The Role of Presented Objects in Deriving Color Preference Criteria from Psychophysical Studies

Michael P Royer[1]
Minchen Wei[2]

[1]Pacific Northwest National Laboratory
620 SW 5th Avenue, Suite 810
Portland, OR 97204
michael.royer@pnnl.gov

[2]Hong Kong Polytechnic University

This is an archival copy of an article published in *LEUKOS*. Please cite as:
Royer MP, Wei M. 2017. The Role of Presented Objects in Deriving Color Preference Criteria from Psychophysical Studies. LEUKOS. 13(3):143-57. DOI: 10.1080/15502724.2016.1271339.

Abstract

Although it is a critical component of any measure of color rendition, a standardized set of color samples can seldom perfectly match a real space or a real set of observed objects. This means there will always be some level of mismatch between predicted and observed color shifts. This article explores how the color distortions of three object sets that could be used in experiments compare to the color distortions predicted using the color evaluation samples of IES TM-30-15 (TM-30). The experimental object sets include those from a recent experiment [Royer and colleagues 2016], a set of produce (10 fruits and vegetables), and the X-Rite ColorChecker Classic. This numerical analysis focuses on the range of differences between viewed and characterized color shifts—using the TM-30 Fidelity Index (R_f), the TM-30 Gamut Index (R_g) and an alternative to R_g based on ΔC in CIECAM02—over a set of 344 spectral power distributions. The differences depended on the average chroma and spectral features of the sample set. The substantial range of differences shown for the produce and the ColorChecker means that design criteria for color rendition derived using these sample sets are less reliable. Specifiers should carefully consider how average measures of color rendition are applied to real spaces, and experimenters should carefully select experimental objects to avoid mischaracterizations.

PNNL-SA- 121334

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <http://www.ntis.gov/about/form.aspx>
Online ordering: http://www.ntis.gov

This document was printed on recycled paper.
(8/2010)

1 Introduction

One of the early human factors experiments to focus on color rendition was conducted by C.L. Sanders in 1959 [Sanders 1959]. Participants viewed six natural objects under different spectra created by mixing fluorescent lamps of varying chromaticities. In his article describing the work, Sanders did not include the spectral power distributions (SPDs) of the lamps, but did include the spectral reflectance functions of the objects, concluding that the range of acceptable colors of an object was very dependent on the object. Although the importance of the object was emphasized in this early work, subsequent research has deemphasized characterizing the spectral reflectance functions of the objects, focusing instead on the qualities of different SPDs. This article takes a fresh look at the role of specific objects, as characterized by their spectral reflectance functions, in psychophysical experiments on color rendition.

As the quest for psychophysically-relevant color rendition measures has surged in the last decade, a number of original experiments have been conducted that ask people about their impressions of a group of objects, relating their responses to established color rendition measures for the sources used in the experiment. Reports on these experiments have typically focused on the values of color rendition measures derived from SPDs, with emphasis on trying to understand if the values can correctly match the rank order of the participants' perceptions. However, the psychophysical stimulus (and independent variable) in studies on color rendition is not the SPD alone, or any derived color rendition measure; the visual stimulus results from the interaction of the SPD and the specific objects being viewed. As an extreme example, one should not expect CIE R_a to predict perceptions of a set of only red objects. The same is true of more modern measures of average color fidelity, such as the Illuminating Engineering Society's (IES) Fidelity Index (R_f), which is part of TM-30-15 (TM-30) [IES 2015].

Color rendition measures—especially ones seeking to characterize average performance—are calculated using standardized sets of spectral reflectance functions, such as the eight pastel test color samples (TCS) used to calculate the Commission Internationale de l'Eclairage's (CIE) General Color Rendering Index (R_a) [CIE 1995] or the 99 color evaluation samples (CES) used in TM-30 R_f or TM-30 R_g [David and others 2015; IES 2015]. In contrast, the objects viewed by study participants have ranged from their own skin, to a selection of food, to packaged consumer goods, to a set of manufactured color samples, such as the X-Rite ColorChecker Classic, also known as the Munsell Color Checker or MacBeth Color Checker. These sets of viewed objects do not match any of the standardized sample sets used for calculating color rendition measures, adding ambiguity to experimental results because the color rendition measures are potentially a weak characterization of the visual stimulus. Even subtle mismatches between the characterized stimulus (based on standardized samples from a color rendition measure) and the experimental stimulus (that is, a set of objects in a room or booth) may indiscreetly lead to results that are not generalizable.

Because a standardized set of spectral reflectance functions may not match a specific group of objects well, it is possible that a source will have a relatively high fidelity value, but render the specific objects with substantially lower fidelity. This may ultimately contribute to erroneous findings about the psychophysical meaning or "accuracy" of various measures, such as CIE R_a or TM-30 R_f. For example, a rank order for perceived fidelity may accurately reflect the fidelity of the visual stimulus, even though it does not match the rank order according to a standardized fidelity measure. With so few research experiments reporting the properties of the observed objects—which appear to be chosen most frequently based on anecdotal evidence of distribution across the hue range—it is impossible to determine the true effect of this oft-overlooked phenomenon. However, it is likely that the mismatch has rarely been considered, based on the analyses presented in journal articles.

1.1 Object Sets from Past Experiments Focusing On Color Preference and Other Subjective Impressions of Color Rendition

The quantity and type of objects included in color rendition experiments has varied considerably. Some narrowly-focused experiments (or parts of experiments) have focused solely on rendition of skin tones [Quellman and Boyce 2002; Teunissen and others 2016; Veitch and others 2002; Wei and others 2014a; Wei and others 2014b], and others have focused on printed images or the X-Rite ColorChecker [Islam and others 2013; Liu and others 2013; Rea and Freyssinier-Nova 2008; Schanda and Sandor 2003; Szabó and others 2009; Veitch and others 2014]. Some have primarily examined fruits and vegetables [Jost-Boissard and others 2009; Liu and others 2013; Ohno and others 2015; Rea and Freyssinier 2010; Teunissen and others 2016; Thornton 1974; Zukauskas and others 2012], while others have included a broader variety of consumer goods [Lin and others 2015; Smet and others 2010; Spaulding 2012; Szabo and others 2014; Wei and others 2014b; Xu and others 2016]. With the exception of several recent studies [Islam and others 2013; Jost-Boissard and others 2014; Royer and others 2016; Smet and others 2010; Wei and others 2016a], most recent literature on color preference includes limited discussion of the attributes of the objects, despite the fact that they are critical to the stimulus being evaluated. Two studies have also demonstrated the varying influence of different types of objects on judgements of color rendition [Royer and others 2016; Wei and others 2016a].

Another important consideration is the number of objects included in the evaluated scene. Most of the previously noted experiments have used a relatively small number of objects (less than 20) presented in a viewing booth, but a few have used more objects presented in full-size spaces [Houser and others 2005; Lin and others 2015; Spaulding 2012; Szabo and others 2014; Wei and others 2014a; Wei and others 2014b]. This contextual factor may also influence participant responses. Rea and Freyssinier point out that with a large number of objects, some observers may focus on one particular color while other observers may give an overall response to the objects in the field of view [Rea and Freyssinier-Nova 2008]. However, applying general color rendition measures to situations that include only a limited set of objects (e.g., only fruit), perhaps not even including all hues, is particularly problematic. Although important, this analysis does not examine how the number of objects presented, or the arrangement of objects, effects outcomes.

1.2 Purpose

The purpose of this article is to illustrate the range of discrepancy, or mismatch, between color rendition measures and experimental stimuli. This was accomplished using three sets of spectral reflectance functions, corresponding to plausible experimental object sets: 1) the reflectance functions of the X-Rite ColorChecker Classic, 2) the spectral reflectance functions of a selection of common fruits and vegetables, and 3) the spectral reflectance functions from a recent experiment [Royer and others 2016]. Using a set of 344 SPDs, custom average color fidelity and average gamut area measures were calculated for each object set using the conceptual methods of TM-30 R_f and R_g, respectively. The custom measures were compared to standard values for the TM-30 measures, as well as CIE R_a.

2 Methodology

2.1 Spectral Power Distributions

This analysis is based on the set of 318 SPDs from the IES TM-30 Calculator Tool Library, which includes a variety of commercial, experimental, and theoretical light sources. This set was augmented with the 26 SPDs used by Royer and colleagues [2016]. The SPDs provide a wide range of possible conditions, with R_f values ranging from 14 to 100 and R_g values ranging from 45 to 125. Characterizations of the SPDs are shown in **Figure 1**. This set of sources was also used in [Royer 2016].

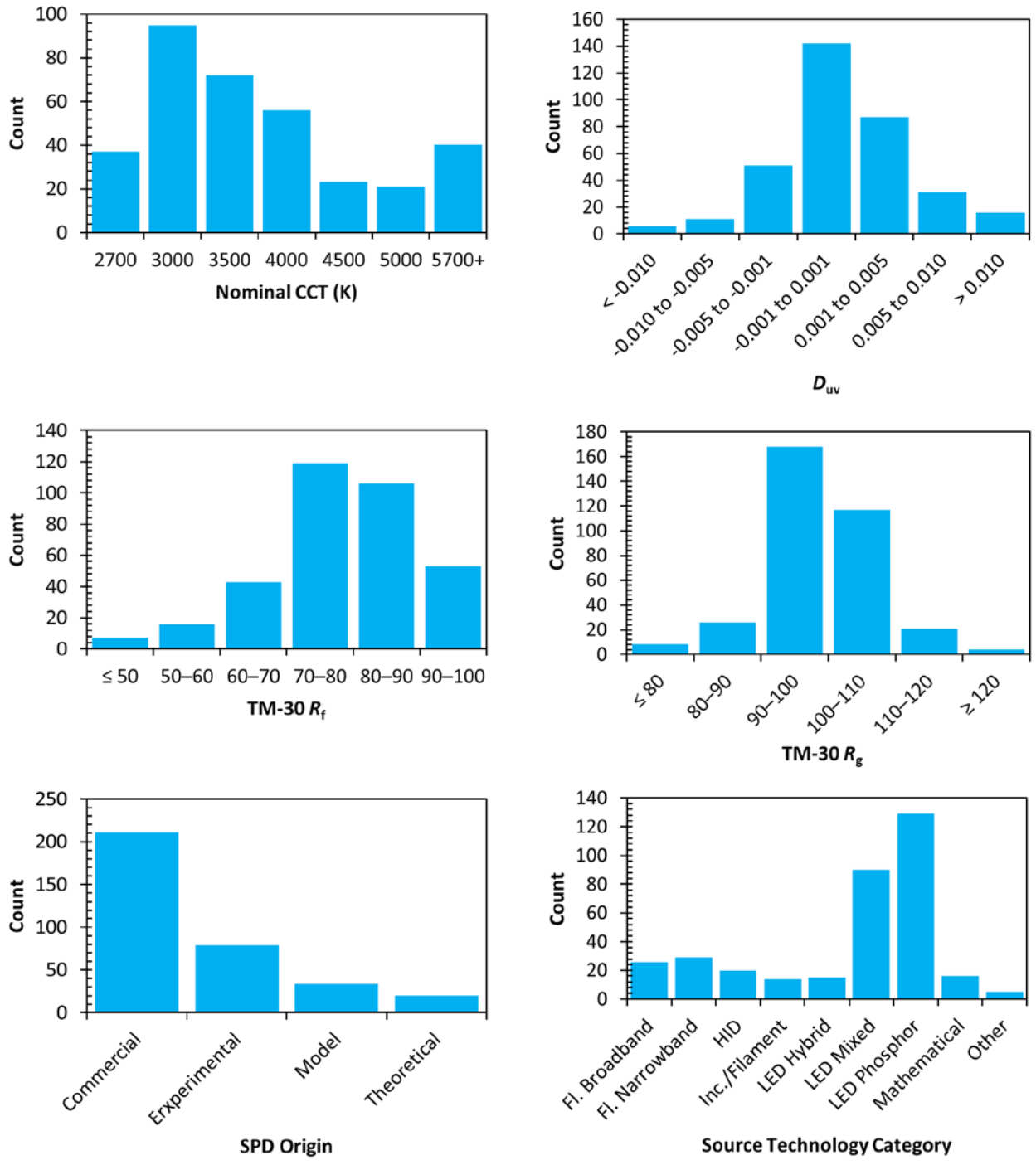


Figure 1. Characterization of the 344 sources comprising the dataset used in this evaluation.

2.2 Object Sets

This analysis is based on three object sets: a set of 122 spectral reflectance functions representing the variety of experimental objects from [Royer and others 2016] (Experiment), a set of 11 spectral reflectance functions for the produce in [Royer and others 2016] (Produce)—including two red apples, a green apple, cranberries, a grapefruit, an orange, dried banana chips, a lemon, a green pepper, blueberries, and a purple cabbage—and the 18 spectral reflectance functions from the non-grey samples of the X-Rite ColorChecker Classic (CCC). All of the spectral reflectance functions were measured using a factory-calibrated Minolta CM-600d spectrophotometer.

2.3 Color Rendition Measures and Calculations

Custom average fidelity measures were calculated for each of the three object sets using the same procedures as TM-30. Given the features of the object sets, it was only possible to calculate custom average gamut area values that adhered to TM-30 R_g procedures using the full set of experimental objects. For the produce and CCC sets, samples were not included in each of the 16 hue angle bins, which are used to determine the vertices of the TM-30 gamut area polygon. For these two sets, average gamut area was calculated using a polygon formed by the furthest outlying color samples. Because this is somewhat different from the TM-30 procedure, a second alternative calculation was performed for all datasets, based on the average difference in chroma (ΔC) of the samples, using the formula for chroma (C) defined in CIECAM02 [Fairchild 2013]. In this analysis, the average ΔC across all samples in the set was scaled by a factor of 6.82, as well as adjusted so that the reference had a value of 100, as shown in Equation 1 (where n is the number of samples in the set). The scaling factor was determined so that the TM-30 R_g and ΔC values for the TM-30 CES across the CIE F Series illuminants were equal (93.0).

$$\Delta C = 100 + 6.82 \frac{\sum_1^n C_{i_{test}} - C_{i_{ref}}}{n} \qquad \textbf{Equation 1}$$

Figure 2 compares values for ΔC to values for average gamut area for the four datasets. As shown, there is strong correlation between the measures, with perhaps 10 to 20 outliers. The outlier points were highly structured SPDs with relatively low fidelity: color-mixed LEDs and mercury vapor lamps. Hue shifts influence the average gamut area, but not the average chroma shift, leading to the noticeable differences between the two measures for sources that induce substantial hue shifts.

The custom values were compared to standard values for the TM-30 R_f and R_g, as well as CIE R_a, for the entire set of 344 SPDs. Average ΔC values were also calculated for the 99 color evaluation samples (CES) of TM-30 in order to serve as the baseline. The level of correlation between the custom and standard values is an indicator of how well the standard measures capture color rendition for the specific object sets. Also important is the magnitude of the residuals, which indicates the error for a given source.

Another important consideration in defining the custom calculations is the scaling factor used for average fidelity calculations. As a default, the scaling factor used for R_f was also used for the custom fidelity calculations, because the intent was to compare potential visual scenes to the characterization of lighting according to standardized CES/TM-30 calculations. This analysis is not focused on evaluating the described object sets for potential use in color rendition measures. Normalizing the scaling factors would reduce the differences in values across the sample sets, revealing only the effects arising from spectral features; however, this normalization would not occur if the samples were viewed in a physical environment, which is the key distinction of this analysis. That is, this analysis investigates how viewed objects differ from standardized sample sets. The implications of maintaining or normalizing scaling factors is shown later, where data using the constant scaling factor is contrasted with data calculated by adjusting the scaling

factor so that the average fidelity scores for the CIE F Series illuminants is equal to that calculated for CIE R_a and TM-30 R_f.

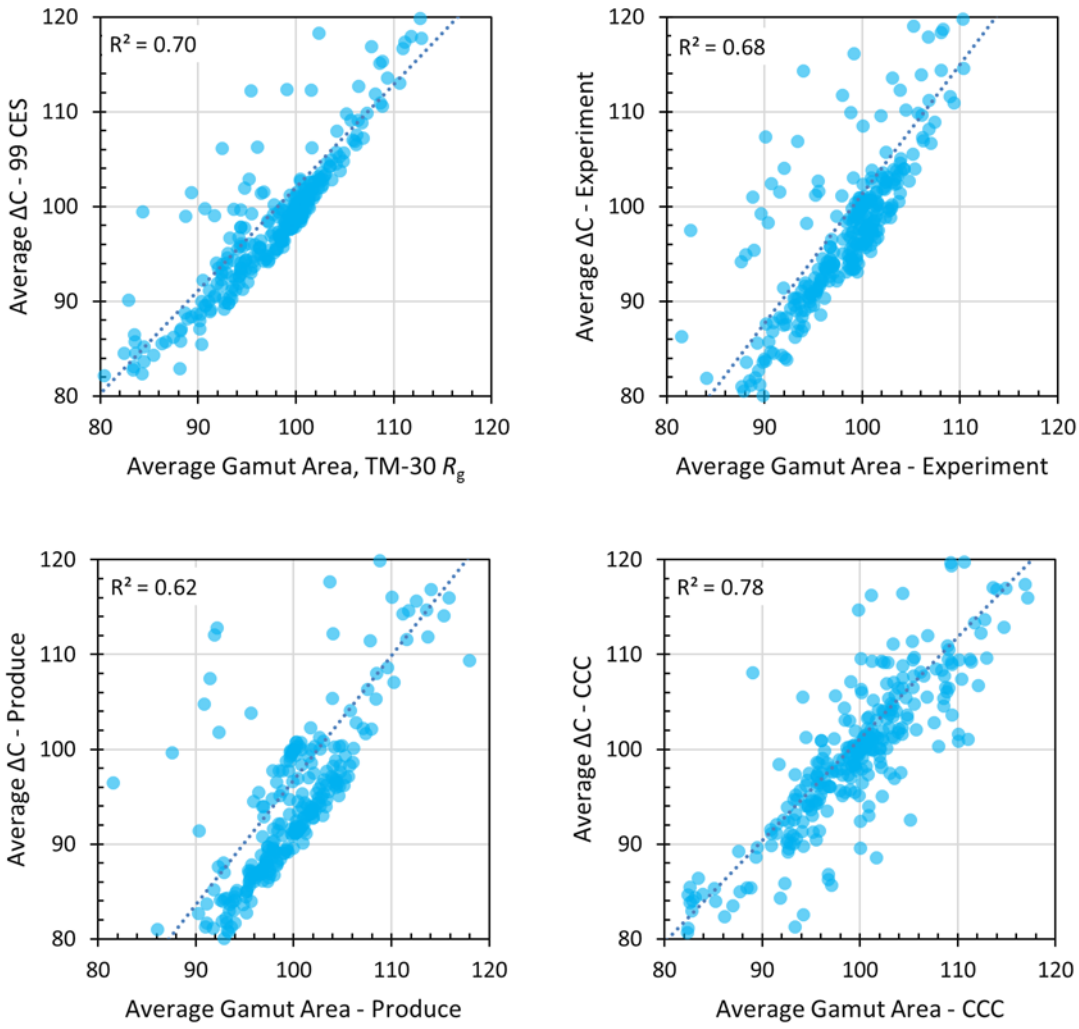


Figure 2. Average change in chroma (ΔC) versus average gamut area for the set of 344 light sources, using the four different sample/object sets. The outlier points are mostly highly structured SPDs (e.g., color mixed LEDs, HID) with low fidelity.

3 Results

Several measures of the difference in average fidelity, average chroma shift, and average gamut area values between the object sets and the 99 CES are provided in **Table 1**. The data are based on subtracting the standard value based on the 99 CES from the custom value based on the identified object set.

Table 1. Statistics for the difference in custom values versus TM-30 R_f, ΔC_{CES}, and TM-30 R_g for each of the 344 SPDs. Positive values indicate the custom value is greater than the standard value.

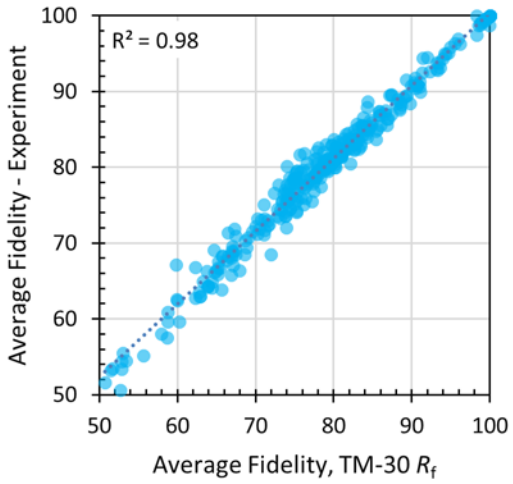
	Difference for Custom Average Fidelity vs. TM-30 R_f			Difference for Custom ΔC vs. ΔC_{CES}			Difference for Custom Average Gamut Area vs. TM-30 R_g		
	Exp.	**Produce**	**CCC**	**Exp.**	**Produce**	**CCC**	**Exp.**	**Produce**	**CCC**
Minimum	-4.5	-16.8	-44.8	-13.8	-24.2	-10.6	-3.5	-10.6	-7.9
Mean	1.2	-1.9	-18.4	-0.9	-4.7	2.0	0.2	0.7	2.7
Maximum	7.3	9.3	1.7	9.8	25.9	17.2	5.0	13.0	33.4
Range	11.8	26.1	46.5	23.6	50.1	27.9	8.5	23.6	41.3
St. Dev.	1.6	4.1	11.3	2.9	7.0	3.8	1.3	3.2	4.3

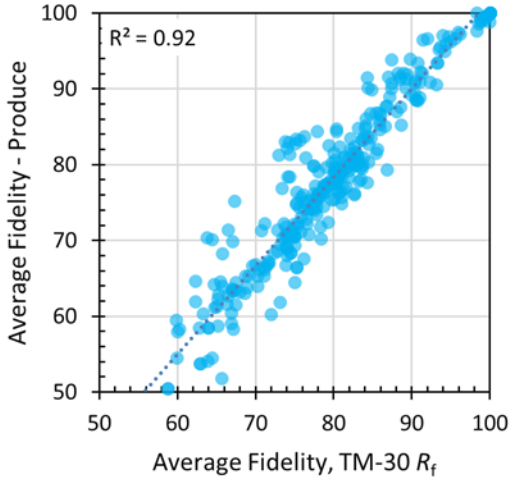
3.1 Average Fidelity Results

For the experiment object set, the custom average color fidelity calculations had a very strong correlation (r^2 = 0.98) with TM-30 R_f, as shown in **Figure 3**. (Note that Figure 3, like others in this report, shows the region of most relevance, rather than all sources that were used to determine the values in Table 1.) This match is not surprising, given that the goal of choosing the objects was to obtain a set with even coverage of color space—a goal in common with the selection of the TM-30 CES. The average difference for the 344 pairs of average fidelity scores was 1.2 points, although average fidelity scores differed by as much as 7.3 points among the 344 SPDs in the dataset. The total range in values was from -3.4 to 7.3 (10.7 points), with positive values indicating the R_f under-predicted the average color fidelity of the experimental objects and negative values indicating that R_f over-predicted the average color fidelity of the experimental objects.

For the produce object set, the custom average fidelity was as much as 16.8 points lower than predicted by TM-30 R_f, and differed by -1.7 points on average. The correlation was still very strong (r^2 = 0.92). The color distortions one would experience if viewing only this bowl of produce were less similar to those generalized by TM-30 R_f than if the full experimental object set was viewed. The total range, or "error", was 26.1 points. This is substantially larger than what is generally considered a meaningful difference in fidelity (about 10 points), even if such rules-of-thumb have been shown to be less useful given the importance of gamut shape [Royer and others 2016]. As is generally the case, the greatest differences were for sources with highly structured SPDs, such as color-mixed LEDs and high-intensity discharge (HID) lamps.

Finally, average fidelity values using the CCC samples showed the greatest disparity with TM-30 R_f. Differences for individual SPDs were as great as -44.8 points, with an average of -18.0 points and a range of 45.5 points. In almost all cases, TM-30 R_f over-predicts the average color fidelity (or color difference) that would be experienced by an observer viewing on the CCC. In addition to color-mixed LEDs and HID lamps, some phosphor-coated LEDs also demonstrated large differences—although they were low-fidelity sources more suited for street lighting than architectural interiors, where color rendition measures are most applicable.





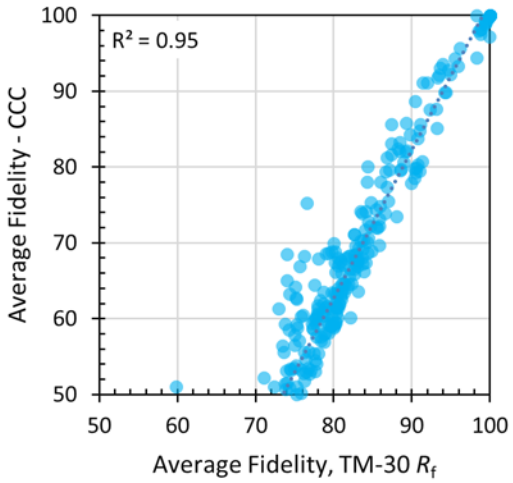


Figure 3. Comparison of average color fidelity for three object sets versus standard TM-30 average fidelity (R_f) calculations. The correlation (identified using the coefficient of determination, r^2) and the magnitude of residuals (the difference between specific values) are dependent on the spectral features and average chroma of the object sets. Note that 7 SPDs (2%) with TM-30 R_f values less than 50 were removed in these plots, because the logarithmic adjustment used in TM-30 to prevent negative scores makes the relationship nonlinear.

These three results can be traced to the specific (a', b') coordinates of the samples in each set (**Figure 4**), which varied substantially in distribution in hue and chroma (**Figure 5**). Using the data from Royer and colleagues [2016], it can be shown that fidelity tends to be related to the chroma of the sample—samples with higher chroma tend to shift more (**Figure 6**), resulting in lower average color fidelity. This is also visible in the data presented by David in analyzing color fidelity over large sample sets [David 2013]. The end result is that object sets with greater average chroma will tend to have lower fidelity values, as was the case with the CCC and produce sets considered here. In practice, study participants viewing a set of objects that is more saturated than the standardized sample set will base evaluations on greater-than-predicted distortions; thus, using their responses to develop specification criteria may be inappropriate.

Importantly, in the different average fidelity measures that have been proposed, this has been addressed through a scaling factor that sets the overall range of values [David and others 2015; Davis and Ohno 2010], yet it is still a practical consideration when choosing objects to be viewed in an experimental setting. The scaling factor addresses the slope of the lines in **Figure 3**, but does not account for the variation from the trend line, which is due to the spectral features of the samples and the distribution within hue space combined with the spectral features of the SPD. **Figure 7** is analogous to Figure 3, but uses custom average fidelity values with sample-set-specific scaling factors.

The variation from the trend line, or residuals, of the comparisons shown in **Figures 3 and 7** are specifically related to the characteristics of the spectral reflectance functions (**Figure 8**) and the resulting CAM02-UCS coordinates. That is, they originate from differences in color space uniformity and wavelength uniformity/spectral features (**Figure 9**) [David and others 2015], two key considerations during the selection of the TM-30 CES [David and others 2015; Smet

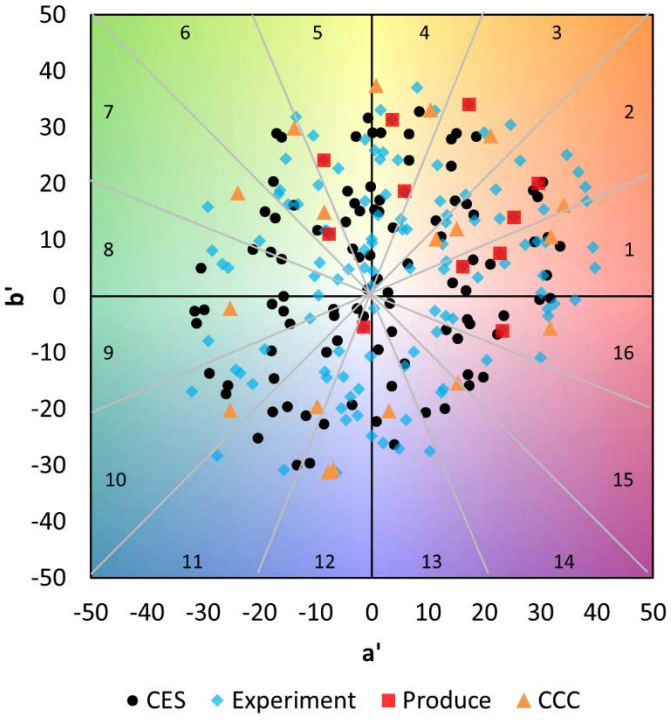


Figure 4. **CAM02-UCS (*a*', *b*') coordinates of the samples in the four sets in the *a*'-*b*' plane of the CAM02-UCS.** The coordinates were determined using 3500 K Planckian radiation, which was an arbitrary choice. The TM-30 hue angle bins and their corresponding numerical identifier are also shown.

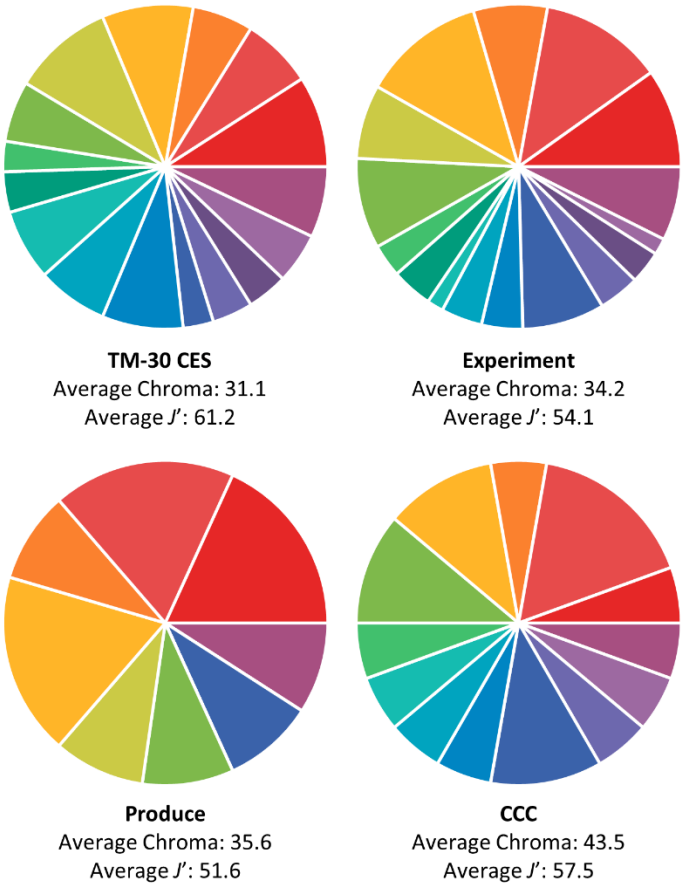


Figure 5. **Visualization of the differences between the four sets of samples/objects.** The pie charts are colored based on approximations for the 16 hue angle bins. Both the distribution in hue space and average chroma vary for the four sets. Note that even for the TM-30 CES, the distribution of samples in the 16 hue angle bins is not even, reflecting the fact that color space is not spherical.

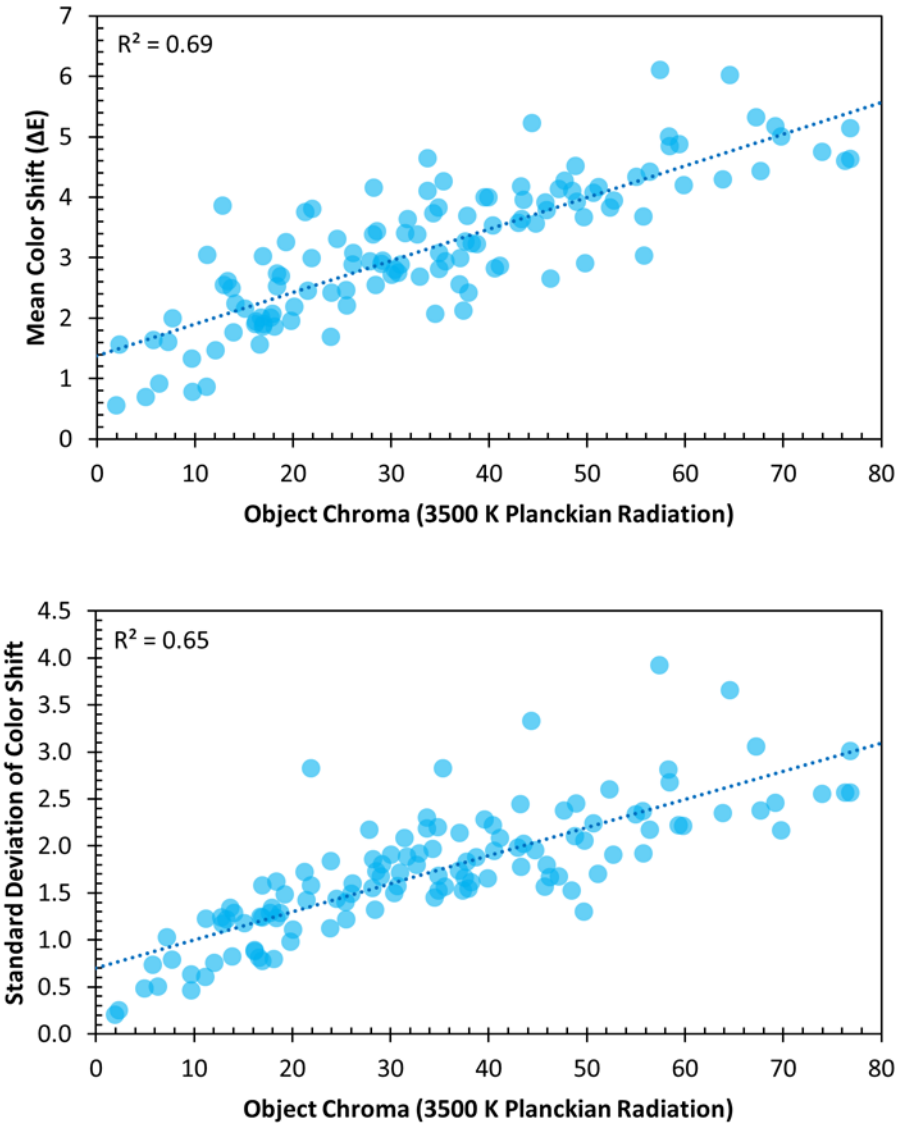


Figure 6. Average color shift (top) and standard deviation of color shift (bottom) for the experimental object set illuminated by the 26 SPDs used by Royer and colleagues [2016]. As object chroma increases, the chromaticity tents to be less stable from light source to light source.

and others 2015]. Because the experiment object set most closely mimics the coverage of color space of the TM-30 CES, it has the smallest residuals. Like all of the object sets, it does not have the same wavelength uniformity, which likely contributes to the differences. The range of the difference between rescaled custom average fidelity values and the standard R_f calculation for the experiment object set is 13.2 points with a standard deviation of 1.6. The CCC dataset also offers a relatively even coverage of hue space, but the samples are not evenly distributed in chroma. As a result, the difference between the rescaled custom average fidelity values and the standard R_f calculation is as 15.7 points with a standard deviation of 2.0 points. Then there is the produce set, which lacks both uniform coverage of color space and has spectral features that are much different from a generalized set of objects. As a result, the range in difference between the rescaled custom average fidelity calculation and the standard R_f calculation for a given SPD is 18.1 points with a standard deviation of 3.0 points.

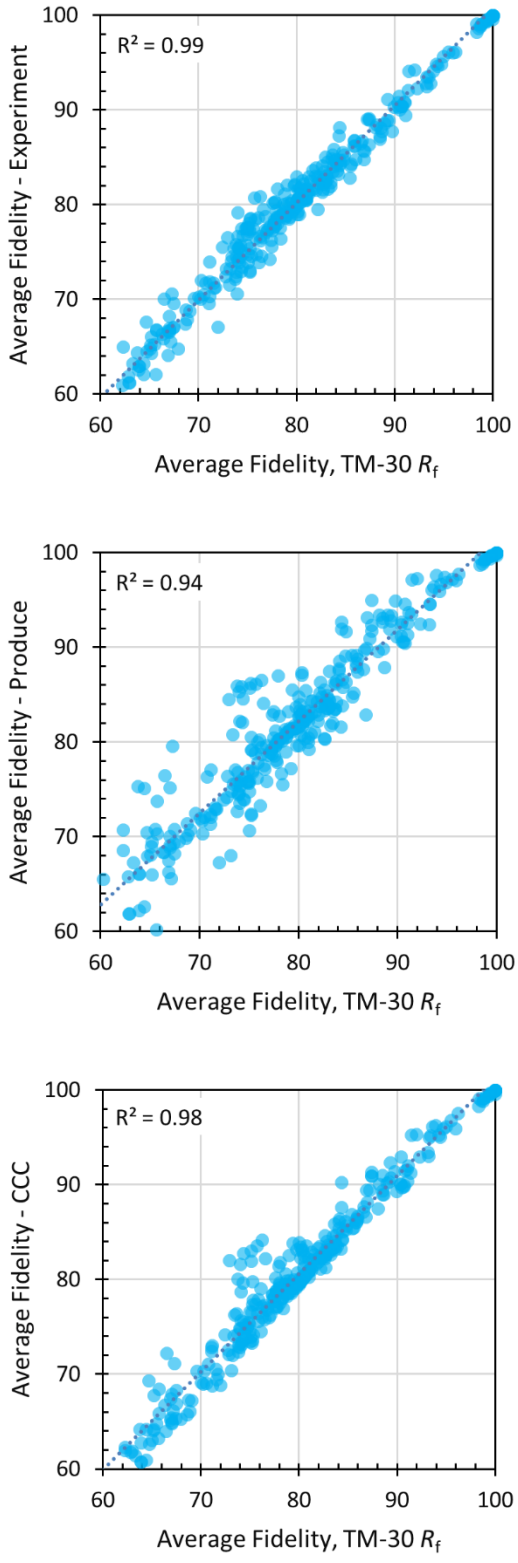


Figure 7. Comparison of average color fidelity for three object sets versus standard TM-30 average fidelity (R_f) calculations. In this case, the scaling factor for each of the custom average color fidelity measures was adjusted using the same procedure that was used to calculate TM-30 R_f. As such, the correlations are about 1:1 on average, but there is still variation due to features of the spectral reflectance functions.

Finally, it is contextually important to see that using the CIE R_a scheme to calculate average color fidelity results in substantially higher score difference ranges for the experimental objects and the produce, but a smaller range for the CCC (**Table 2, Figure 10**). In all cases, the difference is toward over-prediction by CIE R_a, ostensibly because all three object sets have higher average chroma than the pastel TCS employed in the CIE method. Two other key observations emerge:

- The issue of non-uniformity in red region of the CIE U*V*W* color space is readily apparent, with the values calculated using the produce object set—filled with mostly red and orange objects—substantially lower than the standard calculation. Many are familiar with this issue because of the unusual scale of CIE R_9.
- The difference for the average fidelity of the CCC set is lower when compared to CIE R_a instead of TM-30 R_f. This is principally because both the CIE R_a and CCC samples are from the Munsell system, which relies on a limited number of pigments [Cohen 1964]; that is, they have similar spectral features. This further highlights the role of spectral uniformity in sample sets for evaluating color rendition.

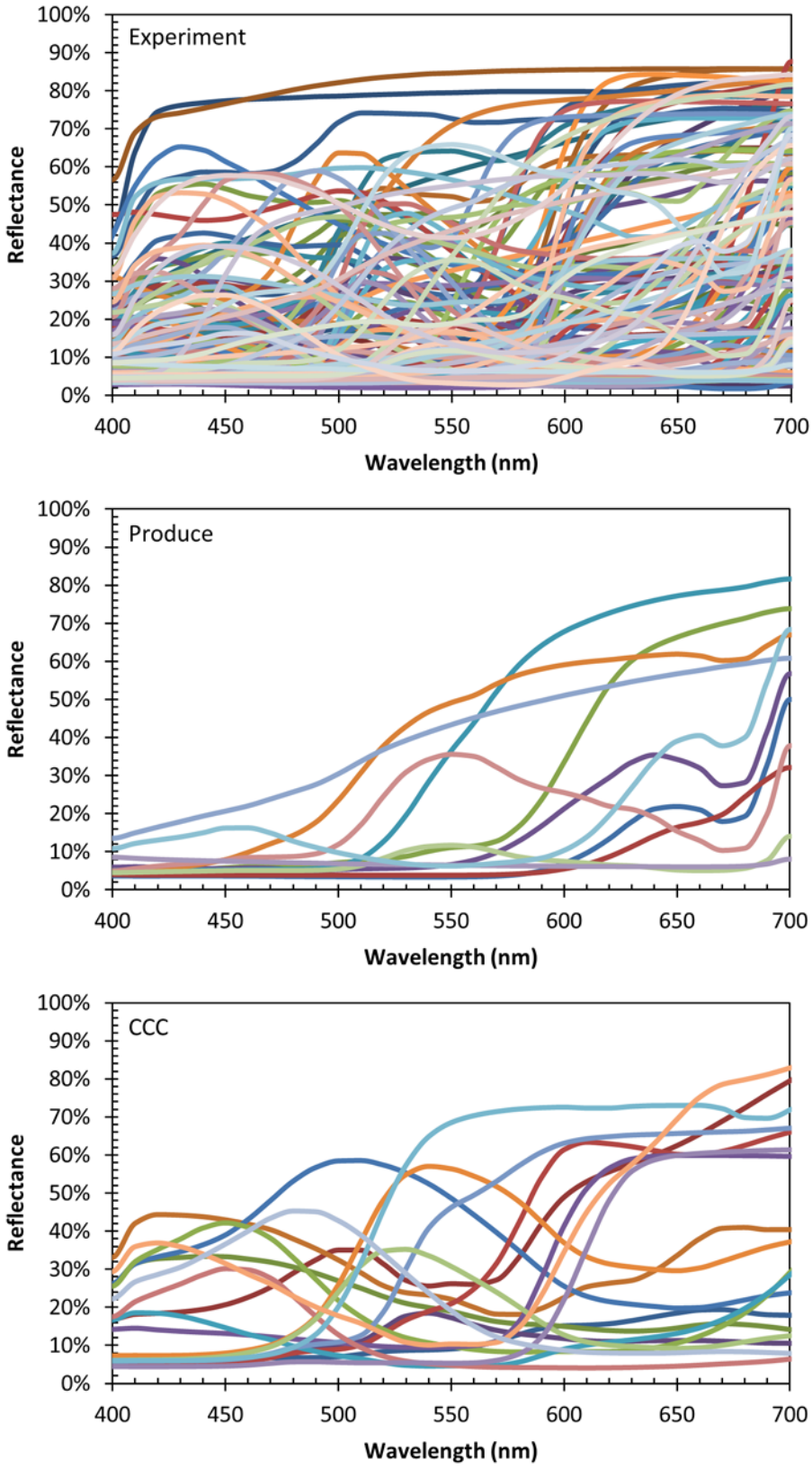


Figure 8. **Spectral reflectance functions for the three non-TM-30 object sets included in this analysis.**

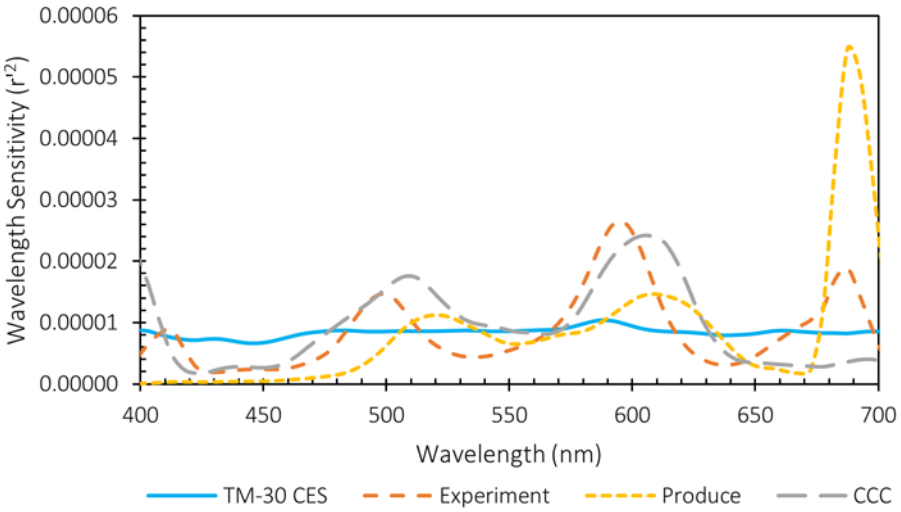


Figure 9. Wavelength sensitivity of the four sample/object sets. For more information on wavelength sensitivity, see [David and others 2015].

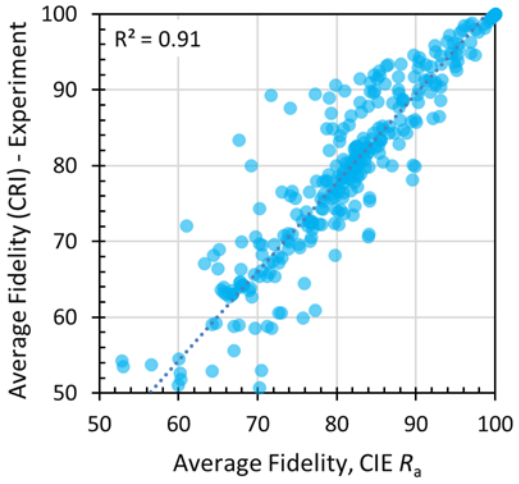
Table 2. Statistics for the difference in values for custom average fidelity, calculated using the CIE CRI procedures but alternative color samples, versus CIE R_a.

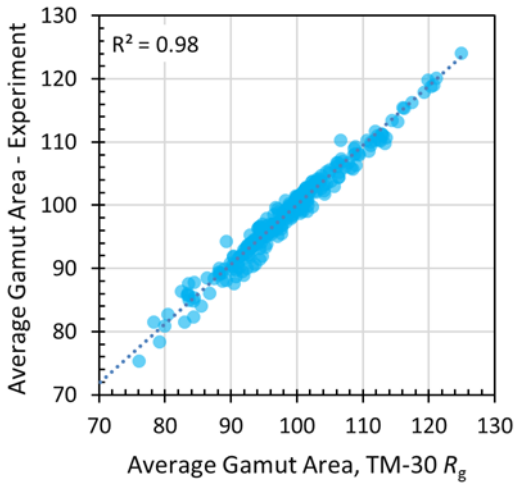
	Difference in Average Fidelity versus CIE R_a		
	Experiment	**Produce**	**CCC**
Minimum	-25.1	-58.4	-27.0
Average	-2.3	-16.7	-8.0
Maximum	17.6	2.5	2.1
Range	42.8	60.9	29.2
Standard Deviation	5.5	11.7	5.4

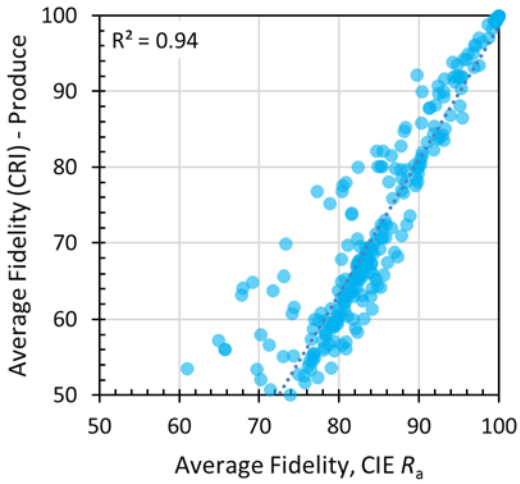
3.2 Average Gamut Area Results

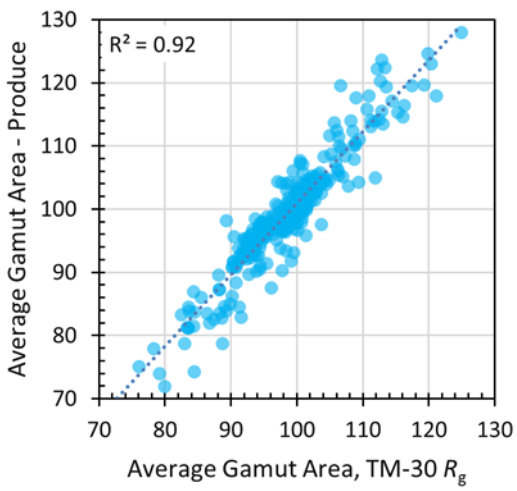
In contrast with the results for average color fidelity, differences in average gamut area calculated for the three object sets versus the TM-30 CES were smaller, on average (**Table 1**). This likely arises because average gamut area calculations are not dependent on a scaling factor, which makes the relationships generally closer to 1:1. In other words, average gamut area is a relative measure, so the average chroma of the color samples has much less influence on the characterization. Plots showing the correlation between average gamut area values calculated using the three object sets versus the standard TM-30 R_g values are provided in **Figure 11**.

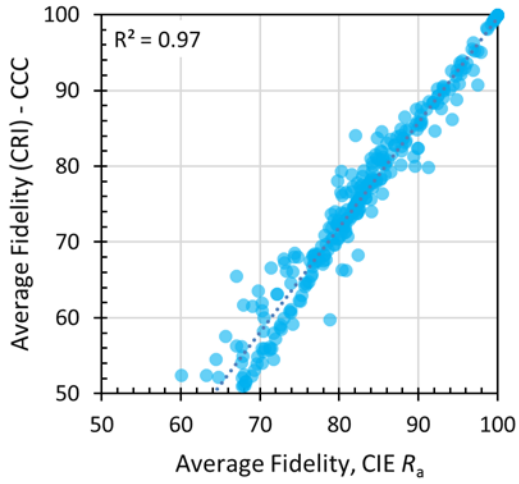
Although the underlying math leads to small average differences in the three comparisons, the range of differences was similar to that found for average color fidelity, with the CCC set having the largest range and the experimental set having the smallest range. For particular light sources, the viewed object set can have a profoundly different average gamut area than is predicted by the TM-30 CES. Especially for the produce and CCC sets, the TM-30 CES tended to under-predict the average gamut area; that is, TM-30 values would tend to be lower than the average gamut area experienced by an observer looking at the three object sets. Again, the largest differences are for highly structured SPDs, which tend to stress color rendition measures. This is











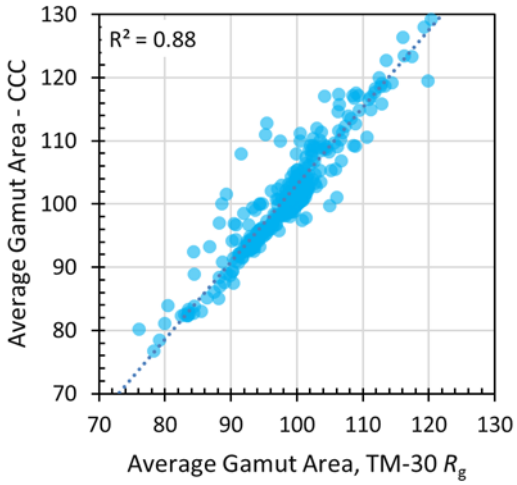


Figure 10. Comparison of average color fidelity for three object sets (using CIE R_a formulas) versus standard CIE R_a calculations. The correlation and magnitude of residuals is dependent on the spectral features and average chroma of the object sets.

Figure 11. Comparison of average gamut area for three object sets versus TM-30 R_g. The correlation and magnitude of residuals is dependent on the spectral features and average chroma of the object sets.

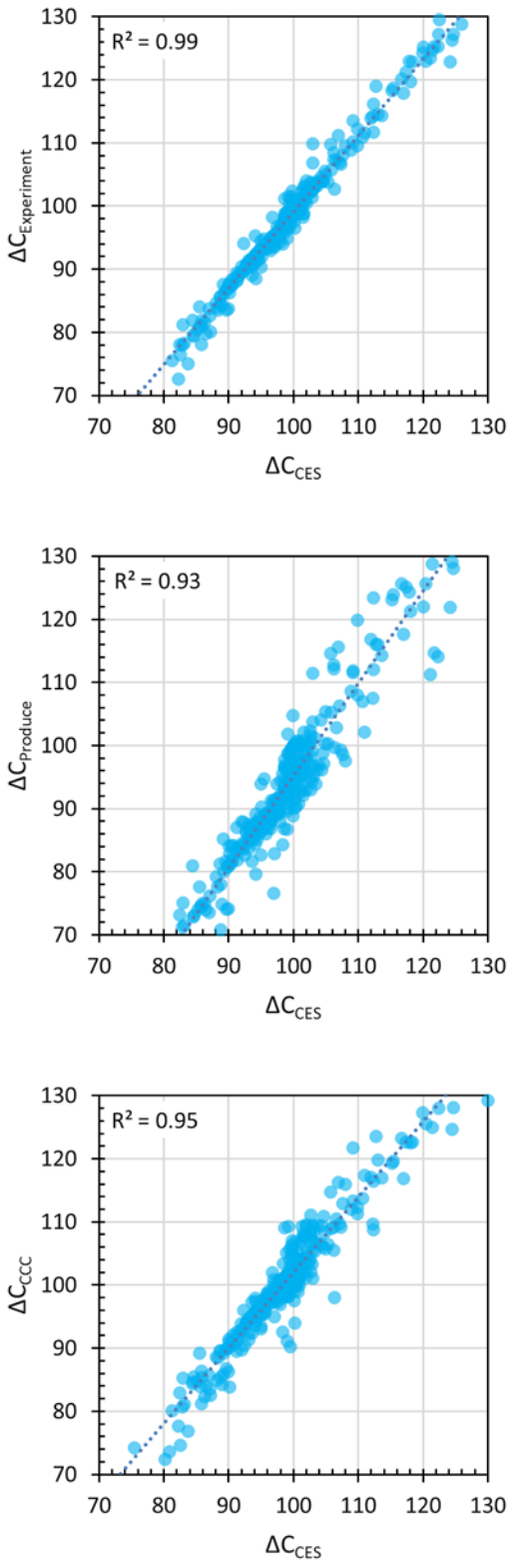


Figure 12. Comparison of average chroma change for three object sets versus average chroma change calculated using the TM-30 color evaluation samples. The correlation and magnitude of residuals is dependent on the spectral features and average chroma of the object sets.

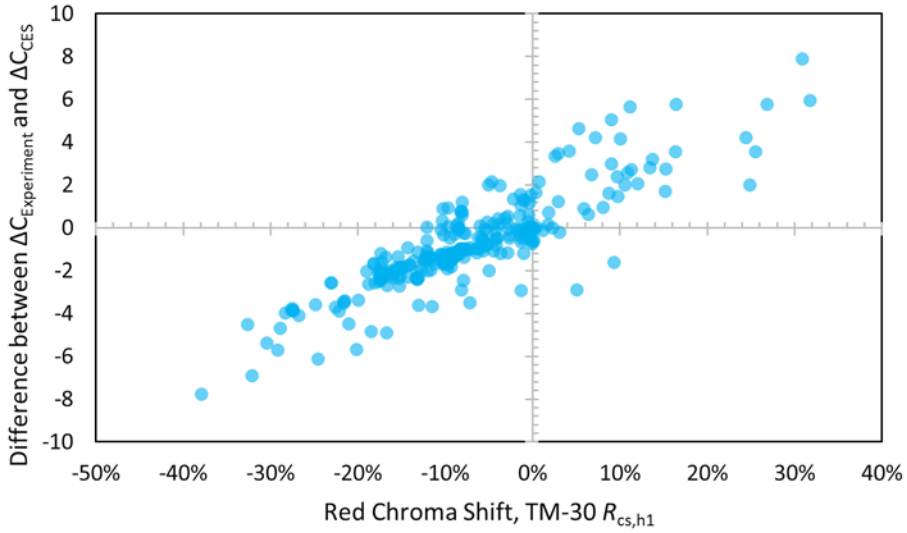
important to keep in mind as such sources become more prevalent, which may come in the form of color-mixed LED sources or laser diodes, for example.

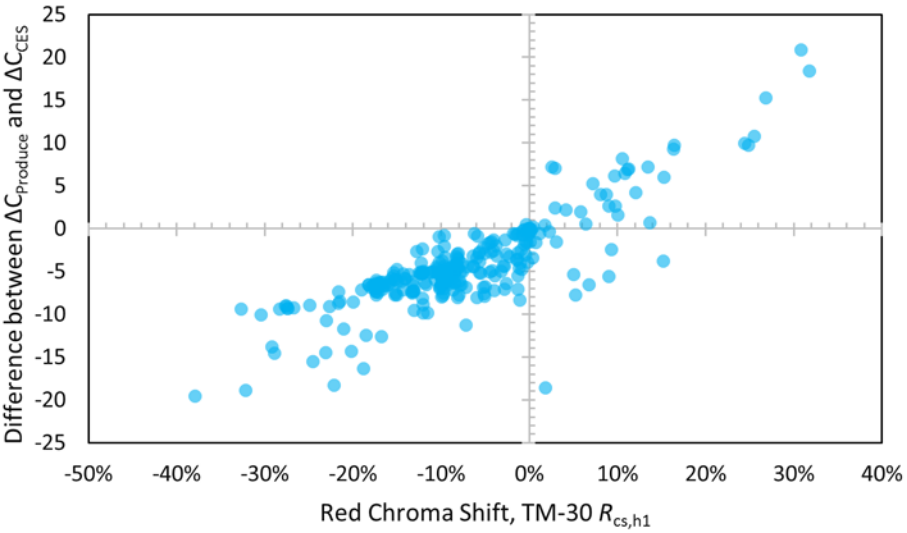
Again, the takeaway is that the visual stimulus of a given object set may vary substantially from what is calculated by TM-30, depending on the exact source. For many sources, the difference is small, but for some highly-structured SPDs, the difference can be quite large. The difference depends on the exact features of the sample/object sets, such as color space and wavelength space uniformity, in combination with the features of the SPD.

3.3 Average Chroma Shift Results

In contrast with average fidelity versus TM-30 R_f and average gamut area versus TM-30 R_g, there are some differences in data patterns among the three object sets when evaluating ΔC. As with the other comparisons, the experimental object set is most similar to the standard TM-30 sample set when determining average chroma shift of the 344 SPDs, with an average difference of -0.8 points and a range of 19.0 points (r^2 = 0.99). Also similarly, standard TM-30 calculations provide a worse prediction of ΔC for the produce, with an average differences of -3.8 points and a range of 40.4 points (r^2 = 0.93). The magnitude of the difference was similar for the CCC dataset, with an average of 3.4 points and a range of 32.2 points (r^2 = 0.88), but the typical direction of the difference was opposite. These relationships are documented in **Figure 12**.

Figure 13 plots the difference between ΔC for each object set and ΔC for the TM-30 CES (ΔC_{CES}) against TM-30 hue angle bin 1 chroma shift ($R_{cs,h1}$). For the experiment and produce object sets, when the SPD increases red saturation (that is, increases $R_{cs,h1}$), the custom ΔC value tends to be higher than the ΔC_{CES}, with the opposite true when the SPD decreases red saturation. There is less of a relationship for the CCC dataset, ostensibly because most of the samples are highly saturated and therefore the custom measure is (almost) always greater than the standard measure. These results are related to the effect of chroma level on color shifts, which was documented in **Figure 6**.





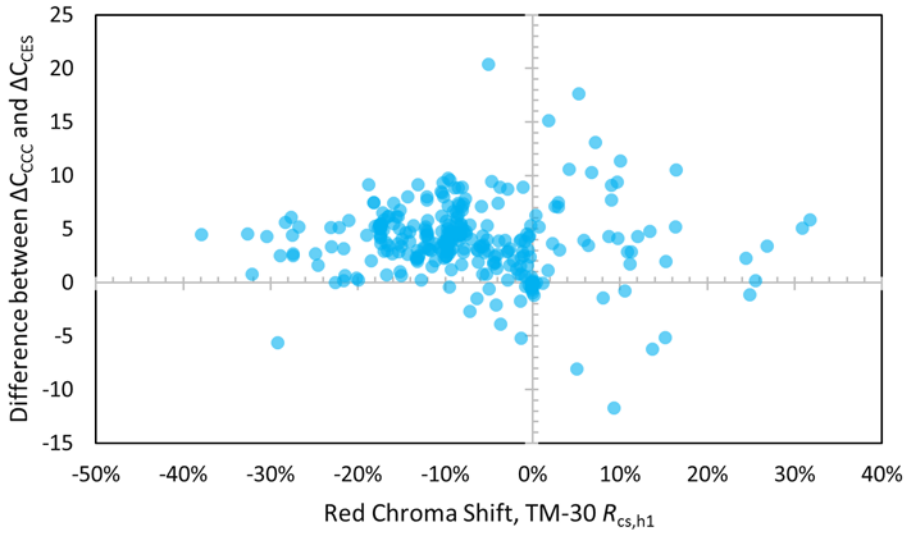


Figure 13. ΔC versus TM-30 $R_{ch,h1}$ for the three object sets. There is a clear pattern for the experiment and produce sets, but not the CCC set.

4 Discussion

Color rendition measures attempt to simulate the effect of light on the color appearance of objects in a real environment, substituting a standardized set of objects for whatever might be present in a given space. This is a tradeoff between utility and accuracy, where color rendition measures have varying levels of applicability to any given environment, depending how similar the objects' spectral reflectance functions are to the set of standardized samples used in the calculation. This makes the sample color set a critical part of any color rendition measure. It also means that using a set of objects in psychophysical experiments to derive meaning for a given color rendition measure must be done with great care.

To date, standardized sets of colors, such as the TM-30 CES or the CIE CRI Test Color Samples (TCS), have been chosen based on predefined goals. In the case of TM-30, the goals included uniformity in color space and wavelength space, as well as the use of a relatively small number of spectral reflectance functions of a variety of real objects [David and others 2015]. In the case of CIE R_a, the TCS were chosen to have moderate chroma and approximately equal spacing in hue space. Other selection criteria could also be justified, as there is no known data that captures the distribution of spectral reflectance functions in real spaces—either on average or for any particular type of space.

In practice, an ideal set of color samples would be one that matches the objects in the space in question. Absent technology to measure (or know) the spectral reflectance functions of the objects present in a given space, calculating space-specific color rendition measures will remain practical for only the very most critical of applications, such as the Sistine Chapel [Schanda and others 2016]. Alternatively, a good set of color samples might be the one that best represents the average of all interior environments, with a sufficient number of samples to predict a complete set of visually different colors with minimal error. However, it would be essentially impossible to quantify an "average" interior environment, nor would that environment be representative of any specific environment. The end result is that any sample set must be chosen using a justifiable set of criteria.

Because the TM-30 CES, like any other samples, do not perfectly represent any given illuminated space, accommodations must be made when interpreting the results and predicting how a given space will appear. If a somewhat detailed color profile of a space is known, a TM-30 user can more carefully evaluate the hue-bin specific sub-indices to better understand how the most important objects will be rendered. One may also just use color rendition measures as a first pass, with visual evaluation ensuring satisfaction. Critically, it is important to recognize that small differences in average values (e.g., R_f, R_g), especially for disparate source types, are likely to be overshadowed by the mismatch induced by the difference between the standardized sample set and the objects in a given space, not to mention the effects of gamut shape [Royer and others 2016; Wei and others 2016a; 2016b]. Two sources with the same average color fidelity and the same average gamut area may make objects look different. In short, standardized color rendition measures are valuable, but the limitations should be understood.

4.1 Implications for Experiments

Perhaps more germane to the analysis presented here is the practical relevance of the described differences to choosing experimental object sets and generalizing research findings, especially when the goal is to correlate perceptual meaning, such as preference or acceptability, with numerical quantifications. Again, this article is not focused on the merits of any given sample set for use in standardized calculations.

As shown in this analysis, there can be substantial differences between viewed color differences for a given object set—such as the CCC, which has been used in numerous experiments—and the quantification of the

light source using TM-30 (or any other measure). If the goal is to understand acceptability at a given fidelity value, for example, the mismatch between calculated and viewed color distortions can cause substantial error, which may be further amplified later when an established value is applied to a real space that is unlike the evaluated colors or the standardized test colors. In other words, a researcher using only the CCC to establish a preference model based on the TM-30 measures may determine that an R_f value of 90 is necessary for a given perception. However, the color shifts of the CCC are typically much larger than predicted by TM-30—and likely larger than any space that is not filled with highly saturated colors—meaning 90 may be an artificially high threshold. These types of erroneous findings may stem from overall differences in the chroma of the evaluated and standardized samples, which tend to affect all SPDs similarly (that is, changing the slope of the lines in **Figures 3, 10, 11, and 12**). They may also stem from differences in the samples' coverage of color space or wavelength space, which are more likely to result in selective error depending on the features of a given SPD (that is, the range of values presented in **Tables 1 and 2**).

When designing color perception experiments, researchers must decide on all contextual factors (culture, application/objects, illuminance, and chromaticity/chromatic adaptation), as well as the source(s) under which the judgment is made, sorting through innumerable options weighed against the constraints of limited resources. This has led to different tradeoffs being made in the numerous experiments in this field. While the tradeoffs are necessary, it is important to understand the resulting limitations. One of the most difficult challenges is deciding on the objects that will be viewed.

Based on the analysis presented here, the selection of color samples for evaluation should be paid careful attention. As a compromise to the ideal of space-specific color rendition measures, a series of application-specific experiments, using objects typical of the space, could be performed in order to develop criteria or models for those applications. This would require substantial work, and may even be too detailed, given the wide variety of spaces within a given application and the wide variety of personal preferences regarding color appearance. But, it would likely produce better results than the current use of a one-size-fits-all criterion for all applications. In the meantime, choosing experimental objects with properties similar to those of the measure being evaluated, such as the TM-30 CES, is a reasonable approach and a promising method for establishing performance criteria or models, because it eliminates one level of discord when the criteria or models are then applied to real applications.

Other important considerations for the selection of experimental object sets include:

1. **The familiarity of the objects**. Preferred color distortions have been linked to how we remember colors [Smet and Hanselaer 2015a; Smet and others 2011; Smet and Hanselaer 2015b; Smet and others 2014; Smet and others 2010], which is typically more saturated than reality [Elliot and Maier 2014]. Whether none, some, or all of the objects presented to an experiment participant are familiar may influence his or her perception of various distortions. No study has specifically examined this variable using *a priori* hypotheses.
2. **The area subtended by different objects.** With IES TM-30 and CIE R_a, each sample is given equal weight. In an experiment endeavoring to use real objects, finding representative items with equal size is not realistic. As such, different objects—representing different hues, chromas, and lightnesses—will subtend different areas of the visual field. Objects subtending larger areas may prove more influential, although this has not been specifically examined. It is also possible that color psychology or object familiarity could overwhelm any effect due to object area.
3. **The placement of objects within the space.** As with the area subtended by the objects, their placement in a room or booth may also influence experimental outcomes. Even the choice of a

room or booth is important, although it has not been explicitly studied, ostensibly because the objects used in a room would not fit in a booth. One might theorize that booths are further removed from the architectural context in which measures of color rendition are intended to be used, and therefore a weaker experimental tool, but this has not been proven.

4. **The total quantity of objects**. Should the number of objects be limited to ensure that each participant keys in on the same information, or should a greater number of objects be presented to have participants consider a more complex stimulus, like they might in a real situation? This is another important consideration for which the implications have not been formally investigated.

Alas, even if the spectral reflectance functions of an experimental object set can be made similar to those of the standardized sample set for the measure that is being evaluated, these other factors could influence the resulting perceptual correlates and any derived specification criteria. As with object sets, there can be no perfect experimental space, because all real spaces will vary in all four of the listed considerations. The only fail-proof method for the specification process is visual evaluation of a light source in the intended space. Given the impracticality of this approach, it is likely that measures of color rendition will always have a place in the lighting industry; it is paramount that the limitations of the measures themselves—and any subsequently derived criteria—are understood. A consensus on specification criteria is only reasonable after multiple experiments that vary the above considerations have been conducted.

5 Conclusions

This paper examines the relationship between a viewed stimulus, which is an interaction between a spectral power distribution and a set of spectral reflectance functions for a group of objects, and the rated stimulus, which is an interaction between the same spectral power distribution and a *standardized* set of spectral reflectance functions. Because of the differences between these two conditions, which can be quite substantial, specifiers must be cognizant of how color rendition measures apply to architectural spaces and researchers must be careful in selecting the objects presented to experiment participants.

A standardized set of color samples used for rating the *properties of a light source*, independent of application, is essential for commerce, and provides a first pass indication of how a source will render objects. However, it is critical for users to understand the differences between the standardized sample set and any given application when interpreting the values of TM-30 measures—or those of any other color rendition measure. Use of TM-30 sub-indices for specific hues may be warranted. In the most extreme cases, custom measures can be calculated based on the objects in the space under considerations.

For researchers, the results of this paper show the substantial effect that the chosen object set may have on the findings of an experiment. If developing application-specific criteria or models is the goal, objects' color characteristics should represent that application as best as possible. If developing generalized criteria or models is the goal, using an experimental object set with color characteristics similar to those of the measure in question is advised. The spectral reflectance functions of the objects is just one of several factors related to experimental objects that can influence the accuracy of resulting specification criteria.

References

CIE. 1995. 13.3: Method of measuring and specifying colour rendering properties of light sources, 3rd Ed. Vienna (Austria): Commission Internationale de l'Eclairage. p. 16.

Cohen J. 1964. Dependency of the spectral reflectance curves of the Munsell color chips. Psychonomic Science 1(1):369-370.

David A. 2013. Color Fidelity of Light Sources Evaluated over Large Sets of Reflectance Samples. Leukos 10(2):59-75.

David A, Fini PT, Houser KW, Ohno Y, Royer MP, Smet KA, Wei M, Whitehead L. 2015. Development of the IES method for evaluating the color rendition of light sources. Opt Express 23(12):15888-906.

Davis W, Ohno Y. 2010. Color quality scale. Optical Engineering 49(3):033602.

Elliot AJ, Maier MA. 2014. Color psychology: effects of perceiving color on psychological functioning in humans. Annu Rev Psychol 65:95-120.

Fairchild MD. 2013. Color Appearance Models. Wiley. Chichester, United Kingdom 472 p.

Houser KW, Tiller DK, Hu X. 2005. Tuning the Fluorescent Spectrum for the Trichromatic Visual Response: A Pilot Study. Leukos 1(1):7-23.

IES IES. 2015. IES-TM-30-15 Method for Evaluating Light Source Color Rendition. New York, NY: The Illuminating Engineering Society of North America. p. 26.

Islam MS, Dangol R, Hyvarinen M, Bhusal P, Puolakka M, Halonen L. 2013. User preferences for LED lighting in terms of light spectrum. Lighting Research & Technology 45(6):641-665.

Jost-Boissard S, Avouac P, Fontoynont M. 2014. Assessing the colour quality of LED sources: Naturalness, attractiveness, colourfulness and colour difference. Lighting Research and Technology 47(7):769-794.

Jost-Boissard S, Fontoynont M, Blanc-Gonnet J. 2009. Perceived lighting quality of LED sources for the presentation of fruit and vegetables. Journal of Modern Optics 56(13):1420-1432.

Lin Y, Wei M, Smet K, Tsukitani A, Bodrogi P, Khanh T. 2015. Colour preference varies with lighting application. Lighting Research and Technology. Online before print. DOI: 10.1177/1477153515611458

Liu A, Tuzikas A, Zukauskas A, Vaicekauskas R, Vitta P, Shur M. 2013. Cultural Preferences to Color Quality of Illumination of Different Artwork Objects Revealed by a Color Rendition Engine. IEEE Photonics Journal 5(4):6801010.

Ohno Y, Fein G, Miller C. 2015. Vision experiment on chroma saturation for color quality preference. In:28th CIE Session; 2015 Jun 28 – Jul 4; Manchester, UK. Vienna (Austria): Commission Internationale de l'Eclairage. Vol 1. 2124 p.Quellman EM, Boyce PR. 2002. The Light Source Color Preferences of People of Different Skin Tones. Journal of the Illuminating Engineering Society 31(1):109-118.

Rea MS, Freyssinier-Nova JP. 2008. Color rendering: A tale of two metrics. Color Research and Application 33(3):192-202.

Rea MS, Freyssinier JP. 2010. Color Rendering: Beyond Pride and Prejudice. Color Research and Application 35(6):401-409.

Royer M, Wilkerson A, Wei M, Houser K, Davis R. 2016. Human perceptions of colour rendition vary with average fidelity, average gamut, and gamut shape. Lighting Research and Technology. Online before print. DOI: 10.1177/1477153516663615.

Royer MP. 2016. What is the Reference? An Examination of Alternatives to the Reference Sources Used in IES TM-30-15 Leukos. Online before print. DOI: 10.1080/15502724.2016.1255146

Sanders C. 1959. Color preferences for natural objects. Illuminating Engineering 47(1):452-456.

Schanda J, Csuti P, Szabo F. 2016. A New Concept of Color Fidelity for Museum Lighting: Based on an Experiment in the Sistine Chapel. Leukos 12(1-2):71-77.

Schanda J, Sandor N. 2003. Colour rendering, Past – Present – Future. International Lighting and Colour Conference. 2-5 Nov. Cape Town, South Africa. p. 76-85.

Smet K, Hanselaer P. 2015a. Memory and preferred colours and the colour rendition of white light sources. Lighting Research and Technology 48(4):393-411.

Smet K, Ryckaert WR, Pointer MR, Deconinck G, Hanselaer P. 2011. Colour Appearance Rating of Familiar Real Objects. Color Research and Application 36(3):192-200.

Smet KA, Hanselaer P. 2015b. Impact of cross-regional differences on color rendition evaluation of white light sources. Opt Express 23(23):30216-26.

Smet KA, Lin Y, Nagy BV, Nemeth Z, Duque-Chica GL, Quintero JM, Chen HS, Luo RM, Safi M, Hanselaer P. 2014. Cross-cultural variation of memory colors of familiar objects. Opt Express 22(26):32308-28.

Smet KA, Ryckaert WR, Pointer MR, Deconinck G, Hanselaer P. 2010. Memory colours and colour quality evaluation of conventional and solid-state lamps. Opt Express 18(25):26229-44.

Smet KAG, David A, Whitehead L. 2015. Why Color Space Uniformity and Sample Set Spectral Uniformity Are Essential for Color Rendering Measures. Leukos 12(1-2):39-50.

Spaulding JM. 2012. Evaluation of Desirability Assessment Techniques for Tunable Solid State Lighting Applications. Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 22-26 Oct. p. 643-647.

Szabó F, Csuti P, Schanda J. 2009. Color preference under different illuminants—new approach of light source coloour quality. Light and Lighting Conference with Special Emphasis on LEDs and Solid State Lighting; 27-29 May; Budapest, Hungary: CIE. p. PWDAS-43.

Szabo F, Keri R, Schanda J, Csuti P, Mihalyko-Orban E. 2014. A study of preferred colour rendering of light sources: Home lighting. Lighting Research and Technology 48(2):103-125.

Teunissen C, van der Heijden F, Poort S, de Beer E. 2016. Characterising user preference for white LED light sources with CIE colour rendering index combined with a relative gamut area index. Lighting Research and Technology.

Thornton WA. 1974. A Validation of the Color-Preference Index. Journal of the Illuminating Engineering Society 4(1):48-52.

Veitch JA, Tiller DK, Pasini I, Arsenault CD, Jaekel RR, Svec JM. 2002. The effects of fluorescent lighting filters on skin appearance and visual performance. Journal of the Illuminating Engineering Society 31(1):40-60.

Veitch JA, Whitehead LA, Mossman M, Pilditch TD. 2014. Chromaticity-Matched but Spectrally Different Light Source Effects on Simple and Complex Color Judgments. Color Research and Application 39(3):263-274.

Wei M, Houser K, David A, Krames M. 2014a. Perceptual responses to LED illumination with colour rendering indices of 85 and 97. Lighting Research and Technology 47(7):810-827.

Wei M, Houser K, David A, Krames M. 2016a. Color gamut size and shape influence colour preference. Lighting Research & Technology Accepted for publication May 4, 2016.

Wei M, Houser K, David A, Krames M. 2016b. Effect of gamut shape on color preference. CIE 2016 Lighting Quality and Energy Efficiency. Melbourne, Australia. p. 32-41.

Wei MC, Houser KW, Allen GR, Beers WW. 2014b. Color Preference under LEDs with Diminished Yellow Emission. Leukos 10(3):119-131.

Xu W, Wei M, Smet K, Lin Y. 2016. The prediction of perceived colour differences by colour fidelity metrics. Lighting Research and Technology.

Zukauskas A, Vaicekauskas R, Vitta P, Tuzikas A, Petrulis A, Shur M. 2012. Color rendition engine. Opt Express 20(5):5356-67.