How psychophysical methods influence optimizations of color difference formulas

**Eric Kirchner,[1,*] Niels Dekker,[1] Marcel Lucassen,[2] Lan Njo,[1] Ivo van der Lans,[1]
Philipp Urban,[3] and Rafael Huertas[4]**

[1]*Color Research Department, AkzoNobel Automotive & Aerospace Coatings, Sassenheim, The Netherlands*
[2]*Lucassen Color Research, Kistenmakerseiland 8, 1121 PD Landsmeer, The Netherlands*
[3]*Fraunhofer Institute for Computer Graphics Research IGD, Fraunhoferstr. 5, 64283 Darmstadt, Germany*
[4]*Departamento de Óptica, Universidad de Granada, Campus Fuentenueva, 18071-Granada, Spain*
**Corresponding author: eric.kirchner@akzonobel.com*

Received August 26, 2014; revised November 24, 2014; accepted January 5, 2015;
posted January 8, 2015 (Doc. ID 221768); published February 2, 2015

For developing color difference formulas, there are several choices to be made on the psychophysical method used for gathering visual (observer) data. We tested three different psychophysical methods: gray scales, constant stimuli, and two-alternative forced choice (2AFC). Our results show that when using gray scales or constant stimuli, assessments of color differences are biased toward lightness differences. This bias is particularly strong in LCD monitor experiments, and also present when using physical paint samples. No such bias is found when using 2AFC. In that case, however, observer responses are affected by other factors that are not accounted for by current color difference formulas. For accurate prediction of relative color differences, our results show, in agreement with other works, that modern color difference formulas do not perform well. We also investigated if the use of digital images as presented on LCD displays is a good alternative to using physical samples. Our results indicate that there are systematic differences between these two media.

OCIS codes: (330.5510) Psychophysics; (330.1730) Colorimetry; (330.1690) Color.
http://dx.doi.org/10.1364/JOSAA.32.000357

1. INTRODUCTION

During the past decades, the measurement of color differences has become a standard procedure. Reflection measurements obtained with a spectrophotometer are converted into numerical color differences by using a color difference formula. Many different color difference formulas have been proposed. A number of quantitative studies showed that the most recent formulas yield more or less equal performance [1–7].

Unfortunately, the color differences calculated with these formulas are often not sufficiently accurate for critical applications, such as quality control in manufacturing. The squared correlation coefficient R^2 between visual data from psychophysical experiments and color differences calculated from reflection data is typically only 70% or less [8–10]. Therefore, this and other current research aims at obtaining maximum correlation between instrumental predictions and visual assessments of color differences, given practical limitations such as measurement uncertainty and observer-variability.

There are several possible reasons for this "unsatisfactory state of affairs" [8]. The psychophysical tests underlying the experimental datasets on which the current color difference formulas are based were carried out under a variety of observation conditions. Furthermore, many of these tests included different types of samples, ranging from textiles to high gloss paints. Therefore, there is a clear need for a new concerted effort to thoroughly and consistently investigate perceived color differences [8]. In this article, we investigate some important methodological aspects that need to be decided on before starting a new investigation.

Most of the data used for developing the modern color difference formulas were obtained by using the psychophysical method of constant stimuli or the gray-scale method. These methods employ achromatic reference sample pairs mainly showing differences in the lightness dimension (lighter/darker). However, a pair of test samples may display a difference in any combination of the three dimensions of color space, i.e., in lightness, chroma, and hue. Thus, with the constant stimuli and gray-scale methods, observers are forced to mentally convert differences in chroma and hue into equivalent lightness differences, which may increase interobserver variability, and introduce a bias toward lightness differences [11]. A method that would avoid this problem is two-alternative forced choice (2AFC). In this article, we will first briefly introduce these three psychophysical methods (Section 2). We will then present a series of psychophysical experiments in order to decide which of these three methods is to be preferred for developing color difference formulas. The experimental setup is described in Section 3, and the results in Section 4. We also investigate if, instead of conducting psychophysical tests with physical samples, digital images may be used as visual stimuli. The latter would bring immense advantages with respect to time and costs, and a number of preliminary studies indicate that this is a serious option [12,13]. We emphasize, however, that it is not our goal to directly compare results of experiments with virtual samples and real samples. Our main goal is to study differences among the three psychophysical methods. In Section 5, we discuss our main results and conclusions.

1084-7529/15/030357-10$15.00/0

2. THREE PSYCHOPHYSICAL METHODS

Here we discuss three different psychophysical methods that can be used for collecting visual observer data necessary for developing color difference formulas.

A. Constant Stimuli

In this method, one anchor pair of nearly achromatic samples with a certain color difference (mainly a lightness difference) is introduced [14,15]. Observers are asked to assess the color differences for a series of sample pairs. The observer visually compares the color difference of a sample pair with the color difference of the anchor pair, and is then forced to choose if it is smaller or larger than the difference displayed by the anchor pair.

To convert these perceptibility data into acceptability data, as required for industrial quality-control procedures, one usually assumes during analysis of the data that the visual scores "smaller" and "larger" can be interpreted as "pass" (color difference acceptably small) and "fail" (unacceptably large), respectively. This requires the color difference of the anchor pair to be chosen carefully. A logit or probit analysis will yield the tolerance ellipsoids for, e.g., 50% likelihood of passing and failing.

B. Gray Scale

Instead of using one anchor pair, as in the method of constant stimuli, in the gray-scale method a series of anchor pairs is used [16,17]. The color differences in the anchor pairs vary from imperceptibly small up to a value significantly larger than the largest color difference occurring in the set of sample pairs. The anchor pairs consist of achromatic samples that show mostly lightness differences, dominating small differences in chroma and hue that may arise in production of the samples. During the visual experiment, a pair of test samples is compared to all anchor pairs. The anchor pairs are usually numbered in order to obtain a numerical scale. The observer reports the number of anchor pair having a difference most similar to the sample pair. If the perceived difference in the sample pair lies in between the differences displayed by two adjacent anchor pairs, the observer reports an intermediate value. Afterward, the data are converted to acceptability data by choosing a numerical value for the pass–fail criterion.

C. Two-Alternative Forced Choice

A 2AFC experiment requires no anchor pairs. Observers are confronted with three samples, A, B, and C. They have to choose which color difference is smaller, the one between A and B or the one between B and C. This method has been applied (under various names) in only a few studies related to color differences [18–21], but as far as we know it was never used to develop color difference formulas. Therefore, it may be less clear how the resulting visual data can be used for deriving tolerance ellipsoids. This will be discussed further below.

D. A First Brief Comparison

The methods of constant stimuli and gray scales utilize one or more anchor pairs. Usually, the color difference in the anchor pair(s) consists mainly of a lightness difference between achromatic samples. Then, when using these methods, the observers could be biased to this particular dimension of the color difference they observe in sample pairs. Indeed, Kuehni already questioned how well observers are capable of mentally converting a color difference they see in a sample pair to a lightness difference they observe in the anchor pair (s), and if this mental conversion is the same for all observers [11]. Obviously, when using 2AFC, no such bias is introduced. Color differences in any direction are visually compared to color differences in any other direction.

3. EXPERIMENTAL

Since this is a study on methodology, we do not aim for representative sampling of color space. Our experiments are based on three different color centers. From the recommendation of the CIE on concerted efforts for developing color difference formulas [22], we chose a red color center, R1, at CIELAB coordinates $L* = 44$, $a* = 37$, $b* = 23$, and a blue color center, B1, at $L* = 36$, $a* = 5$, $b* = -31$. Because of the limited availability of paint samples, we also studied another blue color center, B2, at $L* = 22$, $a* = 10$, $b* = -42$. Note that the blue region is the most problematic for color difference formula performance [16].

All visual experiments were carried out with 10 observers who had been screened for color vision deficiencies. Based on the outcome of the Ishihara color vision test and the Farnsworth–Munsell 100 hue test, all observers were considered color normal. They used binocular and free viewing when viewing the stimuli. The order and relative positions of samples were randomized to avoid systematic errors.

We started our experiments by studying virtual samples on an LCD monitor, using color centers R1 and B1. Thereafter, we switched to using physical samples. Because of limited availability of physical samples, we were forced to switch to color center B2.

A. Virtual Samples Experiments

Experiments with an electronic display used an EIZO CG221 LCD monitor. This display was selected because of its spatial color uniformity and temporal color stability, and its internal-hardware-based color calibration. With a Gretag-Macbeth i1 spectrophotometer and EyeOne software, the monitor was color calibrated. Color accuracy of the generated images was improved by using low-contrast dithering, increasing the standard color resolution from 24 to 33 bits/pixel [23]. Observers watched the display in an otherwise dark room, as shown in Fig. 1(a). The viewing distance in this case was 70 cm, resulting in each color image having an angular size of 3.7°. The colored images were surrounded by a background representing $L* = 20$, $a* = b* = 0$, as shown in Fig. 2. On the display, sample pairs were separated by a dark line of one pixel width, having the same color as the background. Preliminary experiments showed that, by using this separation line, better correlation with experiments on physical samples is obtained, since in the latter case the visual stimuli are also perceptually separated (this was also concluded in [13]).

The color differences for 30 sample pairs were selected based on a statistical design. Color differences varied between $\Delta E^*_{ab} = 0.125$ and 2.5 as calculated with [24]

$$\Delta E^*_{ab} = \left[(\Delta L^*)^2 + (\Delta C^*)^2 + (\Delta H^*)^2\right]^{0.5},$$

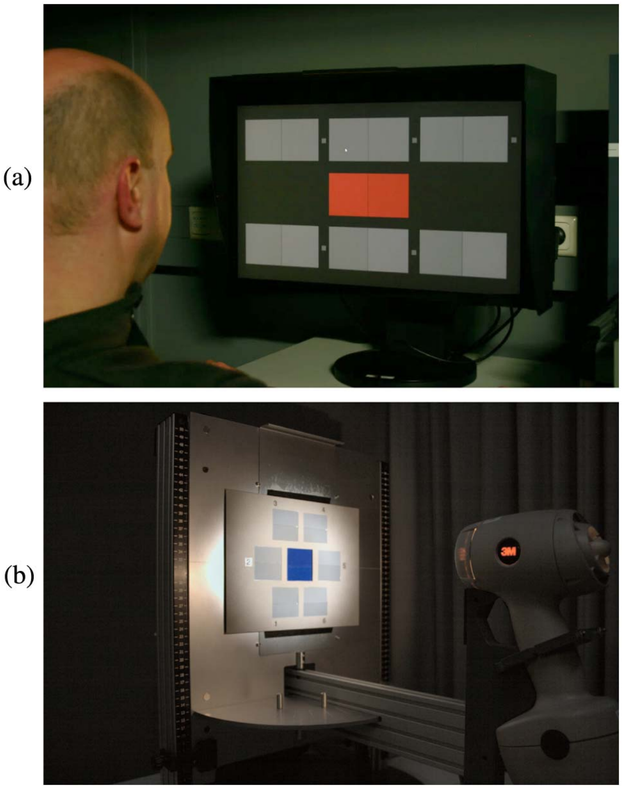

Fig. 1. Experimental setup in the visual test using (a) virtual samples on LCD and (b) physical samples.

which shows how ΔE^*_{ab} is composed of the difference components in lightness (ΔL^*), chroma (ΔC^*), and hue (ΔH^*). We created independent ΔL^*, ΔC^*, and ΔH^* differences as follows. Exactly 2400 random numbers were selected between -1 and $+1$, and these were assigned to columns for ΔL^*, ΔC^*, and ΔH^* for a left-hand pair in 2AFC, and ΔL^*, ΔC^*, and ΔH^* for a right-hand pair. In this way, 400 potential candidate sample sets were created for a 2AFC experiment. Here, each candidate sample set consists of the color center in the middle, to be compared with one sample on the left and one on the right. From the resulting 400 potential candidate sample sets, a IV-optimal design based on the resulting color differences ΔE^*_{ab} led us to select the 15 sample sets that were used for the 2AFC tests (a IV-optimal design minimizes the integrated prediction variance across ΔL^*, ΔC^*, ΔH^* space). For the gray-scale tests, we used the same selected samples, which result in 30 pairs. The average value of ΔE^*_{ab} is 1.27 for this selected set of sample pairs.

B. Physical Samples Experiments

For the set of physical samples, we utilized an existing set of 40 samples possessing small color differences. We found that, within this set, 309 pairs with a color difference $\Delta E^*_{ab} < 2.5$ could be selected. We extracted 30 pairs with a good distribution of color differences ΔE^*_{ab} and components ΔL^*, ΔC^*, and ΔH^*, with color differences similar to the design of the LCD monitor experiment. For the 2AFC method in which three samples are involved per trial, it would be most convenient to keep the same central paint sample throughout an experimental session, and change only the two adjacent test samples from trial to trial. However, we could not find a sample in our set that would then lead to the desired pairs of color differences.

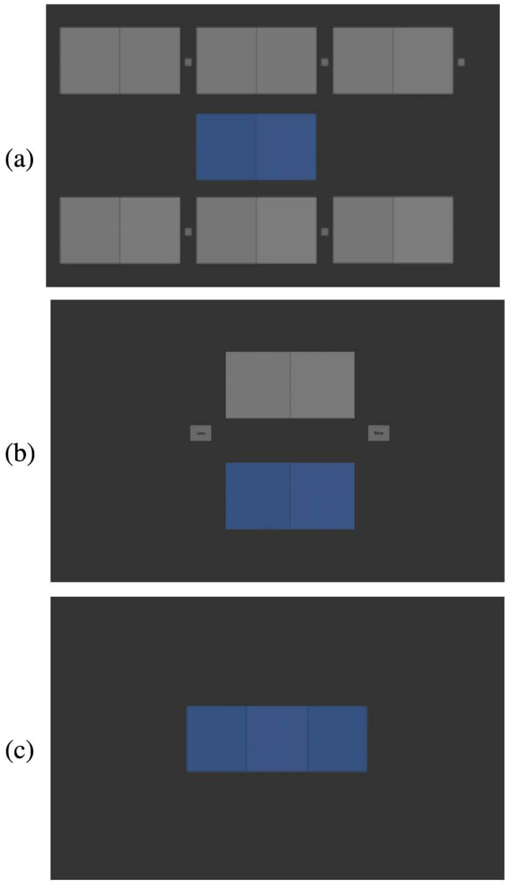

Fig. 2. Screenshots for experiments with LCD-generated colors, using (a) linear gray scale, (b) method of constant stimuli, and (c) 2AFC.

Physical samples were all made from high-gloss automotive paint on a steel substrate. These samples were placed on a vertical sample holder, as shown in Fig. 1(b). The tests were conducted in a dark room with a spotlight providing highly intense directional lighting, at 12.000 lux as measured on the samples. The setup was made in such a way that the physical samples and the anchor pair(s) were uniformly illuminated. More details of this setup are presented in [25].

Special mounts were prepared that included the anchor pair(s) for each psychophysical method, as shown in Fig. 3. We developed mounts to make sure that observers assessed the same area of all samples, and also to control the mid-gray color immediately surrounding the samples ($L^* = 56.5$, $a^* = -1.1$, $b^* = 5.5$). Anchor pairs were also integrated into the mount. Since each of the three psychophysical methods that we tested has a different number of anchor pairs, each method required its own mount. Since we wanted all anchor pairs to be illuminated by the same light spot and with the same illuminance, all pairs had to be positioned quite close to each other. To make everything fit, the angular size of the samples is 7° at the 40 cm observer distance we used.

C. Constant Stimuli Experiments

For the anchor pair in constant stimuli experiments, we used a color difference of $\Delta E^*_{ab} = 1.0$, comparable to the value of $\Delta E^*_{ab} = 1.02$ that was used in the RIT-Dupont dataset [26]

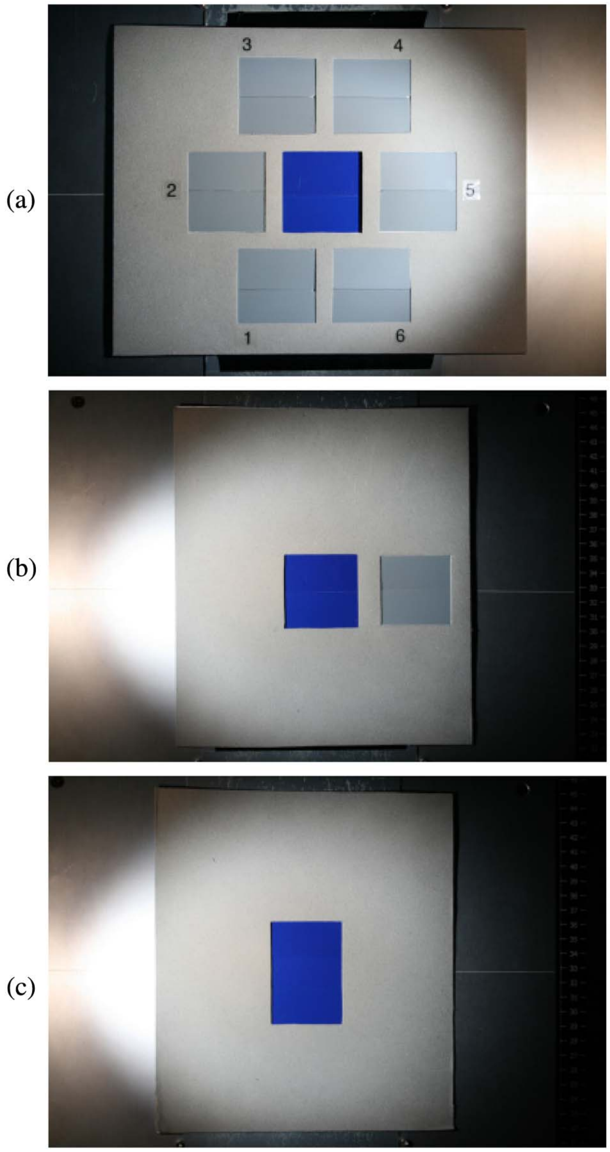

Fig. 3. Three different mounts for visual test using (a) linear gray scale, (b) method of constant stimuli, and (c) 2AFC. Illumination here is from the side, to allow photographing. During experiments, illumination was perpendicular.

underlying the recent color difference formulas ΔE^*_{94} [27] and ΔE_{00} [28]. In our case the color difference in the anchor pair is only in the lightness direction, while in the research for the RIT-Dupont dataset a chromatic component was present.

D. Gray-Scale Experiments

The six anchor pairs in gray-scale experiments were selected according to a linear gray scale. Recently we showed that, for investigating small suprathreshold color differences, mathematical limitations make a geometric gray scale less suitable [29]. We chose the anchor pairs to start with $\Delta E^*_{ab} = 0.0$ and to end with $\Delta E^*_{ab} = 3.0$, thereby covering the full range of color differences displayed by our sample pairs (maximum $\Delta E^*_{ab} = 2.5$). Thus, color differences in the anchor pairs are $\Delta E^*_{ab} = 0.0$ (pair 1), $\Delta E^*_{ab} = 0.6$ (pair 2), $\Delta E^*_{ab} = 1.2$ (pair 3), …, up to $\Delta E^*_{ab} = 3.0$ (pair 6) in the case of the LCD monitor experiments. For the experiments with physical samples, the color differences in the numbered anchor pairs were $\Delta E^*_{ab} = 0.14$, 0.85, 1.42, 2.01, 2.60, and 3.18. For these samples, the ratio of the lightness difference to the total color difference is 91% for the first anchor pair, and larger than 99% for the others. Observers were allowed to give halfway intermediate visual scores as well, so the visual scores are 1, 1.5, 2, 2.5, up to 6.

4. RESULTS AND DISCUSSION

A. Virtual Samples Experiment: Results for Red Color Center R1

For red color center R1 and using the LCD display, we did tests with the gray-scale method and with 2AFC. The 2AFC tests involved 15 sample sets (each set containing three samples, i.e., two color differences), whereas the gray-scale method involved 30 sample pairs (i.e., 30 color differences).

For the gray-scale test, observers gave an average visual score that varied from 1.4 to 5.3 for the different sample pairs. This demonstrates that the range of color differences in the gray scale agrees well with the range for the sample pairs, while for a geometric gray scale this would not have been possible [29].

Next we determined the reproducibility between observers. The average absolute difference to the average score per sample pair is 0.59 units. Given the distribution of visual scores for each sample pair, there are 19 visual scores that differ by more than 1 unit from the mode of the distribution. These represent 6.3% of all visual scores that were given. Furthermore, the average visual score per observer ranges from 2.8 to 3.9. This shows that there is a relatively large inter-observer variation. However, a more detailed reproducibility analysis shows that none of the observers significantly deviates from the group result.

For the 2AFC test, in five out of the 15 sample sets, all 10 observers agreed which of the two color differences they perceived as being the smallest. In five other cases the observers disagreed with a 50%–50% or 60%–40% ratio. For one sample set, the ratio was 70%–30%. Assuming a binomial distribution with 10 observers, and under the null hypothesis that there is no difference present, the chance that this happens is 12%. Therefore, this case is generally considered to be no deviation. In four cases, either one or two observers deviated from the other observers. The chance of this happening when the actual probability would be 0.5 is less than 5%, and, therefore, these cases are considered to be inconsistent choices. This happened in seven from the total of 150 choices that were made, i.e., in 5% of all choices that were made. Therefore, the percentage of deviating scores is approximately the same for the gray-scale and the 2AFC tests.

Figure 4 shows the results of the gray-scale experiment, with the average visual score of each sample pair on the vertical axis, and the corresponding measured color difference on the horizontal axis. The results for color differences evaluated with the CMC formula [30], popular in the textile and paint industries, show that ΔE_{CMC} ($l = 1.5$, $c = 1$) is a poor descriptor ($R^2 = 44\%$) for the perceived color differences assessed in the gray-scale experiment [Fig. 4(a)]. With other formulas, like the original CIELAB [24] and state-of-the-art CIEDE2000 [28], we find similar poor results. In particular, we see that there are six data points (indicated as open diamonds in Fig. 4) lying far from the values predicted by

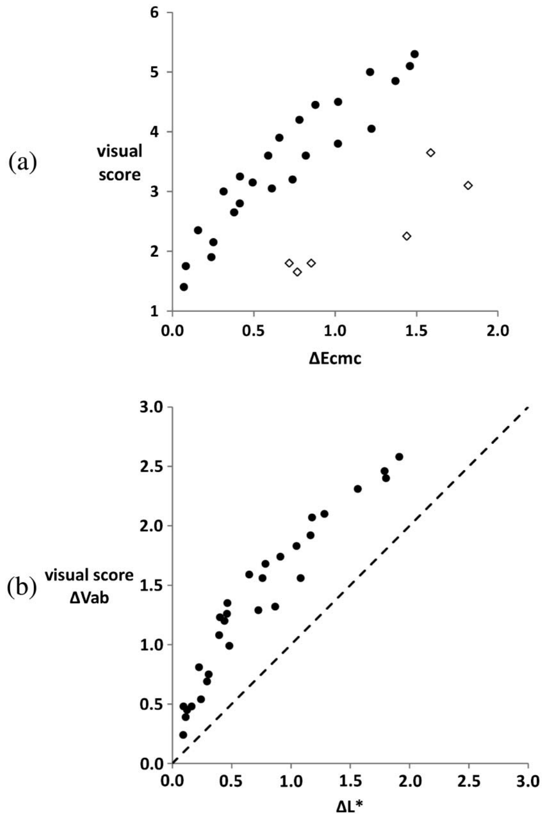

Fig. 4. Virtual samples experiment: results from the gray-scale method on red color center R1. Visual scores are plotted against the color differences within sample pairs, according to (a) ΔE_{CMC} and (b) lightness difference ΔL^*. Open diamonds refer to sample pairs with small contribution of lightness differences to total color difference (see main text).

the calculated color differences. Upon closer inspection, we found that these six data points all represented cases in which the lightness difference ΔL^* constitutes less than 10% of the total color difference ΔE^*_{ab}. Also, the opposite turned out to be true: all cases in which ΔL^* was smaller than 10% of ΔE^*_{ab} proved to be outliers in Fig. 4(a).

Following the usual procedure, we converted the visual assessments of the gray-scale experiment into visual scores ΔV_{ab}, representing the equivalent ΔE^*_{ab} values [16,31]. These visual scores are based on interpolation of the ΔE^*_{ab} values for the anchor pairs. If we now plot the resulting visual score ΔV_{ab} values against the lightness difference ΔL^* within each sample pair, we obtain Fig. 4(b).

Two important observations can be made from Fig. 4(b). The first is that there are no more outliers. Therefore, we conclude that the lightness difference ΔL^* is the main explanatory variable for the gray-scale results obtained in this experiment. Second, Fig. 4(b) shows that all data points lie above the line on which visual scores ΔV_{ab} would be equal to ΔL^* as measured for the sample pairs. Although the lightness difference ΔL^* between sample pairs is the main explanatory variable for the visual scores in the gray-scale experiment, observers apparently assign a larger visual score ΔV_{ab} to account for color differences perceived on other color directions than ΔL^*. We tried to correlate this increase in visual score with the actual color difference components ΔC^* and ΔH^* in the sample pairs, but the correlation is poor. Figure 4(b) shows that observers tend to add a constant value of about $\Delta V_{ab} = 0.6$, i.e., one unit on the gray scales, to account for the other color dimensions. This explains why common color difference formulas do not correlate well with the visual data obtained with this gray-scale experiment on an LCD monitor.

For the 2AFC experiment, visual data is in the form of binary decisions. Therefore, we used logistic regression to analyze the data. On the vertical axis in Fig. 5(a) we have plotted the probability that the “left” sample is assessed to have a smaller color difference with respect to the center sample than the “right” pair. The horizontal axis shows this same probability, using the best logistic model that is based on the difference in color differences ΔE_{CMC} between the left-hand sample pair and the right-hand sample pair. Figure 5(a) shows that the color difference ΔE_{CMC} is a reasonably good predictor for the color differences assessed with 2AFC. Other color differences, like ΔE^*_{ab} and ΔE^*_{ab}, were found to be reasonably good predictors for the 2AFC results, as well. Predictions based on only lightness differences ΔL^* did not result in accurate predictions.

These results indicate that, in the gray-scale experiment on the LCD monitor, the visual assessments are driven mainly by

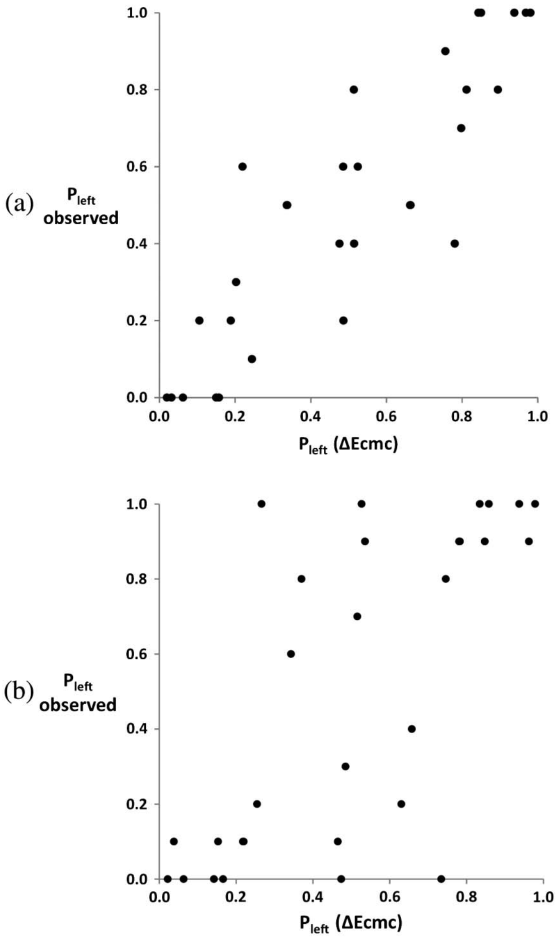

Fig. 5. Virtual samples experiment: results from the 2AFC method for (a) red color center R1 and (b) blue color center B1. The probability that the left-hand pair was selected as having the smallest color difference is plotted against predictions from the best logistic model as based on differences in color difference ΔE_{CMC} between the left-hand and right-hand sample pairs.

lightness differences within the sample pair. Because of the lightness differences in the anchor pairs, the gray-scale method apparently introduces a strong bias toward lightness differences. In the 2AFC experiment, no anchor pairs are used, and indeed no such bias was found.

Interestingly, results from monitor experiments with the gray-scale method published by Montag and Wilber show results similar to ours [32]. According to Fig. 5 and Table 4 in their article, tolerances for lightness were found to be much narrower than for chroma and hue, and their Figs. 4–6 show that, based on their results, it is hardly possible to determine a tolerance for chroma.

In a recent work by Sprow *et al.* we found further confirmation of our results [13]. In Fig. 5 of their article, Sprow *et al.* show that the visual results from their LCD monitor experiment with gray scales correlate less well with calculated color differences if the color differences within sample pairs are not mainly determined by a lightness difference. Our results are also consistent with Kuehni's conclusion that the mental conversion from observed chromatic differences in sample pairs into equivalent lightness differences as observed in anchor pairs varies widely among observers [11]. These individual variations may cause lightness differences to become the main explanatory variable in the gray-scale experiments.

B. Virtual Samples Experiment: Results for Blue Color Center B1

Again using the LCD display, we conducted tests with blue color center B1. This time we used all three psychophysical methods: the gray-scale method, the method of constant stimuli, and 2AFC. The viewing distance was reduced from 70 to 50 cm, resulting in an angular width of 5.1° for the samples. Exactly the same statistical design for the color differences in the sample pairs was used as in the previous case. Eight from the 10 observers of the previous experiment also participated in this test. Two new observers were introduced in the test. Unfortunately, one of them turned out to give visual scores that deviated significantly from the rest. Since this new observer was relatively inexperienced in visual tests, it was decided to exclude the corresponding data from the analysis. Also one score that was given for one sample pair was found to be a statistical outlier, and this one was removed from the analysis as well.

For the gray-scale experiment, the average score given by the observers was found to vary from 2.4 to 3.4. The width of this range is comparable to what we found for the red color center. The average score for each of the 30 sample pairs ranges from 1.2 to 5.3, similar to the results for the red color center.

Regarding the distribution of visual scores for each sample pair, we now find 11 visual scores, i.e., 4.1% of the total, that differ by more than 1 unit from the mode of the distribution. The average absolute difference in visual score to the average value for each sample pair is 0.47. Also in this respect, the results are comparable to the results for the red color center. The number of inconsistent assessments, as defined in the previous discussion for the red color center R1, was 5% for 2AFC and 4% for constant stimuli. Both numbers are comparable to what we found for the red color center.

The results that we find for blue color center B1 are similar to those for red color center R1. For example, for both color centers, the calculated values for ΔE^*_{ab} and ΔE_{CMC} are poor descriptors for the color differences assessed in the gray-scale experiment, and lightness difference ΔL^* is the main explanatory variable for the gray-scale results. Also for the blue color center the data points lying far from the values predicted by those calculated color differences are those with lightness differences ΔL^* constituting less than 10% of the total color difference ΔE^*_{ab}. Figure 6 shows that the data points for the blue color center are surprisingly similar to the corresponding data points for the red color center [Fig. 4(a)]. The other graphs are also very similar for these color centers.

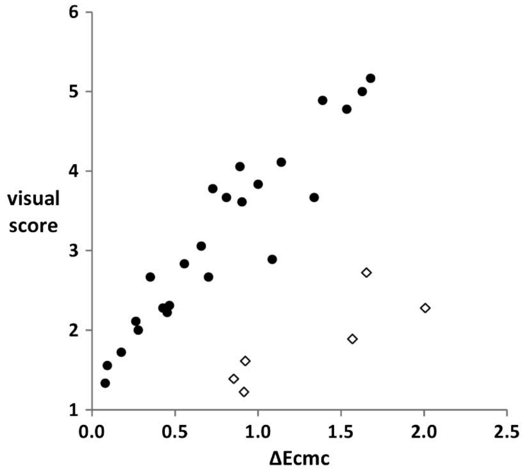

Fig. 6. Virtual samples experiment: results for blue color center B1.

Analyzing the 2AFC experiment with logistic regression, Fig. 5(b) shows that for blue color center B1, the color difference ΔE_{CMC} is not as good a predictor for the color differences assessed with 2AFC as it was for the red color center [Fig. 5(a)]. Similar results are obtained with the color difference ΔE^*_{ab} or ΔE_{00}. When using only lightness differences, ΔL^*, we again find very poor models, as was also the case for the red color center.

The results from 2AFC are actually best modeled using factors that are the square of ΔL^*, ΔC^*, and ΔH^*. This is very different from what we found with the data from the gray-scale method.

Using logit analysis, visual data obtained with the method of constant stimuli is correlated with the best model for predicting if a color difference is scored to be smaller than the anchor pair. Figure 7 shows the results when this probability distribution P is based on color differences ΔE_{CMC} (data for other color difference formulas like ΔE^*_{ab} and ΔE_{00} are very similar), and also with lightness difference ΔL^*. Data points for samples having a contribution of ΔL^* to ΔE^*_{ab} less than 10% are indicated as open diamonds.

Figure 7 clearly shows that the parameter ΔL^* correlates best with the probability function P. It also shows that, in particular, sample pairs having a small contribution of ΔL^* to ΔE^*_{ab} are not well described when probability models are built based on color differences like ΔE_{CMC}, but such points are well described by models based on lightness differences ΔL^*. These results are very similar to what we found when analyzing the gray-scale results. This supports the conclusion we reached in the previous section: that the lightness difference in the anchor pairs is mainly responsible for the bias toward lightness differences in the case of gray-scale methods.

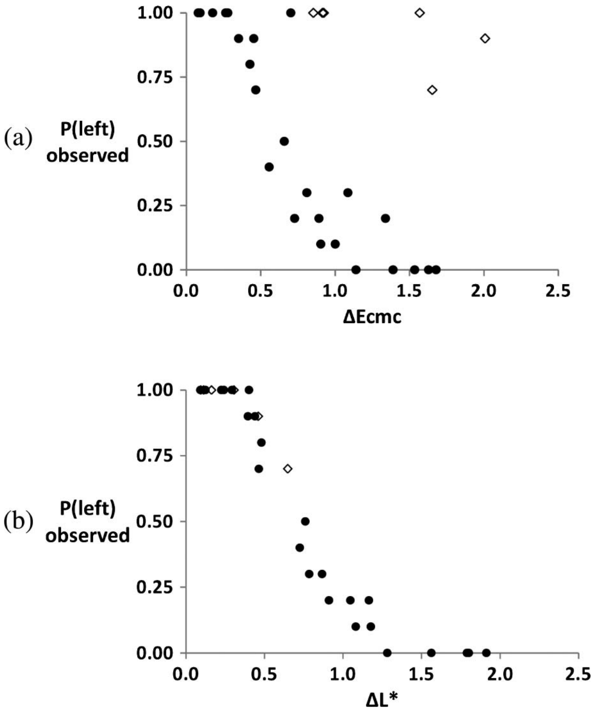

Fig. 7. Virtual samples experiment: results from the constant stimuli method on blue color center B1. Experimentally observed data are plotted against (a) color difference ΔE_{CMC} and (b) lightness difference ΔL^*.

Since the method of constant stimuli also utilizes an anchor pair having a lightness difference, it was indeed expected that the same phenomenon would occur for that method, as well.

C. Physical Samples Experiment: Results for Blue Color Center B2

For blue color center B2, we repeated the tests using all three psychophysical methods, but now for physical samples. For the gray-scale method, the average difference between individual visual scores of sample pairs and their average visual score was 0.61. This is slightly larger but comparable to what we found for gray-scale results obtained with the LCD monitor for blue color center B1 (0.47) and for the red color center (0.59).

The number of inconsistent results that we find with the method of constant stimuli was 5% (15 out of 300 assessments), and 8% for 2AFC (13 from the 170 choices). The latter percentage is slightly larger than what we found for the other color centers.

Results from the gray-scale experiment are shown in Fig. 8. The visual scores ΔV_{ab} correlate better with color differences like ΔE_{CMC}, ΔE^*_{ab}, and ΔE_{00} than with lightness differences ΔL^*. Data points corresponding to pairs with a small contribution of ΔL^* to the total color difference are again recognized as deviating from the main trend, but the deviations are smaller than in the LCD monitor case. The same graphs also show that apart from lightness also the other color dimensions are needed to improve correlation. The highest correlation coefficient is found for the ΔE_{00} formula, giving $R^2 = 85.2\%$ (when accounting only for lightness differences we find $R^2 = 71.7\%$).

As a next step in the analysis, we used the ΔE_{00} expression and optimized its coefficients for this dataset. In this way, we find that visual scores ΔV_{ab} are best predicted by the following expression:

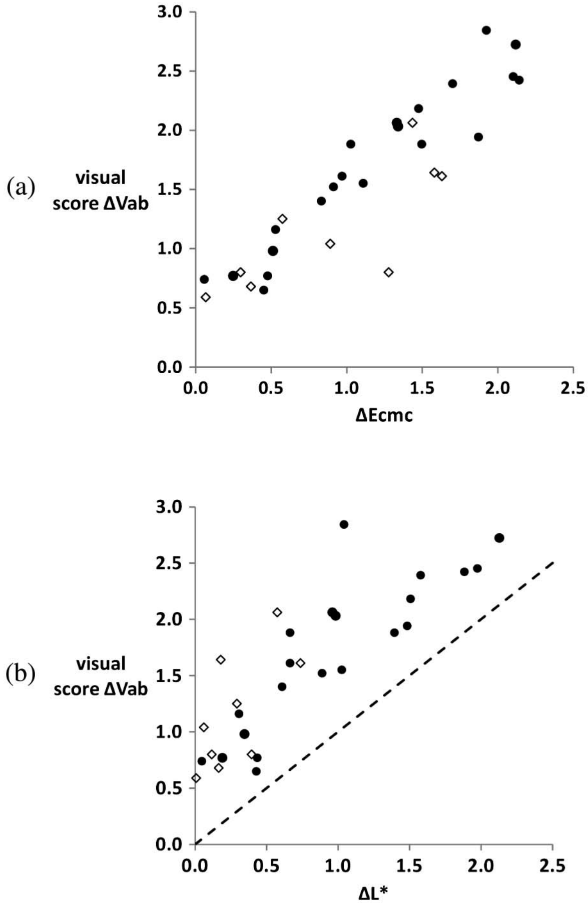

Fig. 8. Physical samples experiment: results from the gray-scale method on blue color center B2. Similar to Fig. 4.

$$\Delta V^2 = 0.458dL^2 + 0.142dC^2 + 0.400dH^2 - 0.423dCdH.$$

This should be compared to the nonoptimized coefficient values for the corresponding terms if the CIEDE2000 formula is used for this color center:

$$\Delta V^2 = 0.407dL^2 + 0.089dC^2 + 0.503dH^2 - 0.338dCdH.$$

The coefficients for ΔL^2 and $\Delta C\Delta H$ that we find are therefore slightly larger than the corresponding values from the CIEDE2000 formula. We found that these larger values improve the fit, especially for the sample pairs with a relatively small contribution of ΔL^* to the overall color difference (the open diamonds in Fig. 8). With the optimized coefficient values we find that correlation improved to $R^2 = 88.6\%$.

The data presented here are best fit with an ellipsoid with a rotation angle of 133°. This rotation angle agrees well with published values for nearby color centers: 131.51° [33] and 129.0° [34] for nearby color centers with similar high chroma values; and 115.4° [34], 122° [33], and 127° [35] for slightly lower chroma centers. With the CIEDE2000 formula a rotation angle of 123° is calculated for this color center.

Also for the method of constant stimuli, we find that our results fit well with what would be predicted with the

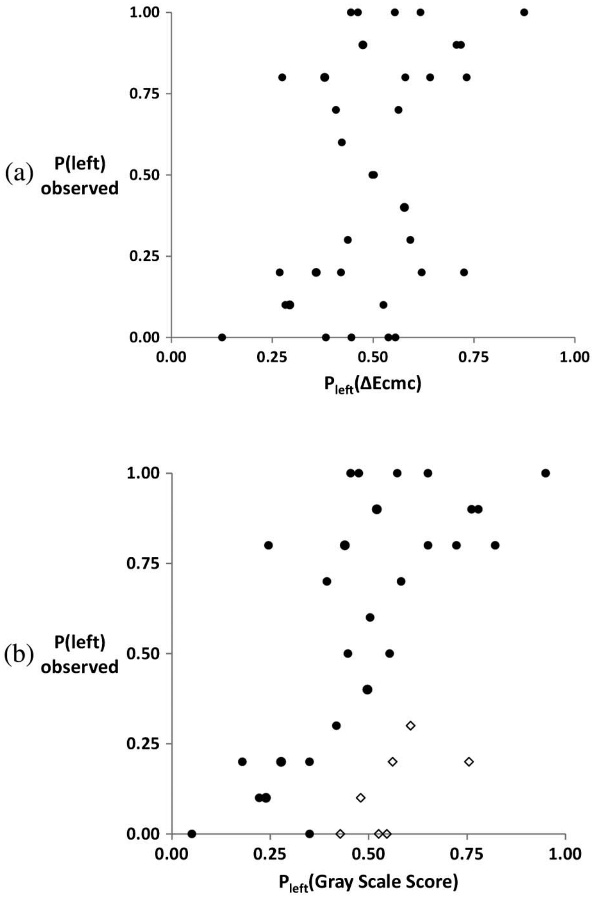

Fig. 9. Physical samples experiment: results from 2AFC method on blue color center B2. The vertical axis shows observed chance of selecting the left sample. The horizontal axis shows (a) the best statistical model based on color difference ΔE_{CMC} and (b) the best model based on visual scores from the gray-scale experiment.

CIEDE2000 formula. This confirms that when using physical samples our results agree with the experimental data on which the modern color difference formulas are based.

Figure 9(a) shows that the choices made by observers during a 2AFC experiment cannot be predicted well by models based on the CMC color difference formula. Similar poor models result for other color difference formulas such as CIELAB and CIEDE2000. Even if we optimize the values of coefficients in the CIEDE2000 formula we do not find a good fit. To explain these poor correlations, in Fig. 9(b) we show the correlation between the probability (on the horizontal axis) of the left sample pair, being selected as having the smallest color difference based on its gray-scale score, versus the probability (on the vertical axis) that the sample pair was actually selected as having the smallest color difference in the 2AFC experiment. The correlation between these two experiments is seen to be small. For example, let us consider the data points represented by open diamonds. These data points correspond to pairs of samples with almost equal gray-scale scores. Therefore, in these cases it is expected that in the 2AFC experiment the probability that a particular pair is perceived as having the smallest color difference is approximately 50%, but the actual 2AFC results show that in that experiment observers tend to agree which of the two sample pairs has the smallest color difference.

To verify these results, we selected eight further combinations of sample pairs for which the scores in the gray-scale experiment were very similar (less than 0.65 units difference on the six-point gray scale defined above). As expected, the calculated color differences were very similar for each combination of sample pairs, with a largest difference of 0.47 in CIEDE2000 units. In a final 2AFC experiment, 10 experienced observers assessed these eight combinations of sample pairs. For two of the combinations, we found that a large majority observers, in a 10:0 and a 9:1 ratio, agreed on which sample pair had the smallest color difference. Based on this result, a statistical analysis showed that we cannot reject the hypothesis that gray-scale scores and 2AFC scores deviate from each other.

Our results indicate that modern color difference formulas such as CIEDE2000 are well able to predict the absolute color difference in a sample pair. Indeed, their predictions correlate well with the data coming from tests that use the gray-scale or constant stimuli method. However, our data also show that the data from tests with the 2AFC method do not correlate well with predictions from the CIEDE2000 formula or with the results from tests with the gray-scale or constant stimuli method.

Several explanations for these results are possible. (i) We may need to distinguish between a quantitative prediction for the magnitude of the perceived color difference within a single sample pair (absolute color difference), versus an accurate prediction of which color difference is perceived to be smallest from two sample pairs (relative color difference). Since observer variability in assessing color differences has a relatively large nonsystematic character, this might explain why predicting relative color differences requires color difference formulas with a larger accuracy than what is needed for predicting absolute color differences. (ii) Our results show that in the gray-scale and the constant stimuli methods, for all sample pairs the observer is biased toward lightness differences. In the 2AFC method, the color difference within a sample pair is compared to another pair, which can be considered a reference pair that changes for each sample pair. Therefore, if there is also a bias in the case of 2AFC experiments, its direction in color space would be different for each sample pair. This may not only increase observer variability, but it may also be impossible to quantify using the mathematical expressions used in modern color difference formulas. (iii) In 2AFC the observer is forced to choose which of the two sample pairs has the smallest color difference, even if no difference in color differences is observed. In such cases, observers may therefore utilize a "plan B" approach for assessing the color differences. For example, the observers may unconsciously fall back to assess color difference preferences rather than color difference perception, or they may decide to give more weight to, e.g., hue differences or lightness differences. Such fall-back plans are not used in tests based on the gray-scale or constant stimuli method, and, therefore, they are not accounted for in the modern color difference formulas that are based on those methods.

5. CONCLUSIONS

The tests that we present for physical samples show that current color difference formulas predict reasonably well the data obtained when using the gray-scale or the constant stimuli method. This is understandable, since to a large part current color difference formulas are based on the method of constant stimuli.

From tests with color stimuli on the LCD monitor, we found that, unlike 2AFC, the gray-scale method and the method of constant stimuli lead to a dominant contribution from lightness differences ΔL^* to the observed color differences. When repeating these tests with physical samples, we found that contributions from ΔC^* and ΔH^* became significant too, but lightness differences still dominate visual assessments of color differences.

Therefore, we conclude that the psychophysical methods underlying current color difference formulas may have overestimated the role of lightness differences in estimating color differences. We plan to investigate this further for more color centers.

We conclude that in 2AFC tests, observer assessments are driven by different factors than in tests that use the gray-scale or constant stimuli method. Since modern color difference formulas are based on data obtained with the gray-scale or constant stimuli method, their predictions do not correlate well with 2AFC data as shown before. We have seen many cases in which observers agree which of two observed color differences is smaller, whereas modern color difference formulas like CIEDE2000 predict the color differences to be equal. The same color difference formulas that describe reasonably well the absolute color differences obtained with the gray-scale method can be poor in predicting relative color differences. Our results indicate that for accurate predictions of relative color differences, new color difference formulas need to be developed. There is a clear need for more accurate predictions of relative color differences.

In many practical applications, a user needs to select which one of a set of candidate colors best matches a reference color. For predicting the outcome of such visual tasks, relative color differences need to be predicted rather than absolute color differences. Based on the results presented here, the actual modern color difference formulas are expected to not always perform well in such cases.

For developing color difference formulas that improve predictions for absolute color differences, we found no clear preference for either the method of constant stimuli or gray-scale. The two methods gave very similar results, confirming an earlier comparison of these methods [14] but in contrast to another study that did show smaller variability for the method of constant stimuli [32]. We found that the percentage of inconsistent data was very similar for the two methods (as it was for 2AFC). We had expected that the gray-scale method would yield higher accuracy, and would require fewer samples and observers than the method of constant stimuli, but these expectations were not confirmed by our results. If we consider only the assessments on single sample pairs, then the gray-scale method does result in a better relative population estimate. Therefore, our results are in line with Kuehni's recent hypothesis that "a scale of achromatic differences has a greater normative effect on the quantitative assessment of the size of a chromatic difference than a single chromatic reference difference" [11].

There are more practical factors that do prefer using constant stimuli over gray scale [32]. In the gray-scale method, producing the anchor pair(s) is more complicated, securing uniform lighting over samples and anchor samples is harder to accomplish, and training naive observers is more time consuming. For these reasons, the use of the method of constant stimuli and physical samples is preferred.

Our results indicate that there are systematic differences between a visual test with color stimuli on an LCD monitor and a test using physical samples. Therefore, we do not recommend the use of LCD monitors for developing color difference formulas for physical samples. Several reasons why LCD monitors may lead to different assessments of color differences have been mentioned in the past. Because of their narrowband character, individual differences in color-matching functions may be emphasized in LCD monitor tests [12]. The luminance level of LCD displays is lower than the luminance of physical samples observed in the tests. In the experiments described in this paper, a reference white has a luminance of 3820 $\mathrm{cd/m^2}$ for the physical sample under the spotlight, but only 80 $\mathrm{cd/m^2}$ for the displayed image on the LCD monitor. Also, in the LCD monitor experiment the surround color was darker than the colors of the anchor pairs, whereas in the experiments with physical samples it was lighter. Although these differences in the experiments may have affected the results we obtained to some extent, they do not seem to explain why the dominance of lightness differences that we found for gray-scale and constant stimuli experiments on the LCD monitor was reduced when we used physical samples.

ACKNOWLEDGMENTS

This work was supported by the research project FIS2013-45952-P ("Ministerio de Ciencia e Innovación," Spain).

REFERENCES

1. M. Melgosa, R. Huertas, and R. S. Berns, "Performance of recent advanced color-difference formulas using the standardized residual sum of squares index," J. Opt. Soc. Am. A **25**, 1828–1834 (2008).
2. R. Shamey, L. M. Cárdenas, D. Hinks, and R. Woodard, "Comparison of naive and expert subjects in the assessment of small color differences," J. Opt. Soc. Am. A **27**, 1482–1489 (2010).
3. J. M. Gibert, J. M. Dagà, E. J. Gilabert, and J. Valldeperas, "Evaluation of colour difference formulae," Color Technol.. **121**, 147–152 (2005).
4. M. R. Luo, M. C. Lo, and W. G. Kuo, "The LLAB(l:c) colour model," Color Res. Appl. **21**, 412–429 (1996).
5. S. Oglesby, "The effectiveness of CIE94 compared with the CMC equation," J. Soc. Dyers Color. **111**, 380–381 (1995).
6. I. Lissner and P. Urban, "Upgrading color-difference formulas," J. Opt. Soc. Am. A **27**, 1620–1629 (2010).
7. R. Shamey, D. Hinks, M. Melgosa, R. Luo, G. Cui, R. Huertas, L. Cárdenas, and S. G. Lee, "Evaluation of performance of twelve color-difference formulae using two NCSU experimental datasets," in *Proceedings of the CGIV* (IS&T, 2010), pp. 423–428.
8. R. G. Kuehni, "Color difference formulas: an unsatisfactory state of affairs," Color Res. Appl. **33**, 324–326 (2008).
9. M. Melgosa, M. J. Rivas, E. Hita, and F. Viénot, "Are we able to distinguish color attributes?" Color Res. Appl. **25**, 356–367 (2000).
10. G. G. Attridge and M. R. Pointer, "Some aspects of the visual scaling of large colour differences–II," Color Res. Appl. **25**, 116–122 (2000).
11. R. G. Kuehni, "Variability in estimation of supratheshold small color differences," Color Res. Appl. **34**, 367–374 (2009).
12. P. Urban, M. Fedutina, and I. Lissner, "Analyzing small suprathreshold differences of LCD-generated colors," J. Opt. Soc. Am. A **28**, 1500–1512 (2011).

13. I. Sprow, T. Stamm, and P. Zolliker, "Evaluation of color differences: use of LCD monitor," in *Proceedings of the 18th Color Imaging Conference* (IS&T/SID, 2010), pp. 115–120.
14. S. S. Guan and M. R. Luo, "Investigation of parametric effects using small color difference pairs," Color Res. Appl. **24**, 331–343 (1999).
15. R. S. Berns, "Deriving instrumental tolerances from pass-fail and colorimetric data," Color Res. Appl. **21**, 459–472 (1996).
16. S. G. Lee, R. Shamey, D. Hinks, and W. Jasper, "Development of a comprehensive visual dataset based on a CIE blue color center: assessment of color difference formulae using various statistical methods," Color Res. Appl. **36**, 27–41 (2011).
17. J. H. Xin, C. C. Lam, and M. R. Luo, "Investigation of parametric effects using medium colour-difference pairs," Color Res. Appl. **26**, 376–383 (2001).
18. P. F. M. Stalmeier and C. M. M. de Weert, "Large color differences and selective attention," J. Opt. Soc. Am. A **8**, 237–247 (1991).
19. R. G. Kuehni, R. Shamey, M. Mathews, and B. Keene, "Perceptual prominence of Hering's chromatic primaries," J. Opt. Soc. Am. A **27**, 159–165 (2010).
20. D. L. MacAdam, "Uniform color scales," J. Opt. Soc. Am. **64**, 504–509 (1966).
21. M. P. Lucassen, T. Gevers, A. Gijsenij, and N. Dekker, "Effects of chromatic image statistics on illumination induced color differences," J. Opt. Soc. Am. A **30**, 1871–1884 (2013).
22. K. Witt, "CIE guidelines for coordinated future work on industrial colour-difference evaluation," Color Res. Appl. **20**, 399–403 (1995).
23. M. P. Lucassen, P. Bijl, and J. Roelofsen, "The perception of static colored noise: detection and masking described by CIE94," Color Res. Appl. **33**, 178–191 (2008).
24. Commission Internationale de l'Eclairage, "Colorimetry," 2nd ed., CIE Publication No. 15.2 (CIE, 1986).
25. E. J. J. Kirchner, L. Njo, and M. Lucassen, "Calculating verbal descriptions of color difference components," in *Proceedings of the 12th International Conference of the AIC* (AIC, 2013), pp. 529–532.
26. D. H. Alman, R. S. Berns, G. D. Snyder, and W. A. Larsen, "Performance testing of color-difference metrics using a color tolerance dataset," Color. Res. Appl. **14**, 139–151 (1989).
27. Commission Internationale de l'Eclairage, "Industrial colour-difference evaluation," CIE Publication No. 116 (CIE, 1995).
28. "Improvement to industrial colour-difference evaluation," CIE Technical Report, CIE Publication No. 142 (Central Bureau of the CIE, 2001).
29. N. Dekker, M. Lucassen, E. Kirchner, P. Urban, and R. Huertas, "Mathematical limitations when choosing psychophysical methods: geometric versus linear grey scales," Proc. SPIE **9018**, 90180G (2014).
30. F. J. J. Clarke, R. McDonald, and B. Rigg, "Modification to the JPC79 colour difference formula," J. Soc. Dyers Col. **100**, 128–132 (1984).
31. S. G. Kandi and M. A. Tehran, "Investigating the effect of texture on the performance of color difference formulae," Color Res. Appl. **35**, 94–100 (2010).
32. E. D. Montag and D. C. Wilber, "A comparison of constant stimuli and gray-scale methods of color difference scaling," Color Res. Appl. **28**, 36–44 (2003).
33. M. Huang, H. Liu, G. Cui, M. R. Luo, and M. Melgosa, "Evaluation of threshold color differences using printed samples," J. Opt. Soc. Am. A **29**, 883–891 (2012).
34. R. Shamey, R. Cao, T. Tomassino, S. S. Zaidy, K. Iqbal, J. Lin, and S. G. Lee, "Performance of select color-difference formulas in the blue region," J. Opt. Soc. Am. A **31**, 1328–1336 (2014).
35. M. Melgosa, E. Hita, A. J. Poza, D. H. Alman, and R. S. Berns, "Suprathreshold ellipsoids for surface colors," Color Res. Appl. **22**, 148–155 (1997).