https://doi.org/10.2352/issn.2169-2629.2021.29.160

A Digital Test Chart for Visual Assessment of Color Appearance Scales

Mark D. Fairchild; Program of Color Science / Munsell Color Science Laboratory, Rochester Institute of Technology, Rochester, New York USA

Abstract

A digital color appearance test chart, akin to a ColorChecker® Chart for human perception, was developed and evaluated both perceptually and computationally. The chart allows an observer to adjust the appearance of a limited number of color patches to allow a quick evaluation of perceived brightness, colorfulness, lightness, saturation, and hue on a display. The resulting data can then be used to compared observed results with the predictions of various color appearance models. Analyses in this paper highlight some known shortcomings of CIELAB, CIECAM02, and CAM16. Differences between CIECAM02 and CAM16 are also highlighted. This paper does not provide new psychophysical data for model testing, it simply describes a technique to generate such data and a computational comparison of models.

Introduction

The X-Rite ColorChecker® chart and its derivatives have a long and important history in the field of color imaging science.[1] So much so that readers of this paper are likely to be able to accurately visualize the color appearance of its 24 patches from memory. While color appearance models such as CIECAM02, CAM16, and others have also become important tools in color imaging and color perception, it is more rare for observers or users to personally evaluate the scales incorporated in such models systematically to get a sense of their performance relative to individual perceptions. The goal of this work has been to develop a relatively simple digital test chart that can be adjusted by an observer on a characterized display to provide scales of color appearance that can then be evaluated with color appearance models of choice. If successful, such a chart could also be used as a psychophysical tool to collect rigorous data on color appearance scales across populations of observers. Such data could then be used to evaluate and improve the mean predictions of color appearance models. This paper describes a prototype of such a chart, a test of its use, and some comparisons of color appearance models using the resulting observational data as points of reference.

Chart Design and Implementation

The goal of the color appearance test chart was to have a collection of stimuli, presented together, to allow an observer to create interval scales of lightness, saturation, and hue as related colors. In addition charts were also made to allow scaling of relative brightness and absolute terminal brightness for unrelated colors. The resulting test images are illustrated in Fig. 1. They were prototyped in Apple Keynote software to allow interactive adjustment of the color patches on a reference display.

Figure 1(a) shows the lightness scaling samples for neutral, red, green, and blue stimuli in the top four rows. The next four rows show saturation scaling samples for approximately unique hues and the bottom row shows hue scaling samples (at equal perceived saturation and lightness). These samples are presented with a white border and a neutral background with relative luminance of 20% of the peak white.

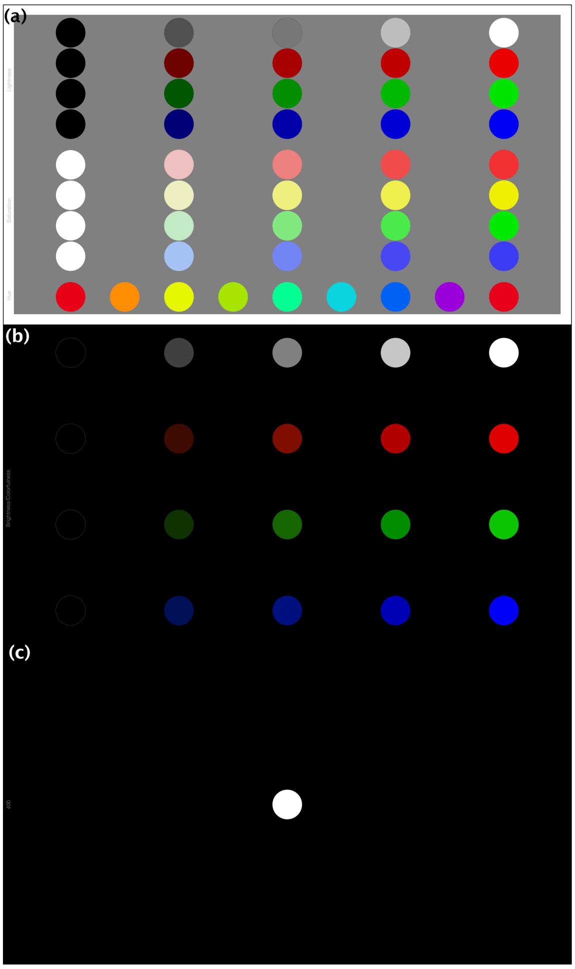

Figure 1. The (a) lightness, saturation and hue, (b) brightness/ colorfulness, and (c) an example of the absolute brightness images in the test chart.

Figure 1(b) shows the brightness/colorfulness scales for essentially unrelated colors that are neutral, red, green, and blue respectively (of constant hue). Lastly, Fig. 1(c) shows one example of a terminal brightness stimulus that was presented at

400nits. Other similar samples were presented at 200, 100, 40, 20, and 10nits respectively.

These charts were presented on a carefully calibrated reference display (Eizo Model CG279X) set up to sRGB primaries and a D65 white point at 400nits and viewed in a darkened room. One observer, the author, interactively adjusted the stimuli on the charts to set to the desired appearances as outlined below. The settings were obtained iteratively over approximately 10 sessions across several weeks until the observer was satisfied that no further adjustments were necessary. All adjustments were made using direct RGB slider controls. These data simply represent an example of the implementation of the test chart and should not be considered as at all similar to mean data across a population of observers. These data allow comparisons of models, but not evaluation of the absolute accuracy of the model predictions.

Lightness scaling was accomplished by starting with the full black and white reference points. The red, green, and blue maximum luminance patches were then adjusted to match the white in apparent brightness/lightness. The remaining samples in the series were adjusted in luminance to be equally spaced in lightness (partition scaling) and match the upper references in hue and saturation. Saturation scaling was accomplished similarly with the four hues adjusted to match in saturation at the maximum possible level and then partition scaling between those anchors and the reference whites. Lastly hue scaling was completed by setting the four patches to the unique/unitary hues (Red, Yellow, Green, Blue, then Red again). Intermediate samples were then set to hues perceptually halfway between the neighboring unique hues. All settings were made at equal apparent lightness and saturation (maximum available).

For the second chart, brightness/colorfulness scaling was completed via partition scaling after the three chromatic bright anchor points were set to equal brightness with the white reference. Lastly, the terminal brightness scales were obtained by magnitude estimation of each luminance level, viewed independently, with the 400nit sample defined as a perceived brightness of 100.

These charts and processes can easily be implemented in custom code as a stand-alone application and managed to any calibrated and characterized display. With this paper, the chart design is being placed in the public domain and anyone is free to replicate it for their own use.

Terminal Brightness

Terminal brightness was defined by Stevens and Stevens [2] as the apparent brightness of unrelated stimuli when the observer is adapted to only the stimulus being evaluated. It represents the maximum perceived brightness for any given adapting luminance.[3] The scaled terminal brightness values as a function of luminance are plotted in Fig. 2 along with predictions by CIECAM02, CAM16, and a terminal brightness function derived from Stevens and Stevens.[2-4] CIECAM02 and CAM16 brightness predictors were computed using a dark background (1%) and adapting luminance equal to the stimulus. The scaled results follow the general sigmoidal functional shape of the Stevens terminal brightness function, but the range of luminance levels explored was small compared to the functional range of human vision and is thus represented by a much smaller output range on the Stevens Function. Likewise, CIECAM02 and CAM16 made essentially identical predictions that also showed less range than the observed results. These results suggest that more adaptation to the stimulus itself is predicted by the models than was observed under the experimental conditions. Also note that the terminal brightness function can be used as a scalar to convert a lightness scale for a given adaptation condition into a proper brightness scale.[3]

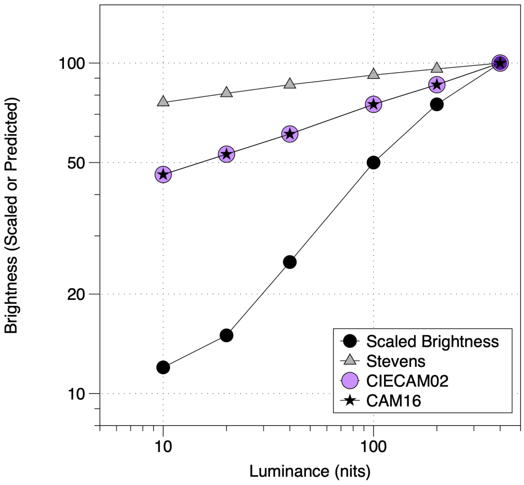

Figure 2. Scaled and predicted brightness as a function of luminance for the six terminal brightness stimuli

Brightness/Colorfulness Scales

Figure 3 illustrates the brightness scaling results from the target represented in Fig. 1(b) along with predictions by CIELAB L* (a), CIECAM02 and CAM16 Q (b), and a G0-relative brightness metric (c).[3] The CIELAB L* values were computed relative to the maximum white stimulus, which was also used as the adapting luminance in CIECAM02 and CAM16 (Yb set to 1.0). The CIELAB L* predictions are linear with the scaled brightness. However, the predicted lightness of the chromatic scales is lower due to the fact that CIELAB L* does not account for the Helmholtz-Kohlrausch Effect (HK Effect; perceived brightness increases with colorfulness/ saturation at constant luminance).[4]. CIECAM02 and CAM16 make similar predictions (though not identical as sometimes suggested) and clearly do not linearly correlate with the scaled results. These models also do not account for the HK Effect. Lastly, the G0-normalized brightness model proposed by Fairchild and Heckaman[3] normalizes the brightness of chromatic stimuli to the brightness of a similar chromaticity that appears to have zero grayness.[5] The G0 normalized brightness prediction does an excellent job at linearly correlating with all the scaled results and the perceptual equality of brightness across the four scales.

Figure 4 Illustrates the prediction of perceived colorfulness for the three chromatic scales using CIECAM02 and CAM16. The predicted scales were normalized to the maximum since the magnitude of the upper anchor is arbitrary (though equal for all three hues). The models make similar, though slightly different, predictions. While they do a reasonable job of predicting the relative colorfulness of the three scales (to one another), the model predictions are not linear with the scaled results. These scales were also set with constant hue. It should be noted that all three models performed well at predicting constant hue for these samples, but that the CIECAM02 and CAM16 hue quadrature predictions varied from each other by up to 15 units with the largest discrepancy in the blue.

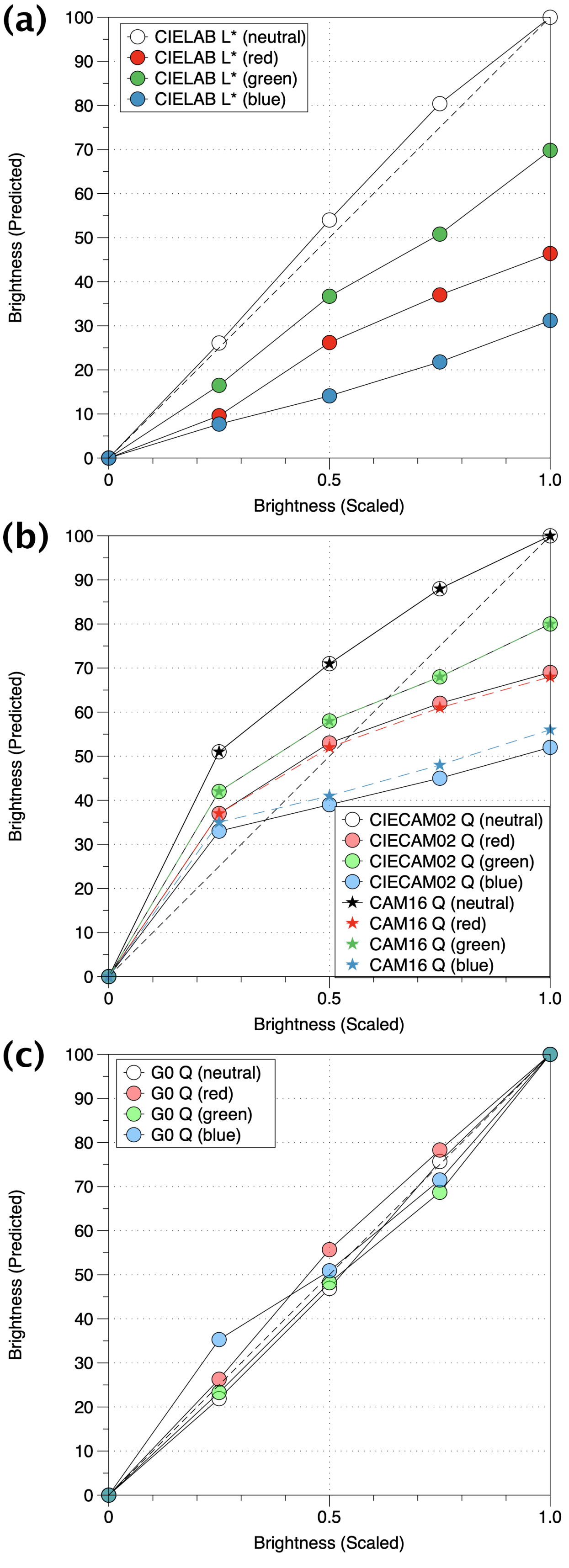

Figure 3. Scaled brightness as predicted by (a) CIELAB L, (b) CIECAM02 and CAM16, and (c) a G0-normalized brightness predictor.*

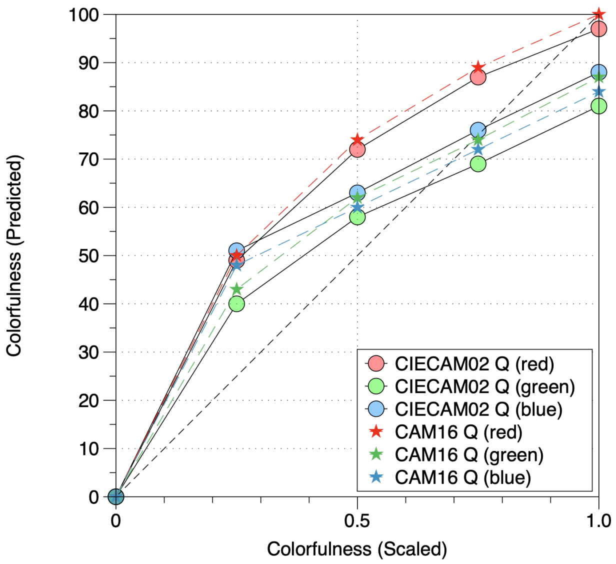

Figure 4. Scaled Colorfulness as predicted by CIECAM02 and CAM16.

Lightness

Figure 5 (analogous to Fig. 3) illustrates the lightness scaling results from the target represented in Fig. 1(a) along with predictions by CIELAB L* (a), CIECAM02 and CAM16 J (b), and a G0-relative lightness metric (c).[3] The CIELAB L* values were computed relative to the maximum white stimulus, which was also used as the adapting luminance in CIECAM02 and CAM16 (Yb set to 20). The CIELAB L* predictions are slightly nonlinear with the scaled lightness. However, the predicted lightness of the chromatic scales is lower due to the fact that CIELAB L* does not account for the HK Effect.[4]. CIECAM02 and CAM16 make similar predictions to one another and more linearly correlate with the scaled results then L*. These models also do not account for the HK Effect. Lastly, the G0-normalized lightness model proposed by Fairchild and Heckaman[3] normalizes the lightness of chromatic stimuli to the lightness of a similar chromaticity that appears to have zero grayness.[5] The G0 normalized lightness prediction does an reasonable job at linearly correlating with all the scaled results and the perceptual equality of brightness across the four scales. These scales were also set with constant hue. It should be noted again that all three models performed well at predicting constant hue for these samples, but that the CIECAM02 and CAM16 hue quadrature predictions varied from each other by up to 15 units with the largest discrepancy in the blue.

Saturation

Figure 6 illustrates the saturation scaling results from the target represented in Fig. 1(a) along with predictions by CIELAB C*/L* (a), CIECAM02 and CAM16 s (b), and excitation purity, Pe (c).[3] The CIELAB C*/L* values were computed relative to the maximum white stimulus, which was also used as the adapting luminance in CIECAM02 and CAM16 (Yb set to 20). The CIELAB C*/L* is recognized as an approximate saturation scale despite not being recommended by the CIE as such.[4] Its predictions are slightly nonlinear with the scaled saturation and vary significantly from hue to hue (illustrating the known non-uniformity of CIELAB in the chromatic dimensions). CIECAM02 and CAM16 make significantly different predictions of saturation despite showing

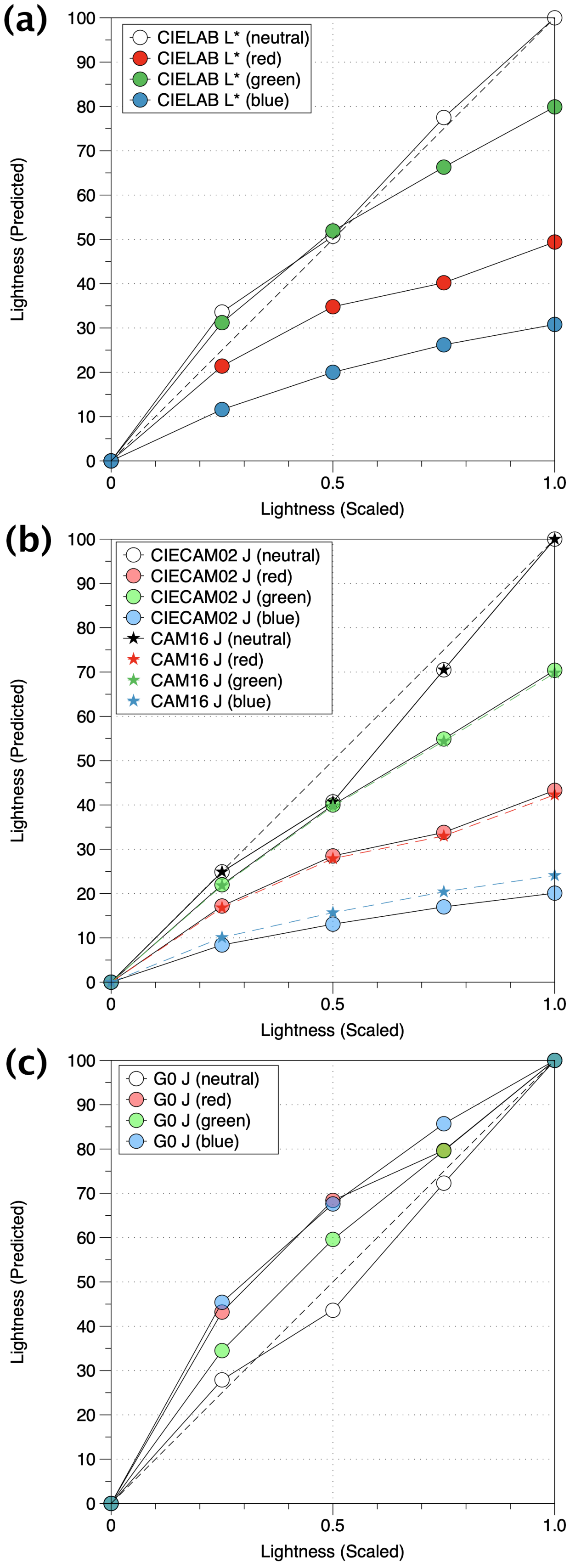

Figure 5. Scaled lightness as predicted by (a) CIELAB L, (b) CIECAM02 and CAM16, and (c) a G0-normalized lightness predictor.*

similar nonlinear trends with the scaled data. Lastly, excitation purity proves to be a reasonable model of perceived saturation as was proposed by Fairchild and Heckaman[3,4]. This very simple model to predict saturation should be explored further since advanced color appearance models like CIECAM02 and CAM16 unnecessarily complicate the prediction of saturation and do not perform well. Excitation purity also performed exceptionally well at predicting the equated perceptual saturation across the four hues.

Figure 7(a) examines the hue uniformity of the predictions from CIELAB, CIECAM02, and CAM16. The models do a reasonable job at making consistent hue predictions for the various saturation levels that were scaled to be equal in hue. However all three models show significant hue non uniformity (largest for CIELAB in the blue hue) and there are small, but significant differences in the hue predictions between CIECAM02 and CAM16. These are caused by a rotation of the chromatic axes in CAM16 (relative to CIECAM02) without a compensatory adjustment in the definition of the unique hues. Hue predictions in CAM16 can be significantly different from those made by CIECAM02. This point is illustrated in Fig. 7(b) where the differences are most significant (up to 10 units in hue quadrature) for the blue and green hues.

Hue

The hue quadrature predictions for CIECAM02, CAM16, and CIELAB (based on the NCS unique hue angles) are illustrated in Fig 8(a) as a function of the scaled hues. All three models show significant error in the prediction of unique yellow (H=100) unique blue (H=300) for this observer and viewing condition. This is not surprising as it has been shown that there are significant individual variations in unique hue settings.[6,7] All three models make reasonable predictions and could be improved with unique hue anchors corresponding to individual observers rather than NCS-based average responses. It is worth noting that the models sometimes make predictions significantly different from one another (again particularly for yellow and blue). Most important among these are differences as large as 10 H units (effectively 10 percent hue composition) between CIECAM02 and CAM16. Figure 8(b) more closely examines those differences between CIECAM02 and CAM16 as well as their discrepancy from this observer's results.

Differences Between CIECAM02 and CAM16

CAM16 is likely to be published as a CIE model soon. It was derived to have essentially identical formulation to CIECAM02 with the exception that the cone responses were changed to avoid negative models and to match the cone responses between the adaptation transform and the color space construction. This leads some to conclude that the models are essentially identical. While their predictions are generally close to one another, they are not identical and in some cases quite dissimilar. Two cases where these differences are most prevalent are in colorfulness (and thus saturation and chroma) of highly chromatic stimuli (such as near the gamut boundaries of wide-color-gamut displays) and in the hue predictions. Users transitioning from CIECAM02 to CAM16 workflows should be aware of these differences.

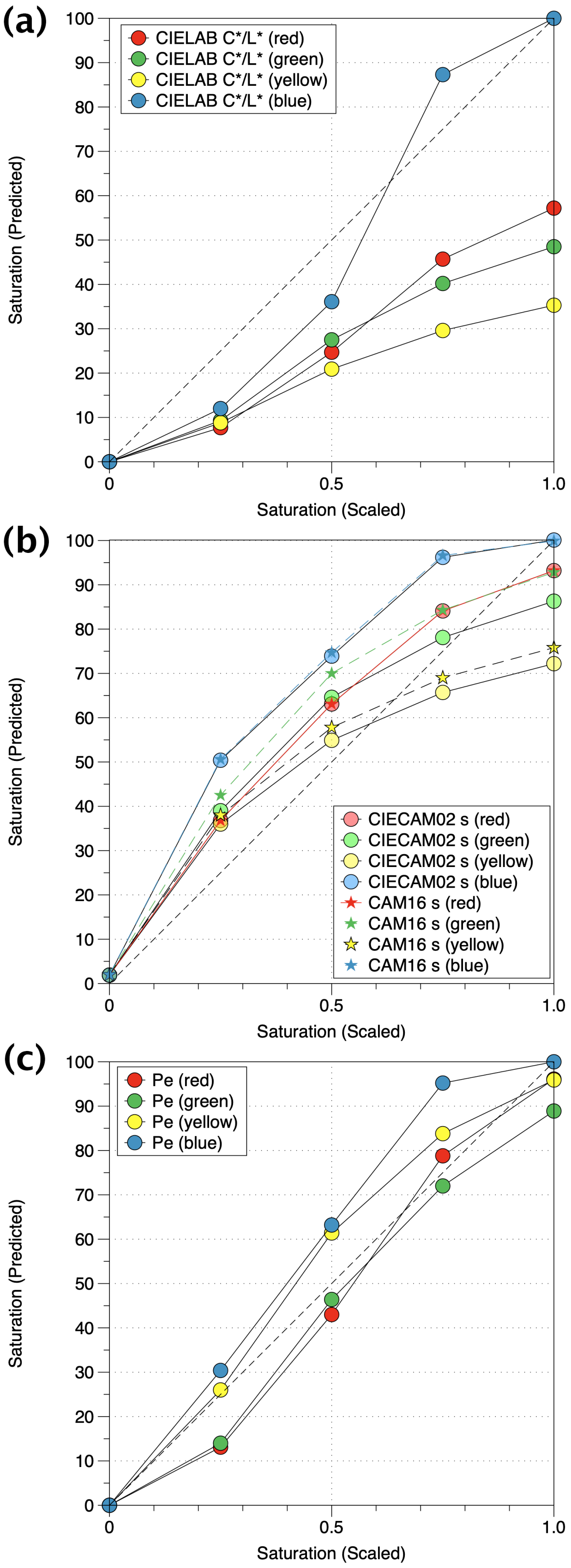

Figure 6. Scaled saturation as predicted by (a) CIELAB C/L*, (b) CIECAM02 and CAM16, and (c) excitation purity, Pe.*

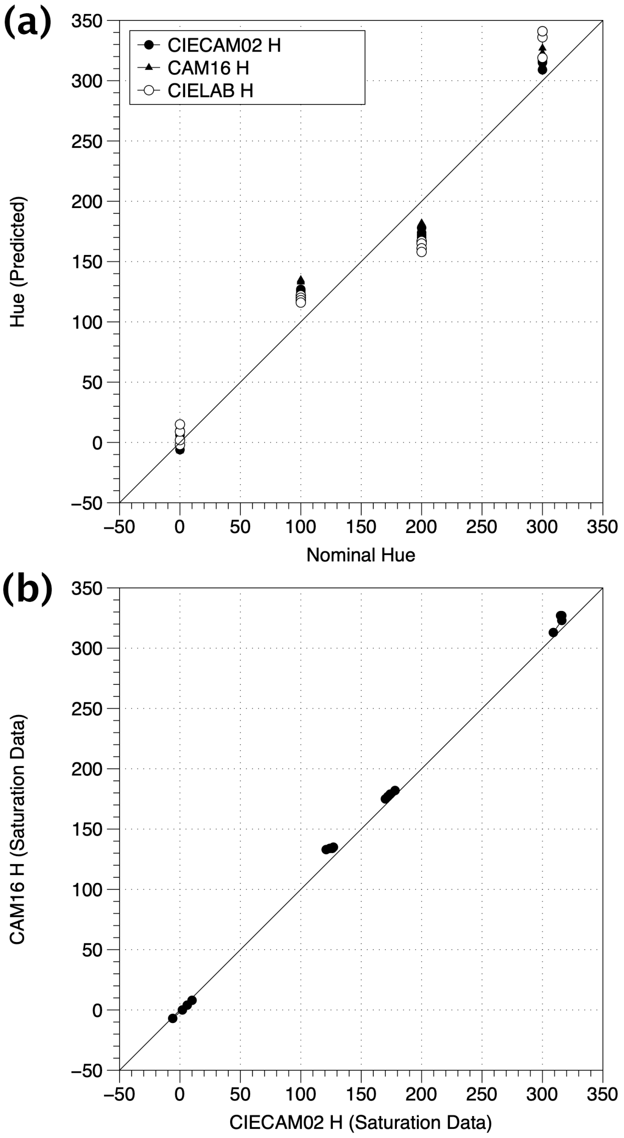

Figure 7. (a) Predicted hue quadrature of the saturation scales (which were constant hue) and (b) comparison of CAM16 hue quadrature with CIECAM02 hue quadrature for these results.

The differences occur because the color space is built on a different set of cone responses, which effectively rotates the hue angle of the appearance correlates. Despite the rotated hue angle, relative to CIECAM02, CAM16 utilizes the same hue angle definitions for the unique hues. Thus hue quadrature calculations can be significantly different when comparing the two models. Additionally, the change in cone responses that removes the negative values introduces nonlinear differences in the colorfulness and constant hue predictions between the two models. Care is suggested in moving between the two models.

Two other issues with CIECAM02 and CAM16 have been recently observed as part of this work and other research. These are that, in both models, saturation is proportional to the square root of colorfulness and brightness is proportional to the square root of lightness. The origination of these square root relationships is unclear, but both are theoretically flawed and not supported by psychophysical results. Application of the models in new situations such as HDR imaging, AR, and VR are revealing such inconsistencies that were previously unnoticed. Remediation of these issues should be considered for future color appearance models.

Despite these differences, CAM16 is a more stable and mathematically well-behaved model, particularly for highly saturated stimuli, and the high visual uncertainty in existing color appearance data suggests it performs similarly to CIECAM02 overall. Also note that the HK effect is not properly predicted by either model while this paper illustrates the potential benefit of a new definition of lightness perception relative to the G0 boundary rather than simply relative to diffuse white.[3]

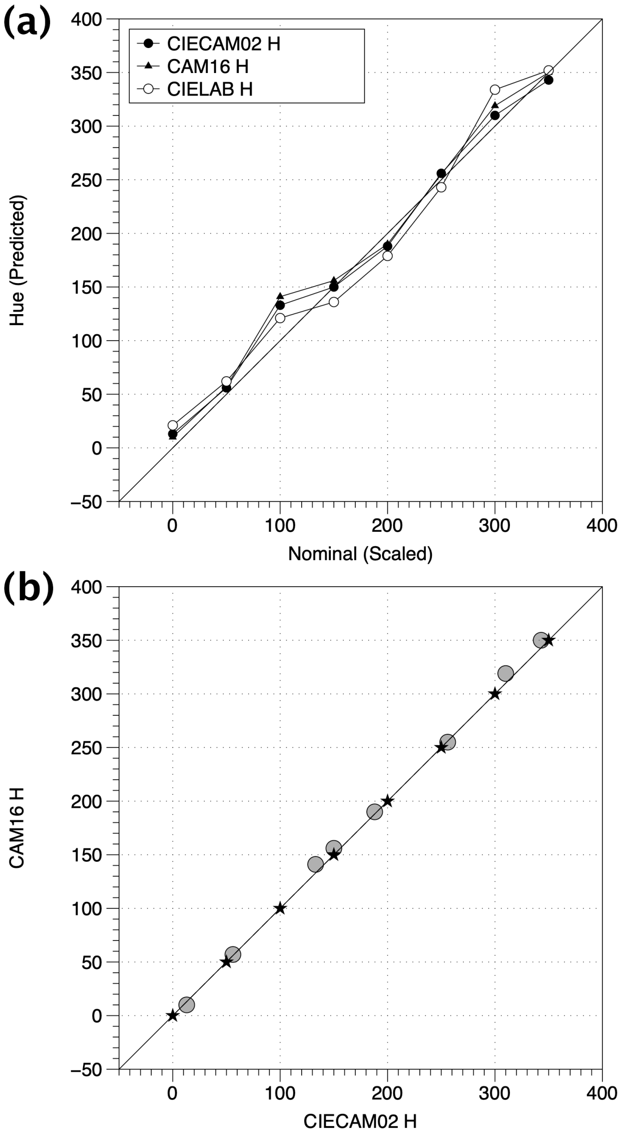

Figure 8. (a) Hue quadrature for the hue scaling results as predicted by CIECAM02, CAM16, and CIELAB and (b) a comparison of CAM16 and CIECAM02 hue quadrature (circles) along with the scaled results (stars).

Conclusions

This paper introduces a set of test charts and procedures to allow individual observers to quickly generate color appearance scales that can be used to compare with the perditions of various models. It is a more robust technique for evaluating the individual effectiveness of a color appearance model than generating equip-spaced samples with the model and trying to visually assess the output (such patterns introduce a confirmation bias that is not present in self-generated stimuli). However, such a chart cannot be used directly to evaluate the overall performance of a color appearance model since it does not represent the average response of a population and explores a rather limited range of viewing conditions.

To do so, the charts could be adapted to proper psychophysical experiments by generating them for large groups of observers and across various levels of luminance, background, surround, and adapting chromaticity. This might be a technique worth future exploration to help eliminate the relative drought of color appearance data that can be used to test and formulate models. Currently the LUTCHI data set[9] together with a few other experimental results is the standard by which models are formulated and tested. While that data set is excellent, it remains limited in scope and range of stimuli. There is clearly room for more data, over wider ranges of viewing conditions, and with more observers.

It is hoped that this small exercise in creating appearance scales and using them to compare appearance models helps motivate even more research and is useful to highlight some of the differences in modern, and not-so-modern but effective, color appearance predictors.

References

[1] C.S. McCamy, H. Marcus, J.G. Davidson, A color-rendition chart, J. Applied Photo. Eng. 2, 95-99 (1976).

[2] J.C. Stevens and S.S. Stevens, Brightness functions: Effects of adaptation, J. Opt. Soc. Am. 53, 375-385 (1963).

[3] M.D. Fairchild and R.L. Heckaman, Deriving appearance scales, IS&T/SID 20th Color and Imaging Conference, Los Angeles, 281-286 (2012).

[4] M.D. Fairchild, Color Appearance Models, 3rd Ed. Wiley-IS&T Series in Imaging Science and Technology, Chichester, UK (2013).

[5] R.M. Evans, The Perception of Color, John Wiley (1974).

[6] R. Shamey, M.G. Sedito, R.G. Kuehni, Comparison of unique hue stimuli determined by two different methods using Munsell color chips, Color Res. App. 35, 419–424 (2010).

[7] M.A. Webster, E. Miyahara E, G. Malkoc, V.E. Raker, Variations innormal color vision. II. Unique hues. J Opt Soc Am A 17, 1545–1555 (2000).

[8] C. Li et al., Comprehensive color solutions: CAM16, CAT16, and CAM16UCS, Color Res. Appl. 42, 703-718 (2017).

[9] R.W.G. Hunt and M.R. Luo, Evaluation of a model of colour vision by magnitude scalings: Discussion of collected results, Color Res. Appl. 19, 27-33 (1994).

Author Biography

Mark D. Fairchild is Professor and Founding Head of the Integrated Sciences Academy in RIT's College of Science and Director of the Program of Color Science and Munsell Color Science Laboratory. He received his B.S. and M.S. degrees in Imaging Science from R.I.T. and Ph.D. in Vision Science from the University of Rochester. Mark was presented with the 1995 Bartleson Award by the Colour Group(Great Britain) and the 2002 Macbeth and 2018 Nickerson Awards by the Inter-Society Color Council (ISCC). He is a Fellow of the Society for Imaging Science and Technology (IS&T) and the Optical Society of America (OSA). Mark received the Davies Medal from the Royal Photographic Society, the 2008 IS&T Raymond C. Bowman award, and the 2021 SID Otto Schade Prize for contributions photography, education, and imaging.