory provides neither an upper nor a lower bound on the variability to be expected in practice, it becomes worthwhile to look at empirical data to see how well it conforms to the predictions of the model. The study of the relationship between mean score on the test and standard deviation of scores is methodologically difficult (because of possible practice effects) and tedious. However, one such study has been reported by Chisholm. Figure 4 shows the actual standard deviations obtained by Chisholm for each of nine subjects and the predicted standard deviations obtained from the subjects' mean scores. A two-tailed t-test was performed upon the fractional differences between the predicted and empirical data for each subject. The difference was not statistically significant (t = 1.35, 8 degrees of freedom; P > 0.2). Although the sample size is small, approximate calculation of the power of the test indicates that if the discrepancy between theory and data were 20% (for example), and the critical significance level was chosen to be 0.05, the probability of correctly rejecting the null hypothesis would be about 0.85. It is clear that the standard deviations predicted by the model do not differ dramatically from those found in practice.

**DISCUSSION**

Figure 5 shows Victor's predictions of error score standard deviation plotted with the data from Figure 1. Any difference between the two sets of results is of negligible practical significance. The two approaches to modeling performance on the 100-hue test therefore can be regarded as lending support to each other. Why the agreement should be so good is not clear. There appears to be some underlying regularity in the way that cap arrangements are converted to test scores that determines the variability in scores almost independently of any assumptions made in the models.

As well as confirming an existing result, the signal detection theoretical model of the 100-hue test observer is valuable in itself. It goes beyond Victor's model in that it models the processes within the subject that limit performance, rather than using an arbitrary method of generating cap arrangements. It yields the result that the variability in scores is not affected by polarity in the distribution of errors. Finally, the techniques involved have had a very successful history in other areas of perceptual research and should provide a useful and flexible tool for further research into the 100-hue test.

**Key Words**

Farnsworth-Munsell 100-hue test, random errors, signal detection theory.

**References**