Journal of the Optical Society of America A
OPTICS, IMAGE SCIENCE, AND VISION

How color difference formulas depend on reference pairs in the underlying constant stimuli experiment

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

[1]Color Research Department, AkzoNobel Performance Coatings, Sassenheim, The Netherlands
[2]Lucassen Colour Research, 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 9 June 2015; revised 2 September 2015; accepted 25 October 2015; posted 28 October 2015 (Doc. ID 242696); published 20 November 2015

For calculating color differences, the CIEDE2000 and CIE94 equations are widely used and recommended. These equations were derived more than a decade ago, based for a large part on the RIT-Dupont set of visual data. This data was collected from a series of psychophysical tests that use the method of constant stimuli. In this method, observers need to compare the color difference within a sample pair to that between a reference pair. In the current investigation, we show that the color difference equation significantly changes if reference pairs are chosen in the underlying visual experiments that differ from what was used when creating the RIT-Dupont dataset. The investigation is done using metallic paint samples representing two color centers, red and yellow-green. We show that the reproducibility differs for three different reference pairs, and that for modeling the visual data for the yellow-green color center, extra model terms are required as compared to the CIEDE2000 equation. Our results suggest that observers differ in their ability to mentally convert a color difference recognized in a sample pair into an equivalent color difference along the color difference direction represented by the reference pair. We also find that in these tests the tolerance to lightness differences is widened by a factor of 1.3 to 1.6, and that for the red color center the tolerance ellipsoid is rotated by 30° as compared to the CIEDE2000 equation. The latter observations are possibly due to the metallic texture in the samples used for the current experiment.

OCIS codes: (330.5510) Psychophysics; (330.1730) Colorimetry; (330.1690) Color.

http://dx.doi.org/10.1364/JOSAA.32.002373

1. INTRODUCTION

For more than a decade, the CIEDE2000 and (to a lesser extent) the CIE94 equations have been the color difference equations for uniform colors recommended by the CIE [1,2,3]. The CIE94 equation was derived from a set of visual data that is usually referred to as the RIT-Dupont dataset, and the same dataset was also part of the data used for developing the CIEDE2000 equation [4]. The RIT-Dupont dataset has several advantages over alternative datasets that have been used in the past to derive other color difference equations. For example, the samples used in the tests underlying the RIT-Dupont dataset form a consistent set of high-gloss paint samples [5]. The numbers of sample pairs, color centers, and observers are larger than for most other datasets. Finally, the color differences investigated are limited to $\Delta E^*_{ab} < 5$. This is important because it is well known that color differences are not scalable over ranges like $\Delta E^*_{ab} < 10$ or larger [6].

For the visual experiments that led to the RIT-Dupont dataset, the psychophysical model of constant stimuli was used [5]. In this method, the observer is asked to compare the perceived color difference within each sample pair to that within a (close to) achromatic reference pair that remains the same for each sample pair. In this way, the observer is forced to mentally convert the differences that are observed in lightness and/or chromaticity and/or hue within the sample pair into an equivalent color difference in the color difference direction that is represented by the reference pair. Since it is unlikely that every observer has the same capability of doing this mental conversion, this procedure may lead to relatively bad reproducibility of visual assessments with this method [7]. Also, when using this

method, one may expect a bias. For example, if the color difference within the reference pair is dominated by a lightness difference, one may expect that assessments for a sample pair with a color difference that is also dominated by a lightness difference are easier to make for observers.

These problems may be avoided by choosing a different psychophysical method, such as the two-alternative forced choice (2AFC) method. We recently showed that the 2AFC method is indeed to be preferred when investigating which of two color differences is perceived to be smallest [8]. However, for developing color difference equations that improve predictions for absolute color differences, the method of constant stimuli (or the related method of gray scales) was found to be still preferable.

This leaves still open the question about the influence of the chosen reference pair on the eventual color difference equation. For creating the RIT-Dupont data that underlies the popular ΔE_{94} and ΔE_{00} color difference equations, a reference pair has been used with a color difference dominated by a chromaticity difference and (to a smaller degree) a lightness difference [4]. In the present investigation, we study the perceived color difference within sample pairs using the method of constant stimuli. We use three different reference pairs, one of which corresponds to the reference pair that has been used when creating the RIT-Dupont dataset. The color difference equations that best fit the visual data for each reference pair are compared with each other to find out the effect of choosing a reference pair on the resulting color difference formula.

In Section 2, we describe how we identified sample pairs for this visual test for two color centers, which are referred to as the red and the yellow-green color centers. In the same section, the visual test setup is also described. Section 3 describes the results that were obtained for both color centers, in terms of reproducibility, repeatability, and optimized models. In the final section we summarize the main conclusion from this work.

2. EXPERIMENTAL

A. Samples for the Red Color Center

In the experimental method that led to the RIT-Dupont dataset, a two-step procedure was used [4,5]. First, samples were created around a color center along six to 14 different color difference directions. Along each direction, the 50% tolerance value was determined by probit analysis of the visual test data. In the second step in the analysis, the 50% tolerance values along the different directions were analyzed to determine the optimum parameter values in the color difference equation. A similar two-step approach was used in a more recent investigation [9].

From a methodological point of view, it is preferable to estimate model parameters directly on the visual data, rather than through the two-step process described above. Also, it is hardly possible to create physical samples that lie perfectly on a straight line in color space. Therefore, we chose a different approach: we aim at including sample pairs that represent all color difference directions. We do not require one of the samples to be the same for all sample pairs, as was done for the RIT-Dupont test. Instead, we require only that all the samples are sufficiently close to each other, with a maximum $\Delta E_{cmc} = 6$ ($l = 1.5$, $c = 1$). The sample set that we used indeed has a maximum $\Delta E_{cmc} = 5.9$, with 95% of the sample pairs having $\Delta E_{cmc} < 4.0$.

Based on the availability of paint samples in our laboratories, we selected a red color center that would allow the use of a statistical design for the samples. This color center has CIELAB coordinates $L^* = 20$, $a^* = 39$, and $b^* = 22$ at the measurement geometry with 45° aspecular angle, as calculated for the 10° standard CIE observer. This is the geometry that best corresponds to the lighting and observation angles in the visual test for this investigation. All samples consist of high-gloss automotive metallic paint on a steel substrate.

The aim of the present investigation is to study the methodology for setting up a visual experiment to develop an improved color difference formula specifically for metallic coatings. An investigation of the effect of texture differences on perceived color differences was not part of the present study (see [10,11,12]). For this reason, we chose a set of values for texture parameters that would allow us to select samples with almost no perceptible texture differences. By using definitions of texture parameters introduced previously [13], the texture values for this color center are quantified by a texture value for diffuse coarseness of 1.6, and values for glint impression of 4.8 (at a 25° aspecular angle), 4.6 (at 45°), and 3.1 (at 75°). These values characterize this color center as a relatively dark red color with a fine coarseness under diffuse illumination conditions, and with a substantial sparkle effect that remains visible until an aspecular angle of at least 45°.

In this investigation, we aim for describing small color differences. Therefore we require a set of sample pairs that all have small mutual color differences, with $\Delta E^*_{ab} < 2.5$, and with the color differences being uniformly distributed over this range. We aim for a set of sample pairs for which the color differences include contributions from all directions in CIELAB space. This makes it possible to estimate different quadratic interaction terms in a color difference formula: ΔL^{*2}, ΔC^{*2}, ΔH^{*2}, $\Delta C^*\Delta H^*$, $\Delta L^*\Delta C^*$, and $\Delta L^*\Delta H^*$.

For the selected red color center, 55 samples were within the specified range. This would make it possible to define 1485 ($55 \times 54/2$) sample pairs. From these, 191 sample pairs have a mutual color difference $\Delta E^*_{ab} < 2.5$. We require the perceived mutual texture difference to be very small, with a maximum difference of 0.6 units on a scale that runs from 0 to 8, which is close to the just noticeable difference in texture space [13]. In this way, 50 sample pairs were left.

For these 50 sample pairs, the values for ΔL^*, ΔC^*, and ΔH^* were used to define a subset of sample pairs suitable for developing a color difference formula with quadratic terms and interaction terms, like the known formulas. Based on a quadratic model with five additional samples to check the model fit quality, a subset of 15 sample pairs was extracted that ensures that the quadratic models can be estimated from the data with least prediction variation. Another 10 sample pairs were added, to make the distribution of color differences uniform. In an initial test, we made this sample selection uniform in ΔE^*_{ab}, but we found that in this way the sample design does not guarantee that the test contains a sufficient number of sample pairs with perceived color differences approximately equal to the difference in the reference pair. For this reason,

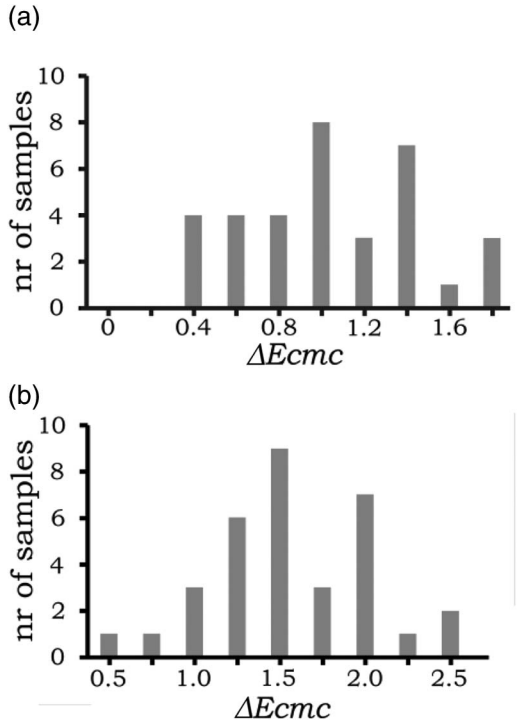

Fig. 1. Distribution of color differences ΔE_{cmc} in sample pairs used in the test for the (a) red and (b) yellow-green color centers.

we selected the sample pairs such that a uniform distribution in ΔE_{cmc} resulted, as illustrated in Fig. 1(a).

If we investigate the parameters ΔL^*, ΔC^*, and ΔH^* for the selected sample pairs, we find no correlation between ΔL^* and ΔH^*, or between ΔC^* and ΔH^* [see Fig. 2(a)]. We do find a weak correlation between ΔL^* and ΔC^*. Upon inspection, we found that this correlation exists for all 1485 sample pairs in our dataset, and, therefore, it could not be avoided. This correlation is probably a result of the set of colorants used for producing the set of samples. As a consequence, for the red color center, the present investigation is not suitable to test the inclusion of a $\Delta L^*\Delta C^*$ term in color difference formulas. Apart from this term, the five other quadratic terms were included in the present study for the red color center.

B. Samples for Yellow-Green Color Center

Based on our experience with producing samples for the red color center, we chose a different approach for the second color center. We chose a metallic paint sample for which we know the composition in terms of colorants and concentrations, and that has CIELAB coordinates $L^* = 43$, $a^* = -3.6$, and $b^* = 6.9$ at the 45° aspecular angle. The texture values for this sample are a diffuse coarseness of 3.7, and values for glint impression of 6.3 (at 25° aspecular angle), 7.1 (at 45°), and 6.6 (at 75°). This means that the sample has a yellow-green color of medium lightness level, with medium coarseness under diffuse lighting conditions and a high sparkle effect that is visible until at least 75° from the specular angle.

With color formulation software, we changed the concentrations of the colorants such that we cover various color

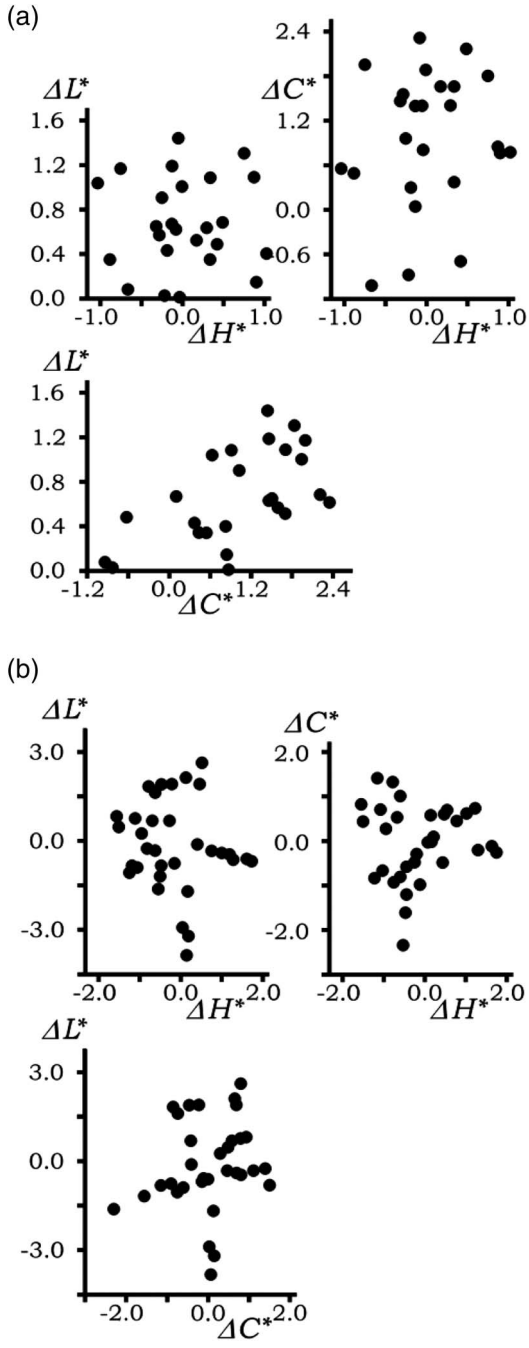

Fig. 2. Distribution of ΔL^*, ΔC^*, and ΔH^* in sample pairs used in the test for the (a) red and (b) yellow-green color centers.

difference directions within a distance of $\Delta E^*_{ab} = 3.0$ from the color center. In this way, we prepared 115 samples. From this set, 33 samples were selected such that we could combine each sample with the same reference sample, thus forming sample pairs that represent all color difference directions and for which we expect a good balance between pairs with a larger versus smaller perceived color difference with respect to the reference pairs. The resulting distributions are shown in Figs. 1(b) and 2(b).

C. Visual Test Setup

Ten experienced observers (eight male, two female, ages between 35 and 50) participated in the visual experiments. All observers have normal color vision as confirmed by the Ishihara color vision test and the Farnsworth–Munsell 100 hue test. All visual tests were executed with binocular viewing in a setup that has been extensively described in Ref. [8], and that will be briefly summarized here.

Each sample pair was placed on a vertical sample holder, as shown in Fig. 3. The samples were illuminated mainly by a spotlight, which provided 8170 lux of highly intense directional light as measured on the samples, with color temperature 4700 K. The setup was designed such that both the sample pair and the reference pair were uniformly illuminated by the spotlight. Samples were also illuminated by fluorescent tube lighting from the ceiling, with a color temperature of 6100 K. Since this additional light produced only 140 lux in the vertical plane at the position of the samples, its contribution to the illumination of the sample pair and reference pair is negligible.

The spotlight was positioned to make the light incident parallel to the surface normal of the sample pair, whereas the observer was at 45° from the surface normal. In this way, we imitate the 45° aspecular angle that is one of the measurement geometries of common multiangle spectrophotometers, and also because at this geometry the influence of small scratches in the samples is small. As shown in Figure 3, the line separating the two samples that constitute the reference pair is parallel (horizontal) to the plane formed by the spotlight, the observer, and the sample. This is important, because a different choice of orientation of the sample pair would lead to a slight difference between the illumination and observation angles of the samples on both sides of this separation line. For metallic samples such as investigated here, a slight difference in observation geometry can already lead to systematic color and texture differences between the samples.

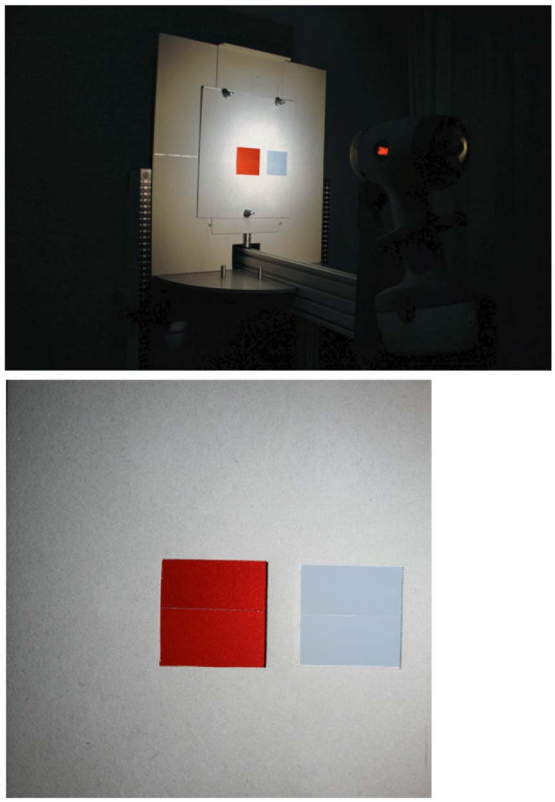

Fig. 3. (a) Setup of the visual test, showing a sample pair and a reference pair, partly covered by a mount, with the position of the spotlight and with the camera at the position of the observer. (b) Mount showing sample pair and reference pair.

For the yellow-green samples, the visual setup was slightly changed by introducing a visor. This visor ensures that the observation distance is fixed at 31 cm; for the red samples, observers were free to choose the observation distance, leading to values ranging between 35 and 45 cm.

To standardize the background and size of the sample pair, we used special mounts prepared for this test (see Ref. [8]). This ensured that, during observations, all samples subtended a visual field angle of 7°, with a well-defined, almost achromatic, color of the background and immediate surround ($L^* = 56.5$, $a^* = -1.09$, $b^* = 5.46$, subtending 23° around the samples). Every mount also contained the reference pair that is used for the constant stimuli method. From Figure 3 it can be deduced that the illumination and observation angles for the reference pair cannot exactly correspond to the 45° aspecular angle, but since the samples of the reference pairs are solid colors (i.e., their color does not change under slight variations in aspecular angle), this is not considered to be a problem.

For this investigation we used three different reference pairs, and separate mounts were prepared for each of them. All three reference pairs show a color difference close to $\Delta E^*_{ab} = 1.0$.

Reference pair A has color difference components similar (but not identical) to those of the reference pair that had been used for producing the RIT-Dupont dataset [4]. For this reference pair, the color difference originates mainly from a difference in chromaticity, and some difference in lightness as well. The corresponding numerical values are shown in Table 1. Reference pair B represents the case of a reference pair where the color difference is caused mainly by a difference in lightness, whereas the difference in hue is predominant in reference pair C.

During the visual test, each observer was asked to assess if the color difference within the sample pair was perceived to be smaller or larger than the color difference within the reference pair. The order of sample pairs was randomly generated for each observer. Separate test sessions were organized for each of the

Table 1. Specification of Color Differences and Color Coordinates for the Three Reference Pairs Used in This Investigation, and the Reference Pair Used for Creating the RIT-Dupont Dataset [5]

	A	B	C	RIT-Dupont
ΔE^*_{ab}	0.91	1.00	1.11	1.02
ΔL^*	0.37	0.99	0.05	0.64
ΔC^*	0.83	0.12	0.21	0.75
ΔH^*	0.07	0.04	1.09	0.26
L^*	49.26	49.60	48.95	49.53
a^*	−1.46	−1.52	−2.57	−0.08
b^*	−4.84	−3.59	−5.18	−5.65

Table 2. Reproducibility and Repeatability of Visual Assessments for the Red and Yellow-Green Color Centers, Using Methods Explained in the Main Text, for Reference Pairs A, B, and C[a]

	A	B	C
Red Color Center			
Reproducibility (average percentage of agreement)	69%	70%	64%
Repeatability (average percentage of agreement)	n.a.	72%	n.a.
Yellow-Green Color Center			
Reproducibility (average percentage of agreement)	67%	74%	70%
Repeatability (percentage of agreement)	71%	67%	75%

[a]Entries with "n.a." refer to cases that are not applicable, because no repeatability data is available.

three reference pairs, where the order of the reference pairs was also randomized for each observer. A period of one week separated test sessions for each observer, in order to avoid possible memory effects. With a bar-code scanner it was verified that observers used the correct reference pair and sample pairs during each session.

3. RESULTS AND DISCUSSION

A. Red Color Center

1. Reproducibility and Repeatability

We express the reproducibility of the visual assessments as follows. First a comparison is made between the individual visual scores that two observers give for each sample. We then calculate the percentage of scores that are equal, and average this over all samples and also over all combinations of observers. The percentage that results shows the average percentage of agreement. We calculate this to be 69% for reference pair A. For reference pairs B and C, the corresponding values are 70% and 64%, respectively (see Table 2).

The results shown in Table 2 indicate that the reproducibility for reference pairs A and B are very similar, whereas we find significantly worse reproducibility for reference pair C. A more detailed analysis of the results showed that, for eight of the 10 observers, reference pair C indeed resulted in the smallest average agreement percentage between the visual scores from one observer and from the other observers.

For reference pair B, an additional slightly modified experiment was conducted to investigate repeatability. This test involved 17 from the 25 sample pairs of the main test. Unfortunately, we could use only eight of the 10 observers of the main test, and also a slightly different spotlight was used that produced a slightly larger illuminance of 12,000 lux on the samples. In this way, we find that individual scores agree for 72% on average, which is very close to the value of 70% found for the reproducibility. This shows that the average agreement in visual scores is very similar if an observation is repeated by the same or by a different observer.

2. Optimized Model

We used the collected visual data to optimize with logistic regression a model containing all quadratic terms and two-parameter interaction terms in ΔL^*, ΔC^*, and ΔH^*. In this way we find that only the terms in ΔL^{*2}, ΔC^{*2}, ΔH^{*2}, and $\Delta C^*\Delta H^*$ obtain coefficient values that differ significantly from zero. For the red color center, none of the three reference pairs led to a model with a significant value for the $\Delta L^*\Delta H^*$ term. For the red color center, the term with $\Delta L^*\Delta C^*$ could not be investigated in the present study, as explained in Section 2.A. The optimized values for the various standardized coefficients are shown in Table 3 (top).

Table 3. Standardized Coefficients for Optimized Models When Using Reference Pairs A, B, and C, for the Red and Yellow-Green Color Centers[a]

	A	B	C	ΔE_{00}	ΔE_{cmc}
Red Color Center					
ΔL^{*2}	0.27	0.51	0.18	0.48	0.61
ΔC^{*2}	0.24	0.16	0.27	0.10	0.09
ΔH^{*2}	0.48	0.33	0.55	0.42	0.30
$\Delta C^*\Delta H^*$	−0.39	−0.35	−0.48	0.0	0.0
$\Delta L^*\Delta H^*$	N.S.	N.S.	N.S.	0.0	0.0
Rotation angle	29.0°	31.6°	30.2°	0°	0°
Somers' D optimized model	0.67	0.70	0.59	n.a.	n.a.
Somers' D CIEDE2000	0.56	0.61	0.45	n.a.	n.a.
Yellow-Green Color Center					
ΔL^{*2}	0.12	0.11	0.13	0.28	0.16
ΔC^{*2}	0.37	0.29	0.29	0.25	0.31
ΔH^{*2}	0.51	0.60	0.58	0.47	0.53
$\Delta C^*\Delta H^*$	0.26	0.18	0.21	0.18	0.00
$\Delta L^*\Delta H^*$	−0.15	−0.14	0.16	0.00	0.00
Rotation angle	149.2°	164.9°	162.0°	160.4°	0°
Somers' D optimized model	0.69	0.76	0.69	n.a.	n.a.
Somers' D CIEDE2000	0.54	0.58	0.60	n.a.	n.a.

[a]The corresponding values of the coefficients for the CIEDE2000 equation are also shown, as evaluated for the color center used in the experiment. The sixth row shows the rotation angle of the tolerance ellipsoid in the ΔC^*–ΔH^* plane. Entries with "n.a." refer to coefficients that were found to not significantly contribute to the model. The last two rows show Somers' D coefficient for the optimized model, as compared to the value when the CIEDE2000 equation is used to evaluate the visual data.

We analyzed if the coefficients for the models corresponding to the three reference pairs are statistically different. In a log-likelihood test we compared one overall model for the data from all three reference pairs with and without a dedicated factor that describes the reference pair used in the test. This statistical test showed that there is indeed a significant interaction between the reference pair factor and the coefficient for the ΔL^{*2} term at the $\alpha = 0.01$ level. This confirms that the optimized models for the three different reference pairs are statistically different.

In Table 3 (top), we also show the values of these coefficients as evaluated for the red color center when using the CIEDE2000 equation (with $k_L = k_C = k_H = 1$) [2,3]. We note that, to calculate these values, care should be taken to convert the primed coordinates $\Delta L'$, $\Delta C'$, $\Delta H'$ that are part of the definition of the CIEDE2000 equation into unprimed coordinates.

An important difference between the results obtained in this experiment with the RIT-Dupont data underlying the CIEDE2000 equation is that we find a coefficient value for the $\Delta C^* \Delta H^*$ term that differs significantly from zero. In the CIEDE2000 model, this coefficient is zero for the red color center. For all three investigated reference pairs, our results lead to a rotation of the tolerance ellipsoid of approximately 30° as compared to the tolerance ellipsoid orientation predicted by the CIEDE2000 equation. The models for all three reference pairs agree with each other in this respect. It is tempting to associate the rotation of the tolerance ellipsoid that we find here with the presence of metallic sparkle in the samples, but this hypothesis would need to be tested further.

The model coefficients shown in Table 3 have been standardized such that the summed coefficients for ΔL^{*2}, ΔC^{*2}, and ΔH^{*2} add up to 1. This step is needed because there is no way to compare an absolute threshold value obtained with one reference pair with an absolute threshold value obtained with another reference pair. For comparing the perceived color differences between two reference pairs, one would need an expression for the color difference for different color difference directions, but obtaining this expression is exactly the goal of the experiment. For this reason, it makes no sense to compare absolute model coefficients with each other. For the same reason, the coefficients for ΔE_{00} and ΔE_{cmc} quadratic expansions have also been standardized. In the analysis, we will compare only the relative contributions from each term with each other, i.e., the shape rather than the volume of the tolerance ellipsoid.

Table 3 (top) shows that the standardized values for the model coefficients differ from those produced by the CIEDE2000 equation (and also for the ΔE_{cmc} equation) for this color center. This confirms earlier results that showed that visual tests on small color differences between metallic samples are poorly predicted by current color difference equations [14]. As an example, the value for the standardized coefficient of the ΔL^{*2} term that we find for reference pairs A and C is appreciably smaller than the value produced by the CIEDE2000 equation. This smaller value is probably due to a wider tolerance to lightness differences, as caused by the presence of metallic texture. This tolerance widening has also been found in earlier publications on metallic samples and other textures [15,12,11]. In a previous publication, we reported that the tolerance on lightness is widened by a factor of up to 1.6 in the presence of metallic texture [12]. From the coefficients that we now find for the ΔL^{*2} term, we calculate that, with reference pair A, the tolerance widening factor is 1.3, and for reference pair C we find 1.6. These values are, therefore, within the range reported before.

On the other hand, for reference pair B, the value for the standardized coefficient of the ΔL^{*2} term that is reported in Table 3 (top) is not smaller than the corresponding value of the CIEDE2000 equation. For this reference pair, the presence of metallic texture does not lead to a widening of the tolerance for lightness differences. Since reference pair B represents a color difference dominated by a lightness difference, we speculate that, for this reference pair, lightness differences within sample pairs are relatively easy to compare with the lightness difference in reference pair B, thus compensating for the widened tolerance for lightness differences that is due to the presence of metallic texture.

The standardized coefficients of the three models in Table 3 (top) show that the models for reference pairs A and C agree quite well with each other, whereas the model for reference pair B is different. This is also concluded from Fig. 4, where we show how the predictions of the three models correlate with each other. Clearly, the models for reference pairs A and C correlate best with each other.

Table 3 also shows the values that we find for Somers' D parameter, which is a measure of the association between model predictions and the visual scores [16]. According to the values shown, predictions with the optimized models agree better with the present visual data than when using the CIEDE2000 equation.

We can use the optimized model for each reference pair to predict for each sample pair the probability that observers assess it as having a larger perceived color difference than the reference pair. In Figs. 5(a) and 5(b) we show that, for reference pairs A and B, this predicted probability correlates well with the percentages of visual assessments that actually do rate the perceived color difference as being larger than for the reference pair. For reference pair C, Fig. 5(c) shows that the optimized model does not estimate well if observers assess a sample pair as having a larger color difference than the reference pair. This is why the Somers' D value for reference pair C is relatively small, as shown in Table 3.

Figure 5(a) shows that, for reference pair A, the optimized model is well able to separate the two main categories of visual data for this reference pair: those sample pairs for which less than 40% of observers rate the perceived color difference as being larger than the color difference in the reference pair, versus those sample pairs for which this percentage is larger than 60%. Although this separation is well described by the optimized model, the order of probabilities for sample pairs within each category is not predicted well by the model.

Based on the optimized models for the three reference pairs, we can plot the tolerance ellipsoids corresponding to predicting a 50% chance for observers to assess a sample pair as having a larger perceived color difference than the reference pair. For the three different reference pairs, these tolerance ellipsoids are shown in Fig. 6(a). To illustrate the different tolerances to lightness differences for the three models, the ellipsoids in Fig. (6) are drawn for a plane in CIELAB-space with a lightness value

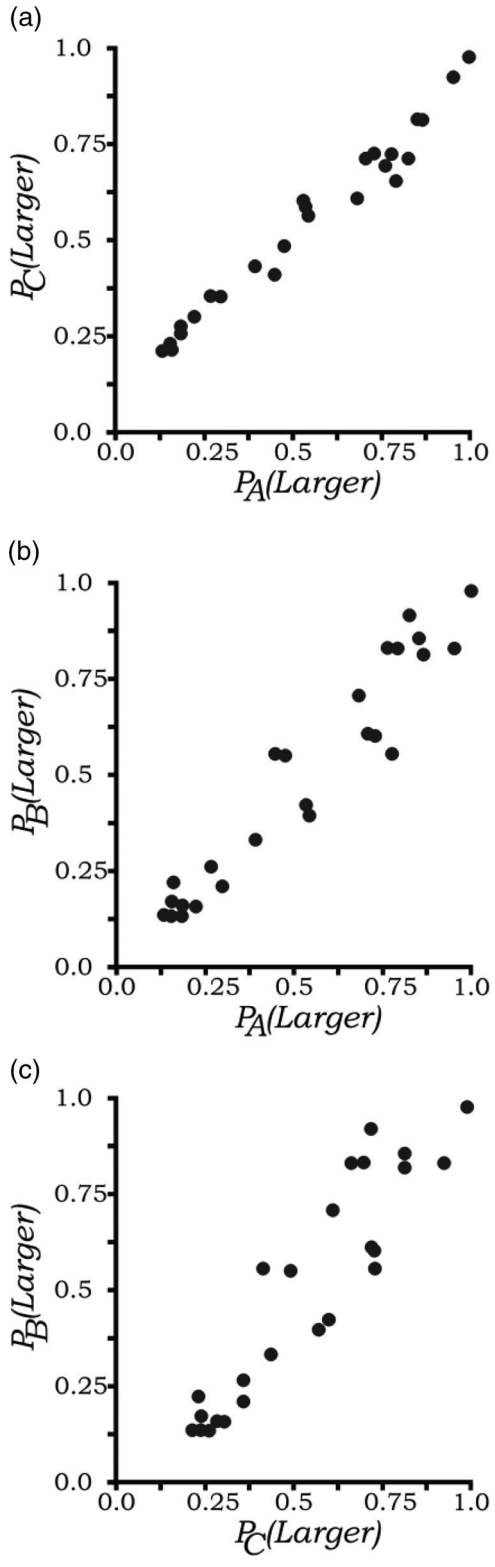

Fig. 4. Predicted percentages of visual assessment stating that the sample pair has a larger perceived color difference than the reference pair, for optimized models for the red color center using (a) reference pairs A and C, (b) A and B, and (c) B and C.

one unit larger than the lightness value of the color center. Models with a tighter tolerance for lightness will then show up with a smaller ellipse.

As another illustration of the tolerance predicted by the optimized models, we use the models to predict the probability that a hypothetical sample pair would be assessed as having a perceived color difference that is larger than the reference pair. This probability is calculated for the case that the sample pair would have the color coordinates and texture values of the red color center, and color differences ΔL^*, ΔC^*, and ΔH^* equal to the difference values of the reference pair. We find a probability of 17% for reference pair A. For reference pairs B and C, the same procedure produces probabilities of 39% and 46%, respectively. Since these percentages are all smaller than 50%, the models consistently predict the average observer to be more tolerant for color differences in the case of textured red samples than in the case of samples with the same uniform gray color as the reference pairs. From the data collected here, it is not possible to determine if this increased tolerance is caused by the sample pairs having a surface texture, or if it is caused by the color center being red.

B. Yellow-Green Color Center

1. Reproducibility and Repeatability

The reproducibility of the visual assessments that we find for the yellow-green color center is very similar to the values for the red color center, as demonstrated in Table 2. While for the red color center we found very similar reproducibility for reference pairs A and B, with reference pair C showing significantly worse reproducibility, for the yellow-green color center all three reference pairs show significantly different reproducibility. Although the average differences are relatively small, it is a consistent finding for most of the observers separately.

For the yellow-green color center, we decided to introduce additional visual sessions in order to better estimate the repeatability of visual assessments. The visual sessions for each reference pair were repeated in another session by four of the 10 observers. For each reference pair, the repeatability of visual assessments is expressed as the percentage of visual assessments that is equal in both sessions. As shown in Table 2, we find that the repeatability is very similar to the reproducibility. The same conclusion was also found for the red color center.

2. Optimized Model

Before optimizing the model coefficients, we first calculated their values when using the CIEDE2000 equation [2,3]. When we optimize a quadratic model to describe the visual data, we find that, for the yellow-green color center, the $\Delta L^*\Delta H^*$ term does obtain a value that significantly differs from zero, in contrast to what we found for the red color center. With log-likelihood tests we found that, for each of the reference pairs, the addition of the $\Delta L^*\Delta H^*$ term is significant at $\alpha = 0.01$. As shown in Table 3 (bottom), the addition of the term in $\Delta L^*\Delta C^*$, which could not be investigated for the red color center, was found to not be significant for the yellow-green color center.

Similar to our analysis for the red color center, a log-likelihood test showed that, also for the yellow-green color center, the optimized models for the three different reference pairs are statistically different. In this case, this is caused by a significant interaction between the reference pair factor and the coefficient for the ΔH^{*2} term at the $\alpha = 0.01$ level.

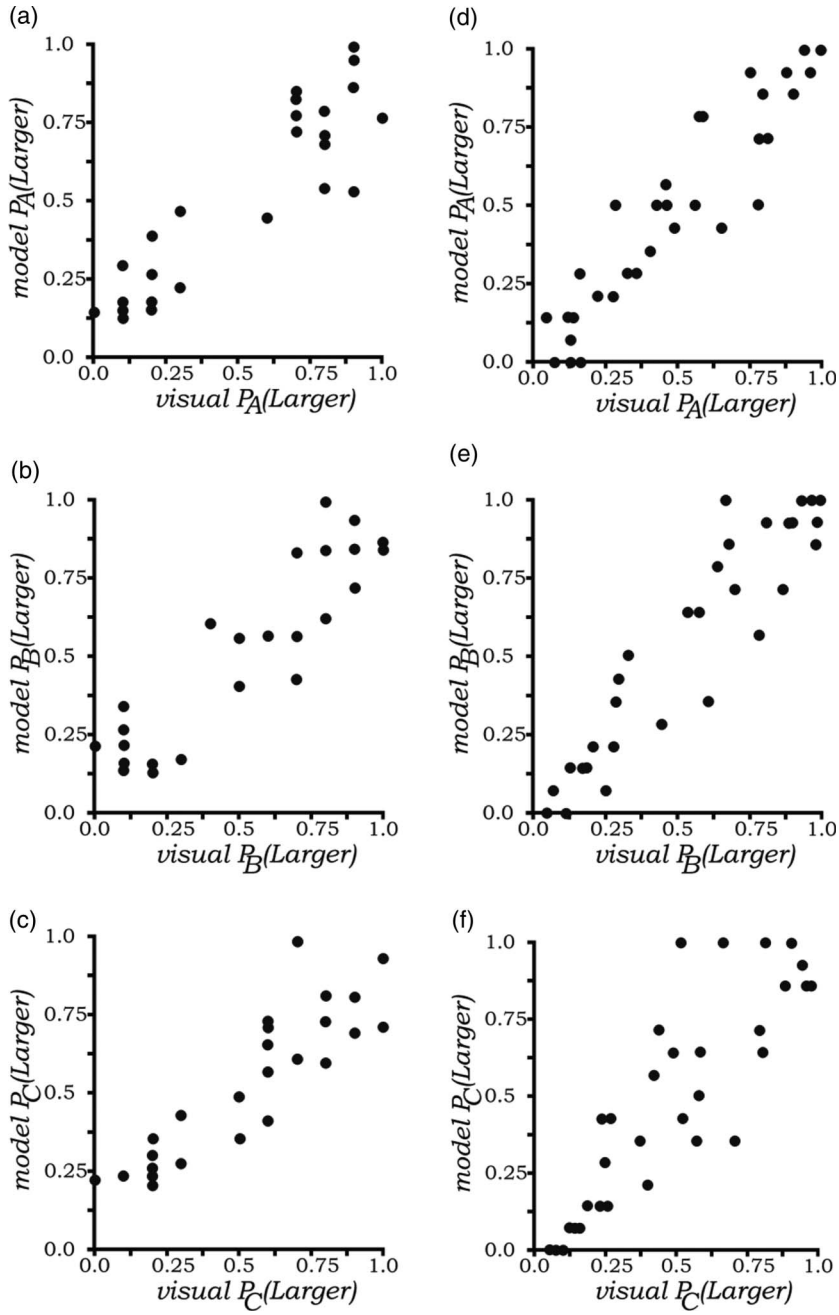

Fig. 5. Predictions by optimized models, compared to visual data, for the red color center and (a) reference pair A, (b) reference pair B, and (c) reference pair C. For the yellow-green color center, similar graphs are shown for (d) reference pair A, (e) reference pair B, and (f) reference pair C.

Table 3 (bottom) shows that, also for the yellow-green color center, the optimized models show smaller standardized coefficients for the ΔL^{*2} term than the corresponding value from CIEDE2000. The values are even smaller than what we found for the red color center, and in this case, the value for reference pair B has also dropped considerably. The smaller values of this coefficient for the yellow-green color center may be explained by the stronger sparkle effect in this set of samples. For this color center, the values shown in Table 3 (bottom) correspond to a widening in lightness tolerance of 1.5 to 1.6 for all three reference pairs. As mentioned before, this is in the reported range of this factor due to the presence of sparkle [12].

The values of the coefficients of ΔE_{cmc} agree better with the optimized model coefficients for ΔL^{*2}, ΔC^{*2}, and ΔH^{*2} than the corresponding local values in the CIEDE2000 equation, but obviously they do not include the rotation terms $\Delta C^*\Delta H^*$ and $\Delta L^*\Delta H^*$ that we do find in our optimized models. Therefore, also for the yellow-green color center, our results confirm earlier conclusions that current color difference equations do not well describe small color differences between metallic samples [14].

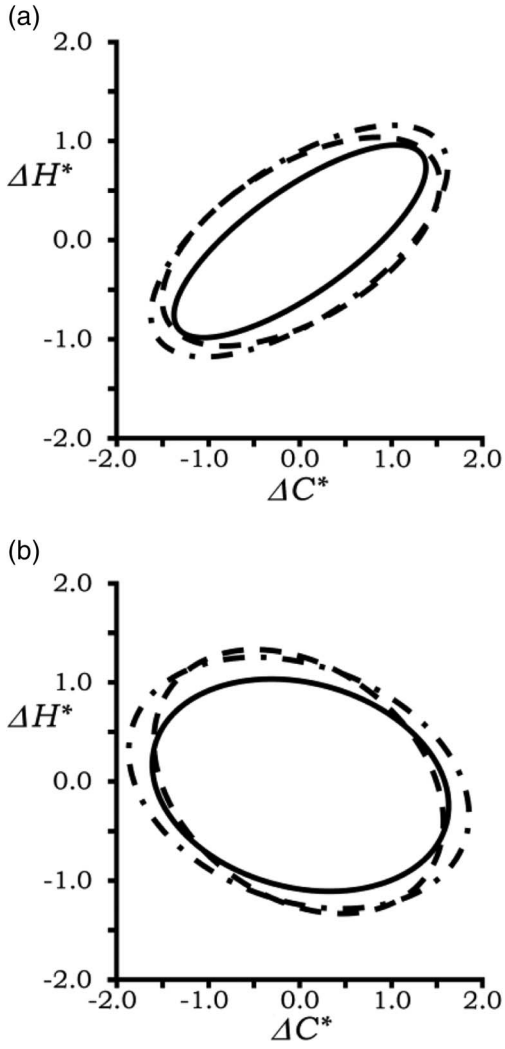

Fig. 6. Predicted 50% tolerance ellipse in the ΔC^*–ΔH^* plane, for the three different reference pairs, for the (a) red and (b) yellow-green color centers. Dashed lines refer to reference pair A, straight lines to reference pair B, and dashed-dotted lines to reference pair C.

Similar to what we described for the red color center, we can again use the optimized models to predict the probability that a hypothetical sample pair would be assessed as having a perceived color difference that is larger than the reference pair, if the sample pair would have the color coordinates and texture values of the yellow-green color center, and color differences ΔL^*, ΔC^*, and ΔH^* equal to the difference values of the reference pair. For reference pairs A, B and C we now find probabilities of only 8%, 4% and 29%, respectively. With all these percentages being substantially smaller than 50%, all these models predict the average observer to be more tolerant for color differences in case of textured yellow-green samples than in case of samples with the same uniform gray color as the reference pairs. The same result was found for the red color center, but it is more pronounced for the yellow-green center with its stronger sparkle effect.

Figures 5(d)–5(f) show the correlation between the predicted probabilities of the optimized models versus the actual probabilities observed in the visual tests, for each of the three reference pairs. These graphs show that for the yellow-green color center, this correlation is quite good for reference pairs A and B, and worse for reference pair C. This is also indicated by the values for the Somers' D parameter, as shown in Table 3. We note that for the red color center, we found the same result.

The tolerance ellipsoids that can be calculated for the yellow-green color center are shown in Fig. 6(b), for all three reference pairs. When compared to the tolerance ellipsoids for the red color center [Fig. 6(a)], it becomes clear that for both color centers the tolerance ellipsoid for reference pair B is the tightest, followed by the ones for reference pair A and then reference pair C. All these tolerance ellipsoids are plotted for the plane defined by a lightness value 1 unit larger than the lightness value of the color center. Therefore the relative sizes of these ellipsoids shows that indeed using reference pair B leads to a tighter tolerance in lightness than when using reference pair A, which in turn leads to a tighter tolerance in lightness than when using reference pair C (as is also clear from the model coefficients in Table 3). Further, Fig. 6 shows that the ellipsoids have different rotation angles with respect to each other.

4. CONCLUSIONS

The main conclusion from this work is that current color difference formulas such as CIEDE2000 and CIE94 would change significantly if a different reference pair would have been used in the constant stimuli experiments underlying these formulas. In other words, color difference formulas such as CIEDE2000 and CIE94 are less universal than what is often hoped for. These equations have been derived from visual data, collected with the psychophysical model of constant stimuli. In this method, the perceived color difference within a sample pair is compared to that for a reference pair. A particular choice of reference pair was made when creating the RIT-Dupont data that underlies the popular ΔE_{94} and ΔE_{00} color difference equations.

In the present study we have investigated how the coefficients in an optimized color difference formula depend on the type of reference pair selected in the underlying constant stimuli experiment. We selected three different reference pairs. Reference pair A represents the reference pair used when creating the RIT-Dupont dataset. Reference pairs B and C represent cases with predominant lightness and hue difference, respectively. In all three cases, the absolute color difference within the reference pair was very similar to the value used in the RIT-Dupont experiments.

Our results show that for the models that are derived for these three cases, the coefficients are significantly different from each other. Therefore, we conclude that the choice for a particular reference pair in the visual tests that produce the experimental data on which the color difference formulas are optimized, has a significant effect on the final expression.

Our results for the red color center show that reference pair C leads to a significantly worse reproducibility. When using the visual data to optimize the coefficients in a color difference equation, we find that the resulting model for visual tests with reference pair C do not predict the visual data well, with a value of Somers' D coefficient of 0.59. This was found to be

significantly worse than for reference pairs A and B, which have a value of Somers' D coefficient close to 0.70. For reference pairs A and B, the model predicts the visual data rather well. The repeatability of visual assessments was found to be of the same magnitude as the reproducibility.

For the yellow-green color center, it is reference pair A that has the worst reproducibility. Based on the resulting value of the Somers' D coefficient, the visual data for reference pair B lead to the most accurate model. Our results therefore imply that visual data can best be modeled with a quadratic function if reference pair B is chosen in the constant stimuli experiment.

For all three reference pairs, the resulting quadratic models show the same terms to be significant. For the red color center, these are exactly the same terms as known from the CIEDE2000 color difference equation. The coefficients for reference pairs A and C are very similar to each other. All three models predict that the tolerance ellipsoid is rotated by approximately 30° as compared to the orientation prescribed by the CIEDE2000 equation. It is tempting to associate this rotation angle to the presence of metallic sparkle in the samples, but this hypothesis needs confirmation from additional visual tests. For the yellow-green color center, the resulting quadratic model adds a $\Delta L^*\Delta H^*$ term to the formula, which is not part of the CIEDE2000 color difference equation.

Our results for both the red and the yellow-green color center show that the tolerance to lightness differences in sample pairs is widened by a factor 1.3 to 1.6. This confirms earlier studies that showed that the presence of metallic texture widens mainly the tolerance to lightness differences. Only for the red color center and reference pair B we found no such influence, possibly because in this reference pair the difference in lightness dominates the overall color difference. For this reason, lightness differences within sample pairs may be relatively easy to compare to the color difference in the reference pair, thus leading to a tightening of the tolerance to lightness differences. This cancels the widening of this tolerance due to the presence of metallic sparkle for the red color center. For the yellow-green color center this cancellation is not found, possibly because the samples that we used for this color center have a stronger sparkle effect than the samples we used for the red color center.

We already concluded that visual data can best be modeled with a quadratic function if reference pair B is chosen in the constant stimuli experiment. There are two possible explanations for this result.

The first explanation is that that lightness differences are easier to assess by observers than other components of color differences (saturation and hue). This is already known from earlier studies [17,18].

An alternative explanation uses the observation that reference pair B is the only reference pair for which the direction of the color difference in CIELAB color space is included explicitly in the quadratic models that we test. Visual data from constant stimuli experiments in which the reference pair has a color difference vector that is not included explicitly in the quadratic terms of the model may show larger scatter, and may therefore lead to models that describe the visual data less accurately. For modeling such experiments, it might be better to use quadratic models in component terms corresponding to color difference vectors parallel and perpendicular to the color difference vector from the reference pair.

The latter explanation may help to explain why in our recent study of the two-alternative forced choice (2AFC) method we found that a quadratic model is not able to accurately fit the visual data [8]. In a 2AFC experiment, observers are confronted with three samples, A, B, and C, and they are asked to assess if the color difference between samples A and B is larger or smaller than between samples B and C. Since after every assessment the three samples are replaced by three other samples, one may consider this method as a constant stimuli experiment in which the reference pair changes for every assessment. The conclusions obtained from the current investigation would predict the visual data obtained with the 2AFC method to be not well described by a single tolerance ellipsoid, which is indeed what we found before [8].

Funding. Ministerio de Economía y Competitividad (MINECO).

Acknowledgment. This work was supported by the research project FIS2013-45952-P ("Ministerio de Economia y Competitvidad," Spain) with support from the European Union (FEDER, European Regional Development Funds).

REFERENCES

1. "Industrial colour-difference evaluation," CIE Publication No. 116 (CIE Central Bureau, 1995).
2. "Improvement to industrial colour-difference evaluation," CIE Publication No. 142 (CIE Central Bureau, 2001).
3. M. R. Luo, G. Cui, and B. Rigg, "The development of CIE 2000 colour difference formula: CIEDE2000," Color Res. Appl. **26**, 340–350 (2001).
4. 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).
5. R. S. Berns, D. H. Alman, L. Reniff, G. D. Snyder, and M. R. Balonon-Rosen, "Visual determination of suprathreshold colour-difference tolerances using probit analysis," Color Res. Appl. **16**, 297–316 (1991).
6. S. Shen and R. S. Berns, "Color-difference formula performance for several datasets of small color differences based on visual uncertainty," Color Res. Appl. **36**, 15–26 (2011).
7. R. G. Kuehni, "Variability in estimation of supratheshold small color differences," Color Res. Appl. **34**, 367–374 (2009).
8. E. Kirchner, N. Dekker, M. Lucassen, L. Njo, I. van der Lans, P. Urban, and R. Huertas, "How psychophysical methods influence optimizations of color difference formulas," J. Opt. Soc. Am. A **32**, 357–366 (2015).
9. S. Shen and R. S. Berns, "Evaluating color difference equation performance incorporating visual uncertainty," Color Res. Appl. **34**, 375–390 (2009).
10. N. Dekker, E. Kirchner, R. Supèr, G. J. van den Kieboom, and R. Gottenbos, "Total appearance differences for metallic and pearlescent materials: contributions from color and texture," Color Res. Appl. **36**, 4–14 (2011).
11. R. Huertas, M. Melgosa, and E. Hita, "Influence of random-dot textures on perception of supra-threshold color differences," J. Opt. Soc. Am. A **23**, 2067–2076 (2006).
12. E. Kirchner, N. Dekker, R. Supèr, G. J. van den Kieboom, and R. Gottenbos, "Quantifying the influence of texture on perceived color differences for effect coatings," in *Proceedings of the 11th Congress of the International Colour Association* (AIC, 2009).
13. E. Kirchner, G. J. van den Kieboom, S. L. Njo, R. Supèr, and R. Gottenbos, "The appearance of metallic and pearlescent materials," Color Res. Appl. **32**, 256–266 (2007).
14. R. Huertas, A. Tremeau, M. Melgosa, L. Gomez-Robledo, G. Cui, and R. Luo, "Checking recent colour-difference formulas with a dataset of

metallic samples and just noticeable colour-difference assessments," in *Proceedings of the Computer Graphics, Imaging and Visualization Conference (CGIV)* (IEEE, 2010), pp. 504–509.
15. E. D. Montag and R. S. Berns, "Lightness dependencies and the effect of texture on suprathreshold lightness tolerances," Color Res. Appl. **25**, 241–249 (2000).
16. R. H. Somers, "A new asymmetric measure of association for ordinal variables," Am. Sociol. Rev. **27**, 799–811 (1962).
17. H. Zhang and E. D. Montag, "How well can people use different color attributes?" Color Res. Appl. **31**, 445–457 (2006).
18. M. Melgosa, M. J. Rivas, E. Hita, and F. Viénot, "Are we able to distinguish color attributes?" Color Res. Appl. **25**, 356–367 (2000).