CIE TC1-34 Testing Colour-Appearance Models

Mark D. Fairchild, Chair

Munsell Color Science Laboratory

Center for Imaging Science

Rochester Institute of Technology

54 Lomb Memorial Drive

Rochester, New York 14623-5604

Testing Colour-Appearance Models: Guidelines for Coordinated Research

I. Purpose

CIE Technical Committee 1-34, Testing Colour-Appearance Models, was formed to investigate the performance of models based on their ability to predict the colour appearance of surface colours in simple and complex scenes under various illumination conditions. The work of this committee is limited only to the study of the colour appearance of surface colours and does not include the appearance of self-luminous colours, aperture colours, or comparison between different media or modes of appearance. This limitation has been imposed to find the best methods for predicting the colour appearance of objects while leaving complexities such as the cognitive interpretation of coloured stimuli for future work. The visual interpretation of non-object-colour stimuli is complex and ill-defined. The plan of work of TC1-34 includes the analysis of colour appearance models using available data and the publication of these guidelines for coordinated research for additional testing of colour-appearance models. The purpose of these guidelines is to encourage researchers around the world to perform experiments that will provide useful data for evaluating the performance of colour appearance models. The guidelines are not intended to limit researchers to a prescribed experimental design, but rather to provide guidance regarding minimal data and viewing conditions that should be included in all experiments to provide useful intercomparison data. Experiments that reach beyond the minimal requirements outlined in these guidelines are certainly encouraged. These data might one day form the basis of a recommendation by the CIE for a color appearance space with broad applications.

II. Summary

These guidelines provide an overview of the many issues involved in generating visual data that can be used to evaluate the performance of colour-appearance models. The three main sections of these guidelines outline the parameters that must be evaluated and controlled in experimental setups for colour-appearance experiments, suggested psychophysical techniques for gathering the data, and some suggested techniques for data analysis. Experimental parameters addressed include models to be tested, illumination conditions, background and surround conditions, types of stimuli to be used, and issues relating to viewing technique. The psychophysical techniques of magnitude estimation, matching, and direct model testing (paired comparison) are described. Data analysis techniques for the evaluation of colour-appearance scales, corresponding-colours data, and model performance scales are suggested.

III. Experimental Setup

The psychophysical evaluation of colour appearance and colour-appearance models is quite complex. This is why, despite widespread interest in colour appearance, relatively few visual experiments have been carried out. This section outlines important variables in the design of such experiments and makes baseline recommendations on how each variable should be treated.

Models

The most extensive colour-appearance models are those that have been published by Nayatani *et al.* [1] and by Hunt.[2] CIE TC1-13, *Colour-Appearance Analysis*, has prepared a final report summarizing the current status of these two models.[3] These models include parameters to account for changes in many of the variables that influence the colour appearance of stimuli and provide output correlates of brightness and colourfulness in addition to those of lightness, chroma, and hue. The two CIE-recommended color spaces, CIELAB and CIELUV,[4] can also be utilized as simple colour-appearance models. However use of CIELAB and CIELUV can only account for changes in illumination colour and have no mechanisms for predicting appearance changes due to change in luminance level, background, or surround. Richter[5] has developed a model, LABHNU, which is a modification of the CIELAB space. LABHNU also only accounts for changes in illumination colour. Lastly, Fairchild and Berns[6] have formulated a colour-appearance model, RLAB, which is an extension of CIELAB incorporating the ability to account for changes in surround and mode of appearance. CIELAB, CIELUV, LABHNU, and RLAB provide correlates of lightness, chroma, and hue only.

CIE TC1-34 has decided to investigate all of the above models in its work. Investigators wishing to perform new research should include the evaluation of at least the Nayatani *et al.* and Hunt models along with CIELAB as a baseline measure. However, researchers are encouraged to include CIELUV, LABHNU, and RLAB as well as any other colour-appearance models with which they might be familiar. The minimum requirement for a model to be included is that it includes predictors for the relative appearance attributes of lightness, chroma, and hue.

Illumination

The colour and illuminance level of the illumination used for the visual evaluation of colour appearance are of primary importance and must be carefully controlled. Researchers should make spectral power measurements of their light sources and evaluate the stability and warm-up requirements of their viewing booth. The actual spectral power distributions of the light sources under which viewing is performed must be used for all colorimetric calculations. The colour of the light source should be specified using either xy or u'v' chromaticity coordinates and the illumination level of the light source should be specified either as illuminance in lux or as the luminance of a reference white in cd/m^2. The CIE 1931 Standard Colorimetric Observer (2°) should be used for all colorimetry.

The majority of research has been performed using sources that simulate CIE Illuminant D65 and CIE Illuminant A. The change in adaptation between these two illuminants is quite large and therefore represents a severe test of the colour-appearance models. CIE TC1-34 recommends that researchers include at least these two sources in their work in addition to any others that might be of particular interest.

A variety of illuminance levels should also be used. Since colour appearance is a function of illuminance level, experiments that can include only one combination should examine illuminance differences of at least a factor of 10 to best accentuate differences in the appearance models. If the number of illuminance levels to be investigated is limited, they should be kept in the range of 100 - 10,000 lux.

Background and Surround

The background is defined as the area immediately adjacent to the test stimulus. The surround includes the remainder of the visual field. In all experiments, the background and surround should have the same chromaticity as the light source. The minimum recommendation is to use backgrounds and surrounds with reflectance factors of approximately 0.20. Some researchers might want to change the reflectance factor of the background and surround as experimental variables. In these cases, dark backgrounds and/or surrounds with reflectance factors of approximately 0.10 or less and light backgrounds and/or surrounds with reflectance factors of approximately 0.80 or greater are recommended.

Stimuli

Reference stimuli for these experiments should be physical samples (*e.g.*, printed or painted material) since TC1-34 is interested only in surface colours. *Simulated object colours, such as CRT displays, should not be used.* The stimuli should be uniform and subtend a visual angle of not less than 0.5° and not more than 2°. A variety of test colours should be used. CIE TC1-34 would like to encourage the inclusion of test stimuli from (or approximating) the following set of colours to facilitate intercomparison between studies. These colours were specified (D65, 10° observer) by Robertson[7] for coordinated research on colour differences. It might prove useful to have both colour-difference and colour-appearance data for similar object colours.

(1) Y = 30.0, x = 0.314, y = 0.331 (Gray)

(2) Y = 14.1, x = 0.484, y = 0.342 (Red)

(3) Y = 69.3, x = 0.388, y = 0.428 (Yellow)

(4) Y = 24.0, x = 0.248, y = 0.362 (Green)

(5) Y = 8.8, x = 0.219, y = 0.216 (Blue)

Colourists in the surface colour industries often prefer to use the CIE 1964 Supplementary Colorimetric Observer (10°) for their colorimetry and conduct colour quality tasks using large colour samples with viewing angles greater than 2°. Recognizing this, CIE TC1-34 would like to include data generated with such samples in its future analyses and suggests that researchers using larger samples report colorimetric values for both the CIE 1931 Standard Colorimetric Observer and the CIE 1964 Supplementary Standard Colorimetric Observer to facilitate comparison of the various colour appearance models that have been formulated based upon the 1931 observer. This recommendation also applies to those wishing to include the above colors, which were specified for the 10° observer.

Haploscopic Viewing

Ideally, observers will make binocular (*i.e.* both eyes adapted to a single viewing condition) judgements of visual stimuli while completely adapted to a single visual environment. The necessary comparisons across a variety of viewing conditions are made either by memory matching or magnitude estimation. However, some experimental designs require more direct comparisons across viewing conditions. In these cases a haploscopic viewing environment is acceptable. An extensive study has been completed by the Color Science Association of Japan[8] which details problems and reasonable solutions for haploscopic experiments. Investigators might also consider a successive-haploscopic viewing situation.[9] Reference 9 includes a review of experiments that have utilized various adaptations of the haploscopic viewing technique.

IV. Psychophysical Experimental Methods

Previous successful experiments have been based on three general types of design: magnitude estimation, matching, and direct model testing. Experimenters are encouraged to consider one of these techniques for their visual experiments. Each of the experimental designs is described briefly, along with their advantages and disadvantages, below.

Magnitude Estimation

This method has been widely used in studying colour appearance by various researchers such as Bartleson,[10] Pointer,[11] Luo *et al.*[12-15] The experiment is often divided into different phases according to the viewing parameters outlined above. The results provide absolute perceived colour attributes under a set of fully adapted viewing conditions. Hence, the parametric effect can be easily revealed by comparing results from a pair of phases. The experimental setup is also much simpler than that in haploscopic viewing, *e.g.* it does not require any specially designed viewing equipment. A viewing booth is sufficient for the experiment. It is preferred that the experiment should be conducted in a darkened room, *i.e.* with a minimum ambient lighting.

Each colour stimulus is assessed by a panel of observers in terms of colour attributes given in the CIE Lighting Vocabulary[16] such as *lightness*, *chroma*, and *hue*. Reference 12 gives the detailed experimental setup and procedures for a study in which lightness, colourfulness, and hue were scaled. It is necessary to receive a training session prior to the real experiment. To ensure that each observer has a clear concept of these attributes. In the real experiment, some reference samples are required. The lightness scale ranges from 0 (observer's imaginary black) to 100 (a reference white sample). Chroma is an unconstrained scale with 0 as its neutral origin. A chroma reference sample (a mid-range colour) for each phase of the experiment should be provided to assist observers to produce more accurate results. A different reference chroma sample should be used for each experimental phase. An experiment might also need to be conducted to correlate the chroma scales from different phases. No reference colour is required for scaling hue. Hue can be scaled between 0 and 400 by defining four imaginary primaries (red (0), yellow (100), green (200), and blue (300)) which form two pairs of opponent hues. Each hue is described as the percentage of two primaries with the combinations of red-yellow, yellow-green, green-blue, and blue-red. Neutral colors do not have a hue. It is recommended that each stimulus be assessed in two sessions done on different days by the same observer to evaluate reliability. In each session, each stimulus should be assessed twice to evaluate repeatability.

Matching

Matching experiments, particularly asymmetric matching, have a long and valuable history in the colour-appearance literature. Examples of such experiments include the work of Wright,[17] Hunt,[18,19] and Breneman.[20] In such experiments a reference stimulus, presented to one eye, is matched by adjusting the colour of a test stimulus presented to the other eye, typically adapted to a different viewing condition. Another technique involves matching a stimulus to a colour that has been previously memorized. Wright[21] has reviewed this technique and it has been successfully applied to the generation of unambiguous memory colours.[22] While these experiments have been invaluable in producing data for the development and testing of colour-appearance models, they have typically been performed with non-object-colour stimuli to facilitate colour adjustment by the observer. Matching techniques have been implemented with object-colour stimuli. The work of McCann *et al.*[23] and the Colour Science

Association of Japan study[8] provide examples. Precautions should be made to exclude systematic differences between the two eyes. One such technique is to make an asymmetric match followed by a symmetric haploscopic-match such that the test and reference colours and conditions have been evaluated by the same eye.

For object colours, the topic of interest for TC1-34, experiments are implemented by having observers choose a test stimulus that matches the reference stimulus as closely as possible from a collection of potential matching stimuli. For example, the **Munsell Book of Color**, or some other color order system, could be used as a source of potential matching stimuli. In such cases, it is important that observers use a mask to provide a constant surround such that only a single potential matching stimulus is viewed at any moment. Additionally, observers are sometimes given the ability to interpolate between potential matching stimuli to provide more precise responses. Such interpolations can be treated as a magnitude scaling in which the observer reports a fractional distance between the two adjacent (in colour space) stimuli. It is also feasible to produce continuously variable object colour stimuli with appropriate surface attributes using a Maxwell disk device.[24] Once the matches are made, the colorimetric coordinates of the reference and perceptually equivalent test stimuli are recorded along with the experimental parameters of the viewing environments. The results provide data of corresponding-colours that can be used to evaluate colour-appearance models. It is recommended that each match be made in two sessions done on different days by the same observer to evaluate reliability. In each session, each match should be made twice to evaluate repeatability.

Direct Model Testing

Recently, a series of experiments have been undertaken to directly perform visual tests of colour-appearance-model predictions.[9,26,27] In these experiments, a series of reference stimuli are chosen along with a defined set of reference viewing conditions. Then the colour-appearance models to be evaluated are used to predict corresponding colours for one or more sets of test viewing conditions. These results are then used to produce test stimuli that are colour-appearance matches to the reference stimuli according to each of the models. The differences between the models are typically great enough to provide a large range of predicted match colours. The test stimuli are then evaluated using a paired-comparison procedure. In this procedure, each possible pair of test stimuli (*i.e.* pairs of model predictions) are compared. The observers are asked to respond by choosing which sample in each pair is a closer visual match to the appropriate reference stimulus. The comparison can be made either through some form of haploscopic viewing or via memory matching. The data collected are the frequencies that the corresponding colours predicted by each model are chosen as superior to the predictions of each of the other models. Data must be collected for a fairly large number of observers to provide the statistical variability required for the analysis.

Alternatively, direct model tests could be performed using magnitude scaling in which observers assign a scale value to the quality of the colour-appearance match (*e.g.* 0 means total mismatch and 10 means perfect match) or a rank-order procedure in which the observers rank the performance of each model. These techniques will provide useful data, but the scales are not likely to be as precise (in the case of magnitude scaling) or mathematically useful (in the case of ordinal scaling) as an interval scale derived via the method of paired comparisons.

V. Data Analysis

Each type of experiment produces a particular type of results. This section suggests techniques for analyzing the results of the various types of experiments to perform model tests that are consistent with previously published results and most useful to CIE TC1-34.

Colour Appearance Scales

To obtain the mean visual results, the arithmetic mean should be used for lightness and hue (both are fixed-end scales) and the geometric mean should be used for chroma (an open-ended scale). The subsequent data analysis mainly includes the comparisons between two sets of results (visual data and model predictions). The statistical measures such as correlation coefficient (r) and coefficient of variation (CV) can be used to indicate the goodness of fit. Scatter plots should also be provided to reveal systematic trends in the differences.

Each observer's accuracy and repeatability performance should be initially investigated to understand the typical error involved in the experiment. The accuracy performance can be obtained by computing the above measures between each individual's results and the mean visual results, and also between each observer's two repeated assessments.

Corresponding Colours

Data consisting of corresponding-colours generated by asymmetric matching experiments can be used to test colour-appearance models in a rather direct way. However, it is necessary to be able to reverse the calculations of the models. The technique is to use a colour-appearance model to calculate the appropriate appearance parameters (usually lightness, chroma, and hue[25]) of each reference stimulus. These appearance parameters are then used, along with the viewing parameters in the test field to determine the colorimetric coordinates of stimuli in the test field that would match the reference stimuli in the reference field according to the particular colour-appearance model being evaluated. Then the analysis becomes a matter of comparing the corresponding colours predicted by the appearance model with those obtained in the visual experiment. If interpolation is required of the observers, the method used and its accuracy should be carefully specified. All of the techniques described above for colour-appearance scale data can be applied to these data as well. It is useful to specify the visual and model-predicted corresponding colours in an approximately uniform colour space such as CIELAB. This allows analysis in units that are more familiar and perceptually uniform than tristimulus values or chromaticity coordinates Once the data are expressed in such coordinates, the absolute mean and RMS deviations in the L*, a*, b*, C^*_{ab}, and h_{ab} coordinates should be calculated. Overall mean ΔE^*_{ab} values provide a useful summary statistic for model comparison. The absolute deviations should be used to evaluate systematic deviations that could be masked by an RMS metric. Again, it is useful to view scatter plots. The visual uncertainty of the experiment can be evaluated by calculating similar metrics for repeated matches or for sets of symmetric matches (*i.e.* no change in adaptation state).

Model Performance Scales

Preference frequency data collected via paired-comparison direct model tests can be converted into interval scales of the quality of model predictions using Thurstone's Law of Comparative Judgements.[28,29,30] This analysis converts frequency of preference data to an interval scale by converting the probability that one stimulus (model, in this case) is preferred over another stimulus into linear perceptual distances based on the inferred probability distribution of the perceptual magnitude of each stimulus. Typically a normal distribution is assumed and the frequency data are converted to an interval scale by first converting the frequencies to probabilities (proportions), converting the proportions to normal deviates (z-scores), and then averaging the normal deviates for each stimulus to determine the mean scale value. The appropriateness of the normal deviate model for particular data can be evaluated through a χ-squared test comparing average predicted proportions with experimentally determined proportions.[29,30] The significance of differences between models can be evaluated by converting the average scale value differences between the models into a probability that one model will perform better than another. In addition confidence limits can be placed on the scale values by realizing that the resulting scale is proportional to the standard deviations of the visual uncertainty and accounting for the number of observers.

Data collected via magnitude scaling can be directly averaged to provide an overall scale of model performance as described above in the section on colour-appearance scales. The uncertainty in the scale values can be evaluated by analyzing the dispersion of the individual results. However, a precise analysis of confidence in the data cannot be made without further assumptions. Rank order data is best analyzed by averaging the observer-assigned ranks and then re-ranking the results. It is not valid to treat the averaged rank order data as in interval scale. A rank order experiment can not be used to determine how much better one model is than another unless the data are treated as if they were collected in a paired-comparison experiment.

Publication of Data

Too often, the results of visual experiments are published in a form that is of little value to other researchers. Only the results of various analyses on the data are published and more fundamental data that could be used for future formulation and testing of colour-appearance models are lost forever. TC1-34 encourages any investigators performing visual experiments research to publish, whenever practical, tables of appropriate colorimetric data that can be used by others to develop and test models. This practice alone, would provide an immeasurable service to the color science community and those anxiously awaiting the results of current research activity and the resulting CIE recommendations.

The required data include the colorimetric coordinates of all samples, the background, the surround, and the illumination, the absolute illuminance, the size, shape, and configuration of the samples, and the visual results (matches or scale values). All data, particularly visual results, should include an uncertainty estimate. All colorimetric coordinates should be determined for the viewing environment used in the experiment rather than through the use of standard illuminants. Experimenters should also record (and publish if practical) the spectral power distributions of their stimuli. Such data will become useful if it ever becomes necessary to evaluate alternate sets of colour-matching functions.

VI. Conclusions

CIE TC1-34, Testing Colour Appearance Models, has published these guidelines to facilitate and promote additional research in the important area of colour-appearance modeling. Applications such as colour rendering of light sources, colour reproduction, image evaluation, and colour difference measurement are in need of a solution to the problem of predicting colour appearance across a variety of viewing conditions. While the pressure for a solution is great, a reliable one cannot be achieved without careful testing of the models with sound, quantitative visual data. While some data have recently become available, more are needed. Researchers in all areas of color science are urged to contribute existing data or perform new experiments to advance the state-of-the-art in this field. Anyone wishing to contribute to the work of TC1-34 by performing experiments as outlined in these guidelines or in other ways is encouraged to contact the committee chair with questions or results.

VII. Acknowledgements

The following members of CIE TC1-34, Testing Colour Appearance Models, took part in the development of these guidelines:

Paula J. Alessi, USA

Mark D. Fairchild, USA, Chair

K. Hashimoto, JAPAN

Robert W.G. Hunt, ENGLAND

M. Ronnier Luo, ENGLAND

Leo Mori, JAPAN

Yoshinobu Nayatani, JAPAN

Thorstein Seim, NORWAY

H. Sobagaki, JAPAN

Klaus Richter, GERMANY

VIII. References

1) Y. Nayatani, K. Takahama, H. Sobagaki, and K. Hashimoto, Color-appearance model and chromatic adaptation transform, *Color Res. Appl.* **15,** 210-221 (1990).

2) R.W.G. Hunt, An improved predictor of colourfulness in a model of colour vision, *Color Res. Appl.* **19**, 23-26 (1994).

3) Colour Appearance Analysis, CIE Pub. No 103, *Technical Collection*, Vienna, 1993.

4) *Colorimetry*, CIE Pub. No. 15.2, Central Bureau of the CIE, Vienna, 1986.

5) K. Richter, *Farbempfindungsmerkmal Elementarbuntton und Buntheitsabstände als Funktion von Farbart und Leuchtdichte von In- und Umfeld*, Bundesanstalt für Materialprüfung (BAM) Forschungsbericht 115, Berlin, (1985). (See also, *Color Res. Appl.* **5,** 25-43 (1980).)

6) M.D. Fairchild and R.S. Berns, Image color appearance specification through extension of CIELAB, *Color Res. Appl.* **18**, 178-190 (1993).

7) A.R. Robertson, CIE guidelines for coordinated research on colour-difference evaluation, *Color Res. Appl.* **3**, 149-151 (1978).

8) L. Mori, H. Sobagaki, H. Komatsubara and K. Ikeda, Field trials on CIE chromatic adaptation formula, *Proceedings of the CIE 22nd Session*, 55-58 (1991).

9) Fairchild, M.D., Pirrotta, E., and Kim, T.G., Successive-Ganzfeld haploscopic viewing technique for color-appearance research, *Color Res. Appl.* **19**, 214-221 (1994).

10) C.J. Bartleson, Changes in color appearance with variations in chromatic adaptation, *Color Res. Appl.* **4**, 119-138 (1979).

11) M.R. Pointer, The concept of colourfulness and its use for deriving grids for assessing colour appearance, *Color Res. Appl.* **5**, 99-107 (1980).

12) M.R. Luo, A.A. Clarke, P.A. Rhodes, A. Schappo, S.A.R. Scrivner, and C.J. Tait, Quantifying colour appearance. Part I. LUTCHI colour appearance data, *Color Res. Appl.* **16**, 166-180 (1991).

13) M.R. Luo, A.A. Clarke, P.A. Rhodes, A. Schappo, S.A.R. Scrivner, and C.J. Tait, Quantifying colour appearance. Part II. Testing colour models performance using LUTCHI color appearance data, *Color Res. Appl.* **16**, 181-197 (1991).

14) M.R. Luo, X.W. Gao, P.A. Rhodes, H.J. Xin, A.A. Clarke, and S.A.R. Scrivner, Quantifying colour appearance. Part III. Supplementary LUTCHI colour appearance data, *Color Res. Appl.* **18**, 98-113 (1993).

15) M.R. Luo, X.W. Gao, P.A. Rhodes, H.J. Xin, A.A. Clarke, and S.A.R. Scrivner, Quantifying colour appearance. Part IV. Transmissive media, *Color Res. Appl.* **18**, 191-209 (1993).

16) *International Lighting Vocabulary*, CIE Pub. No. 17.4, Central Bureau of the CEI, Geneva, 1987.

17) W.D. Wright, The measurement and analysis of color adaptation phenomena, *Proc. Roy. Soc. (London)* **115B**, 49-87 (1934).

18) R.W.G. Hunt, The effects of daylight and tungsten light-adaptation on color perception, *J. Opt. Soc. Am.* **40**, 362-371 (1950).

19) R.W.G. Hunt, Light and dark adaptation and the perception of color, *J. Opt. Soc. Am.* **42**, 190-199 (1952).

20) E.J. Breneman, Corresponding chromaticities for different states of adaptation to complex visual fields, *J. Opt. Soc. Am.* A **4**, 1115-1129 (1987).

21) W.D. Wright, Why and how chromatic adaptation has been studied, *Color Res. Appl.* **6**, 147-152 (1981).

22) M.D. Fairchild and P. Lennie, Chromatic adaptation to natural and incandescent illuminants, *Vision Res.* **32**, 2077-2085 (1992).

23) J.J. McCann, S.P. McKee, and T.H. Taylor, Quantitative studies in retinex theory, *Vision Res.* **16**, 445-458 (1976).

24) M.D. Fairchild and E. Pirrotta, Predicting the lightness of chromatic object colors using CIELAB, *Color Res. Appl.* **16** 385-393(1991).

25) Y. Nayatani, K. Takahama, H. Sobagaki, and K. Hashimoto, Color-appearance model and chromatic adaptation transform, *Color Res. Appl.* **15**, 210-221 (1990).

26) E. Pirrotta, *Testing Chromatic Adaptation Models*, M.S. Thesis, Rochester Institute of Technology, (1994).

27) T.G. Kim, R.S. Berns, and M.D. Fairchild, A comparison of color appearance models using pictorial hardcopy images, *IS&T/SID Color Imaging Conference* Proceedings, Scottsdale, 72-77(1993).

28) L.L. Thurstone, A law of comparative judgements, *Psych. Review* **34**, 273-286 (1927).

29) W.S. Torgerson, *Theory and Methods of Scaling*, Wiley, New York, 1958, pp.159-204.

30) C.J. Bartleson and F. Grum, *Optical Radiation Measurements, Vol. 5, Visual Measurements*, Academic Press, Orlando, 1984, pp.455-467.