A Second Box-End Scoring Artifact in the Farnsworth-Munsell 100-Hue Test

B. J. Craven

Purpose. To investigate and describe a hitherto unreported scoring artifact in the Farnsworth-Munsell 100-hue Test, arising from the grouping of the caps into four boxes, which causes caps near the ends of a box to score less than caps near the center of a box. This artifact is in addition to a previously reported one, which causes caps near the end of a box to score more than caps in the center of the box.

Methods. Two different statistical simulations were used to generate synthetic cap sequences, which were scored in the normal way.

Results. For error scores less than about 500, the new artifact, which depresses scores at the ends of boxes, was found to dominate the pattern of scores.

Conclusion. The existing published correction for the box-end scoring artifact is inappropriate for scores less than about 500, and therefore should be applied cautiously. Invest Ophthalmol Vis Sci 1993;34:503–506.

Although they make up a complete, unbroken hue circle, the 85 caps of the Farnsworth-Munsell 100-hue test1 are, for physical convenience, divided into four separate boxes. The subject is not permitted to transfer caps between boxes. As Aspinall2 has pointed out, this restriction on the placing of the caps by the subject produces an artifactual nonuniformity in the distribution of error scores for the caps, which in some circumstances might lead to false diagnosis of certain color vision deficiencies. Victor3 assessed the extent of this nonuniformity by considering a very large number of random arrangements of the caps. The average scores for caps situated at the ends of boxes were almost twice those for caps situated in the middle of boxes.

Victor proposed a correction for the nonuniformity of cap scores based upon the distribution of scores obtained for completely random arrangements of caps. His suggestion was to calculate a relative error score by dividing each cap’s score by the score expected for that cap based on completely random cap arrangements. He stated that such a correction was rigorously justified only in the limit of near-random cap arrangements, but that for lower scores, one could make a correction intermediate between the full correction and no correction at all. (He also argued that this second refinement was, for most purposes, unnecessary.)

Unfortunately, however, the assumption that the box-end artifact for low error scores is simply a scaled down version of the artifact for very high error scores is not correct. In this report, I will show that the pattern of the box-end artifact changes qualitatively as the general level of performance changes, and that applying the suggested correction to cap arrangements scoring 500 or less would exacerbate the artifactual nonuniformity rather than correct for it.

Box-End Artifacts

There are at least two different kinds of box-end artifact in the scoring of the 100-hue test. The first, which
was reported previously, causes an increase in scores for caps whose correct positions are near the box ends, the mechanism of which is best explained with an example. Consider a box with caps numbered 1 to 21 in it, with the caps randomly arranged. The nearest neighbors of cap 1 (assuming that cap 1 is not placed at either end of the box) will be drawn from a population of caps (numbers 2 to 21) whose cap numbers differ from 1 by an average of 10.5. The nearest neighbors of cap 11 (again, not placed at a box-end) will be drawn from a population (caps 1 to 10 and 12 to 21) whose cap numbers differ from 11 by an average of 5.5. Thus, for randomly placed caps, the expected score for cap 1 is almost twice that for cap 11. When the caps happen to be placed exactly at the box-ends, the argument is slightly altered, but the result is very similar. Victor3 covered this artifact in some detail.

The second box-end artifact does not appear to have been reported before and causes a reduction in the average scores of caps at the box ends. Consider the extreme case where the subject swaps a single pair of adjacent caps, obtaining the lowest possible nonzero score of 4. For a subject whose hue discrimination is uniformly good, it is no more likely that the swap would involve one particular pair of adjacent caps than any other. To work out the pattern of average cap scores over a large number of repetitions of the test, we can, in this simple case, calculate the scores corresponding to all possible swaps of two adjacent caps and average the scores obtained for each cap. Now, a cap placed in the interior of a box can participate in two different swaps, one with its left-hand neighbor and one with its right-hand neighbor. In contrast, a cap at the very end of a box can participate in only one swap—that with its single neighbor in the same box. The consequence is that caps in the interior of boxes are twice as likely to be swapped with their neighbors as caps at the ends of boxes. It follows that the scores of box-end caps will be depressed compared to those of caps from the interior of the box.

**Simulation of 100-Hue Scores**

As the score rises from the minimum nonzero score of 4, it becomes impractical to enumerate each possible arrangement of the caps. Figure 1 shows the results of a computer simulation in which disordered cap arrangements were created by successively swapping randomly selected nearest-neighbor pairs (a method used by Victor3). The degree of disorder was varied by varying the number of nearest-neighbor swaps made within each box. The average scores for each cap are plotted for each degree of disorder. It can be seen that for very small numbers of swaps the scores are lowered toward the end of each box. As the number of swaps rises, this lowering of scores becomes more prominent. In addition, however, a small increase in scores starts to become apparent at the box-ends, presumably corresponding to the previously reported artifact. As the number of swaps rises further, this end-of-box increase in scores rises to completely dominate the distribution of scores. Note that it is only for scores well in excess of 500 that the end-box caps score higher than those in the interior of the boxes.

As a check that the result obtained was not an artifact from using nearest-neighbor swaps to disorder the cap arrangement, the result was verified by a radically different simulation in which disorder was obtained by assuming a Gaussian random error in the subject's assessment of the hue of each cap. (For details of methods, see Craven.4) These results are plotted in Figure 2. To make it easier to compare the data with that plotted in Figure 1, the magnitudes of the random error were chosen to yield similar overall mean cap scores to the data obtained by the swapping method. The pattern of cap scores in Figure 2 differs only in detail from that in Figure 1.

Figure 3 shows data for cap arrangements calculated as for Figure 2, but with the cap scores ordered as advocated by Kinnear5. In this system, the scores obtained for the caps are plotted in order of the caps as ordered by the subject, not in numerical order of cap number. Using the Kinnear convention, the artifact described in this report dominates the shape of the score distribution for all scores less than about 900. In addition, it can be seen that the previously reported artifact is now restricted to the extreme end positions of each box.

![Figure 1: Mean cap score (averaged over many tests) plotted as a function of cap number, at a range of levels of performance.](https://iovs.arvojournals.org/pdfaccess.ashx?url=/data/journals/iovs/933397/)
Victor suggested a correction for artifacts in which each cap score is divided by the corresponding score for that cap expected on the basis of completely random cap arrangements. Figure 4 shows such relative error scores plotted as a function of cap number. The lines are labelled with the mean total score, as in Figure 1. The cap scores for the completely random performance (labelled 1202) are necessarily all 1.0. Application of this correction actually increases the variation of cap score with cap number, even for quite high scores. Scaling the expected scores according to the level of performance (a further refinement suggested by Victor) changes the height of each curve but does not change the ratio of the depth of modulation to the mean score.

Discussion

The 100-hue test is a complex system, and any attempt to explain why one artifact dominates in one range of scores and another dominates in another range of scores must be considered conjectural. However, we can observe that the box-end artifact described in this report occurs because of the initial position of the caps: It is caps whose rightful position is at the ends of boxes that have their scores depressed because of limited mobility. When scores are high, the caps will stray so far from their original positions that the (initially) end-of-box caps will not experience significantly different treatment to caps from the interior of the box, and we would expect this artifact to have very weak effects. The artifact discussed by Victor arises because of differences in the populations of potential nearest neighbors available to each cap. In the explanation of this artifact given earlier, it was assumed that all the caps in a box were equally likely to end up as the nearest neighbors of a given cap (in the limit of very high scores, this is true). However, when the error score is low and the cap arrangements are more nearly correct, the population of potential nearest-neighbor caps to a given cap becomes effectively restricted to those close to that cap in cap number. Thus, the possible variation with position in the box of the difference between the number of a cap and those of its nearest neighbors will become increasingly small as the error score gets smaller. Any artifact resulting from this source of variation therefore will have weak effects where scores are low. Thus, we have a tentative explanation for the dominance of one box-end artifact when scores are low, and another when scores are high.
Summary
The distribution of cap scores in the 100-hue test departs from uniformity, because the caps are grouped into four boxes. Aspinall\(^2\) and Victor\(^3\) have discussed an artifact that arises for very high cap scores, in which caps near the ends of boxes have average scores almost twice those of caps near the middle of boxes. Victor published a method for correcting for this artifact, based upon the pattern of cap scores obtained in the limit of completely random cap arrangements. However, this artifact occurs only when error scores are very high. Where error scores are lower, a second, hitherto unreported artifact occurs, and unfortunately, it results in a lowering of cap scores near the ends of boxes. This artifact becomes dominant for scores of about 500 or less (using the Farnsworth scoring convention). As a result, application of the published correction to scores in this range will exaggerate the artifactual nonuniformity of scores rather than correct for it.

Key Words
color vision, Farnsworth-Munsell 100-hue test, scoring artifacts

References