HOW CLOSE IS CLOSE ENOUGH?
SPECIFYING COLOUR TOLERANCES FOR HDR AND WCG DISPLAYS

Jaclyn A. Pytlarz, Elizabeth G. Pieri

Dolby Laboratories Inc., USA

ABSTRACT

With a new high-dynamic-range (HDR) and wide-colour-gamut (WCG) standard defined in ITU-R BT.2100 (1), display and projector manufacturers are racing to extend their visible colour gamut by brightening and widening colour primaries. The question is: how close is close enough? Having this answer is increasingly important for both consumer and professional display manufacturers who strive to balance design trade-offs. In this paper, we present “ground truth” visible colour differences from a psychophysical experiment using HDR laser cinema projectors with near BT.2100 colour primaries up to 1000 cd/m^2. We present our findings, compare colour difference metrics, and propose specifying colour tolerances for HDR/WCG displays using the ΔIC_TC_P (2) metric.

INTRODUCTION AND BACKGROUND

From initial display design to consumer applications, measuring colour differences is a vital component of the imaging pipeline. Now that the industry has moved towards displays with higher dynamic range as well as wider, more saturated colours, no standardized method of measuring colour differences exists.

In display calibration, aside from metamerism effects, it is crucial that the specified tolerances align with human perception. Otherwise, one of two undesirable situations might result: first, tolerances are too large and calibrated displays will not appear to visually match; second, tolerances are unnecessarily tight and the calibration process becomes uneconomic. The goal of this paper is to find a colour difference measurement metric for HDR/WCG displays that balances the two and closely aligns with human vision.

Colour Difference Metrics

All perceptual colour difference metrics aim to predict colour differences as closely as possible to the way humans see them. It is possible that the same metric performs accurately for certain colours and poorly for others, so it is important to consider how well a metric adheres to human vision across a wide range of colours and luminance levels. As the industry transitions to HDR and WCG displays, it is essential to take into account this wider range.

Many specifications for monitor calibration exist today using different metrics. For example, the EBU (3) specification lists colour tolerances for grades 1, 2, and 3 monitors with the

Δu*v* metric. However, the most common metric used for measuring colour differences in displays has been a variation of the ΔE metric (4). ΔE has evolved from a simple Euclidean distance between the CIE L*a*b* coordinates of two stimuli (ΔE_{ab}) to more complex extensions (ΔE_{94} and ΔE_{00}). The CIE L*a*b* space was used because it was designed to mimic the response of the human visual system. The closer a colour representation mimics the human visual system, the simpler and better the colour difference metric can be. Despite being commonly used for display calibration today, ΔE_{00} was not designed for use with emissive colours. The CIE does not recommended ΔE_{00} for use on light-emitting or specularly reflecting colour stimuli (5).

Another colour difference metric for measuring colour differences is ΔIC_TC_P (2). ΔIC_TC_P is roughly the Euclidean distance between the IC_TC_P coordinates of two stimuli (with a scalar on C_T). As with ΔE, the IC_TC_P colour representation (defined in BT.2100) was used because it was designed to mimic the human visual system. However, this representation, instead of being optimized using reflecting surfaces as was done for CIE L*a*b*, was instead optimized for emitting colours found commonly in HDR and WCG displays. In previous research (2) it has been shown that ΔIC_TC_P predicts colour differences more accurately for displays than ΔE_{00}, and is therefore expected to perform more reliably.

JND Thresholds and Colour Difference Datasets

The driving force behind calibration is defining the threshold where humans can perceive a difference between two colours and where two colours appear identical. This concept is called a just-noticeable-difference (JND), which is the smallest difference the human eye can detect. Although a large amount of data has been collected over the years to establish the thresholds at which humans perceive colour differences, existing datasets do not cover all of the colours reproduced by modern displays.

One such dataset is the well-established MacAdam (6) dataset run by David MacAdam in 1942. This dataset includes 25 colours at about 47 cd/m^2. The MacAdam dataset covers many colours but tests only a single luminance level, is heavily influenced by a single observer, and does not test the range of colours reproduced by BT.2100 primary colours.

The RIT-DuPont dataset was extended by Hou (7) in 2010. This experiment was run on an LCD display and included 20 near-neutral and 20 high-chroma colours. These test colours ranged in luminance from roughly 45-90 cd/m^2, and while that covers more dynamic range than the MacAdam dataset, it still does not test dark, bright, or highly saturated colours. Furthermore, this experiment tested supra-thresholds rather than JNDs, which are less useful for calibration.

The 3M dataset collected by Hillis et al (8) in 2015 specifically tested colours near the BT.2020 (same primary colours as BT.2100) boundary using a laser-based Digital Light Processing display. This dataset includes 27 colour pairs, but unfortunately, as with the MacAdam set, the luminance levels of the colours are very limited. In addition, the experiment used shape detection, which makes the thresholds inherently higher.

Due to the limited availability of HDR and WCG colour difference data, we conducted our own JND user study. In this experiment we tested 75 colour pairs using two Christie HDR laser projectors. We tested colours from 0.1 to 1000 cd/m^2 out to the colour gamut boundary of BT.2100.

JUST-NOTICEABLE-DIFFERENCE USER STUDY

The main goal of this experiment was to gather a JND dataset of emissive colours across a wide range of saturation and luminance levels. This data would then be used for objective assessment of existing colour difference metrics with HDR/WCG colours.

Apparatus

The experiment took place in a small cinema using two Christie E3LH HDR laser cinema projectors. As shown in the spectral power distribution plot in Figure 1, the primaries of the two projectors differed slightly. This feature allows for 3D display by means of suitable glasses. We utilized the dual system to increase the peak luminance. The two projectors were fed a 12bit 4:4:4 RGB signal and reached a total combined luminance of 1575 cd/m^2. The display characteristics were modelled and the resulting prediction yielded an error less than one ΔE_{00}.

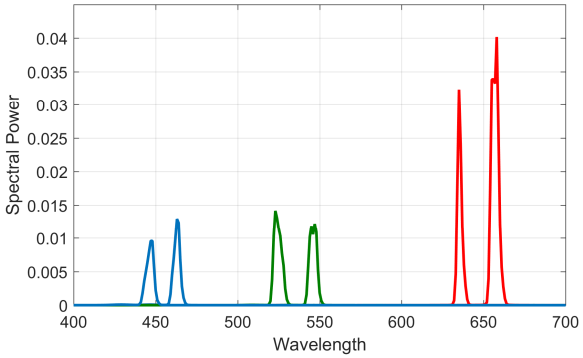

Figure 1 – SPD of Christie E3LH projectors

Each observer sat centred three picture heights away from the screen. The stimulus was concentrated in the middle of the screen to limit uniformity concerns.

Reference and Test Stimuli

Seven colours at three luminance levels were tested using the native primaries of the Christie laser projectors:

- Red, green, blue, cyan, magenta, yellow, white
- 0.1, 25, 1000 cd/m^2

The luminance of the 1000 cd/m^2 stimuli for each colour differed based on the combined capability of the projectors (for example, blue could not reach 1000 cd/m^2). As shown in Figure 2, the primary and secondary colours (red, green, blue, cyan, magenta, and yellow) were tested towards white and adjacent primary colours, whereas white was tested solely towards primary colours. A small change in luminance level was also tested for primary colours and white. The test was repeated for each luminance level. Observers were split into six testing groups to reduce the time per session to 30 minutes.

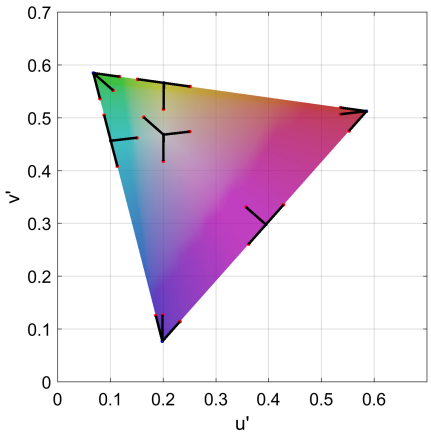

Figure 2 – test/reference points

Matching Procedure

We used a four-alternate-forced choice method because this has been shown to improve results for naïve observers (9). The observer was shown four squares simultaneously (Figure 3) similar to a discrimination experiment by Smith and Pokorny (10).

Three squares were identical and one was different. The observer was asked to select the square that was unique. The response controller had buttons corresponding to the location of the squares (Figure 4). It fit comfortably in the observer's hands so the observer did not need to look down when submitting a response.

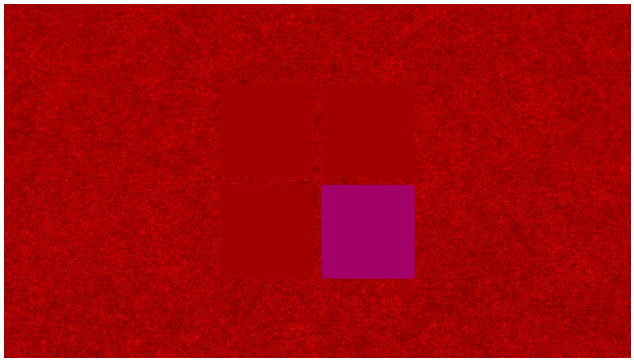

Figure 3 – stimulus pattern

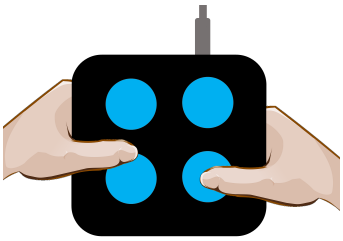

Figure 4 – user control

For convergence, we used the QUEST (11) method because it reduced the required number of trials by concentrating the test points around the threshold. Because adapting to the test colour produces the highest colour difference sensitivity, the adapting noise pattern and the background stimulus was, on average, equivalent to the test colour. 60 seconds of adaptation occurred before each trial, allowing observers to be over 80% adapted to the stimulus at the start of each trial (12).

Each square subtended three degrees of the visual field with a 0.3 degree gap between. The procedure included four steps, with steps 2 through 4 repeated 25 times:

1. 60 seconds of adaptation to the colour noise pattern
2. 4 squares appear (randomly arranged)
3. The observer chooses the square that he/she believes is unique
4. 0.5 seconds of adaptation

Results

The experiment had 27 male and 29 female participants. 23 were in the age range 20-29 years, 17 were 30-39 years, 12 were 40-49 years, and four were 50-59 years. Expertise was split between naïve and knowledgeable/expert observers. There was an average of seven participants for each test point/direction pair.

Previous ΔIC_TC_P optimization research (2) did not include colours that changed in luminance level. We are now able to verify the ΔIC_TC_P equation for the I channel. To match the ΔIC_TC_P results for equiluminance colours, no scale factor is required. The resulting ΔIC_TC_P equation is shown in Equation 1. To fairly compare the ΔIC_TC_P and ΔE_{00} metrics, normalization was required.

$$\Delta IC_TC_P = \sqrt{(\Delta I)^2 + 0.25*(\Delta C_T)^2 + (\Delta C_P)^2}$$

Equation 1 – Scalar on I in ΔIC_TC_P

This normalization centred the metrics on one JND and is shown in Table 1. We found that a scalar of 240 on ΔIC_TC_P equates the average measurement to ΔE_{00}. Then, a scalar of 3 on both metrics relates the measurements to a JND. This means that, on average for HDR/WCG colours, a ΔE_{00} value of 1/3 is roughly one JND. The adaptation point used for ΔE_{00} was D65 at the luminance of the reference point.

Metric	Scalar
ΔE_{00}	3
ΔIC_TC_P	3*240 = 720

Table 1 – JND scalars

For objective comparison of the ΔIC_TC_P and ΔE_{00} metrics, the colour difference data was segmented into categories based on luminance level and saturation as shown in Table 2. The Root Mean Squared Logarithmic Error (RMSLE) was used for this calculation in place of the Root Mean Squared Error in order to weight under-predictions as highly as over-predictions. The combined analysis is given at the bottom of Table 2. Each metric was used to measure the predicted perceptual difference between the reference and test point for each test/direction pair using the scalars from Table 1. A perfect perceptual metric would have consistently measured a value of one – corresponding to one JND – and would yield an RMSLE of zero. The metric's deviation from a perceptual match is reflected in the RMSLE error statistics.

Colours	ΔE_{00} RMSLE	ΔIC_TC_P RMSLE
0.01 cd/m²	0.65	0.34
25 cd/m²	0.33	0.23
1000 cd/m²	0.39	0.22
Saturated	0.45	0.28
Neutral	0.59	0.23
All	0.48	0.27

Table 2 – segmented error analysis

Each reference point was also analysed separately. The median JND results are shown in Figures 6-26. Figure 5 is a diagram of the components of each plot. The JND thresholds are plotted as black lines in u'v' from the reference point. At the end of these lines are colour coded circles (red, green, and blue in Figure 5) that indicate the test direction from the reference point. The mean absolute error (MAE) is displayed as grey circles surrounding the JND threshold points. Only one direction was measured, but the MAE was interpolated to form circles for ease of viewing. Ellipses of ΔE_{00} and ΔIC_TC_P are shown in magenta and cyan accordingly. These ellipses represent 1/3 ΔE_{00} and 1/720 ΔIC_TC_P. A perfect perceptual metric would intersect all three JND threshold points including the MAE circle regions.

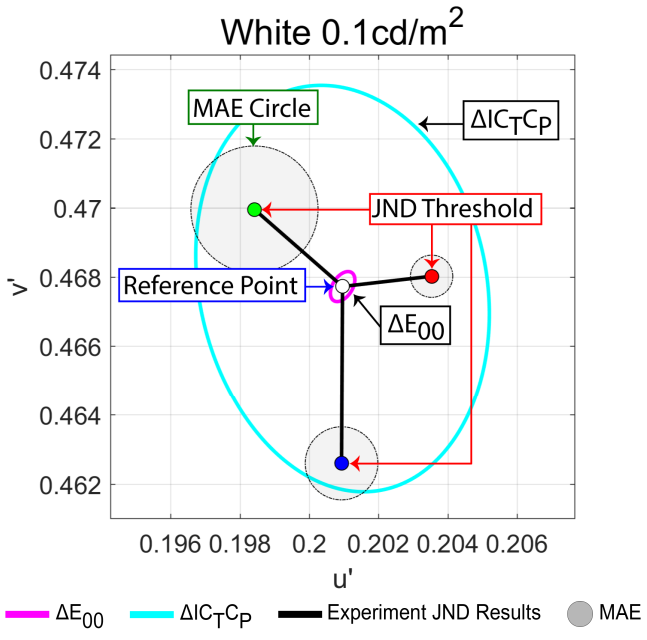

Figure 5

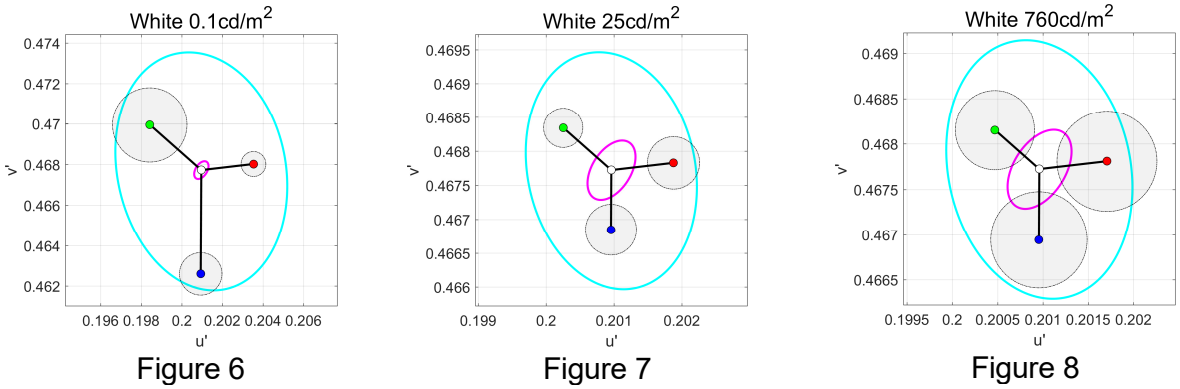

Figure 6 Figure 7 Figure 8

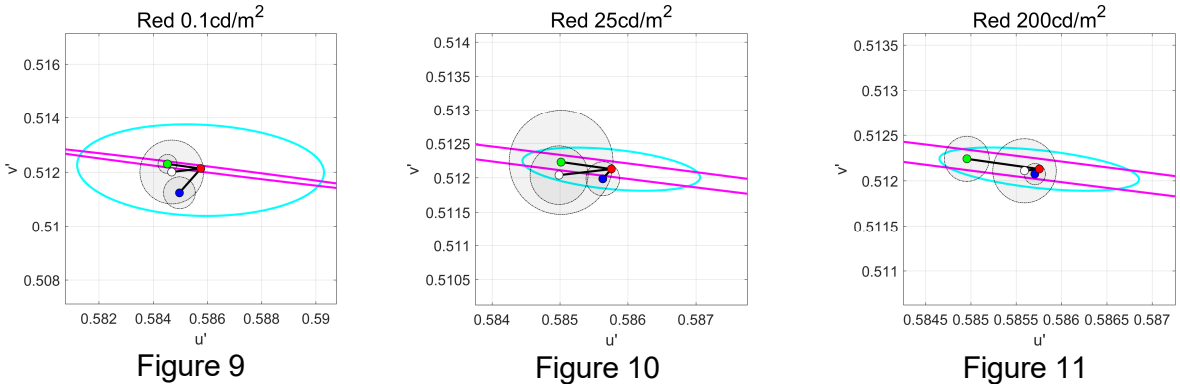

Figure 9 Figure 10 Figure 11

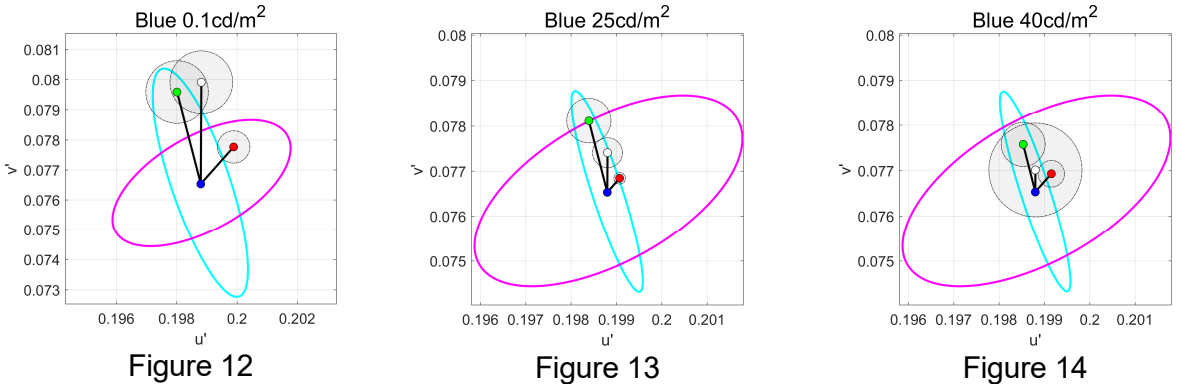

Figure 12 Figure 13 Figure 14

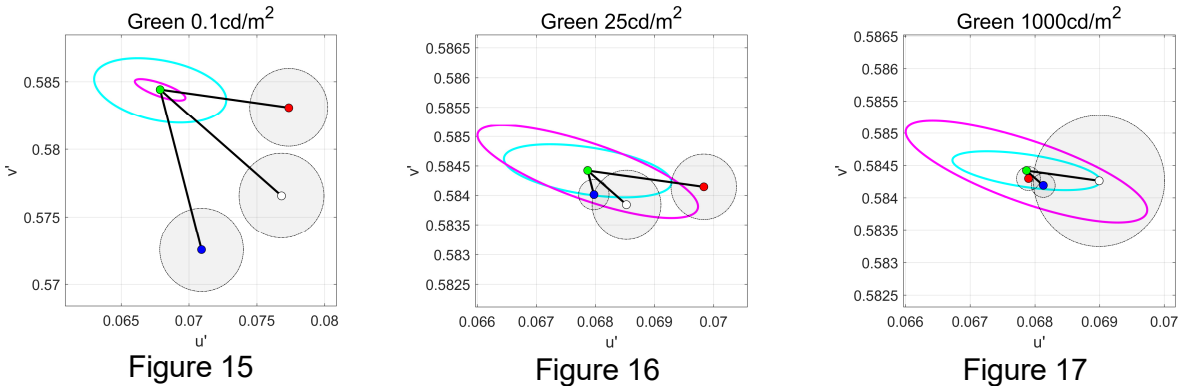

Figure 15 Figure 16 Figure 17

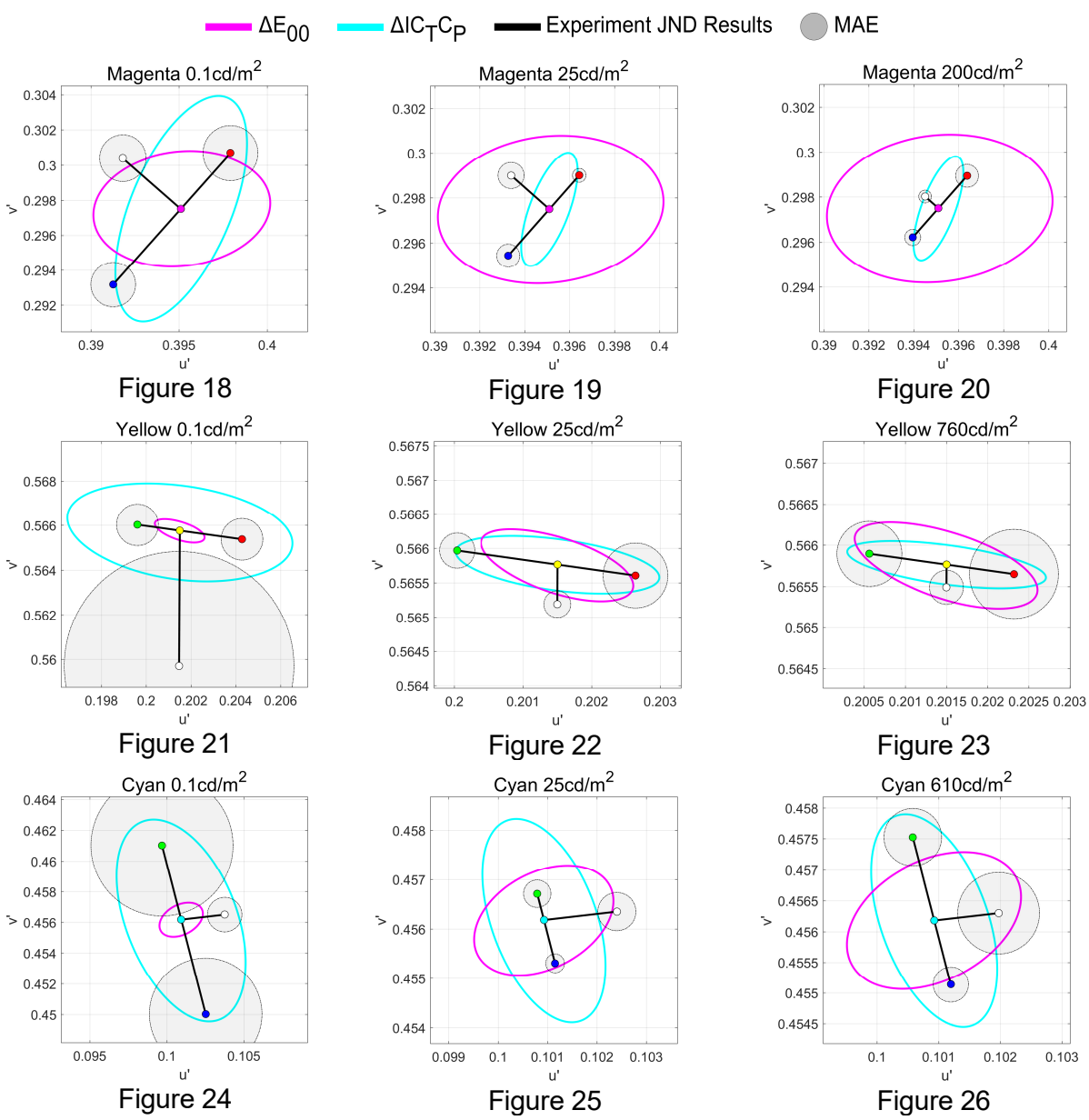

Figure 18 Figure 19 Figure 20

Figure 21 Figure 22 Figure 23

Figure 24 Figure 25 Figure 26

Discussion

The goal of defining an HDR/WCG calibration tolerance is to specify the maximum amount of variability that results in either an invisible, or at least acceptable, difference to a user. We want to define the tolerance as loosely as possible for efficiency of calibration, while still retaining a colour match. This means that we care just as much about over-prediction as we do about under-prediction. We are looking for the metric that most closely aligns with the reference/test pair magnitude. We want the edge of the ellipse to align exactly with the JND test points. If the metric ellipse is smaller than the JND results in a particular test direction,

this means that the metric will over-predict the colour difference. Likewise, if the metric ellipse is larger than the JND results in a particular test direction, this means that the metric will under-predict the colour difference.

In figures 6-26, for each colour, both size and shape of the JND results may be compared across the 0.1, 25, and 1000 cd/m^2 range. We expect consistency, and in most cases this is true. One surprising result is the inconsistency of the low luminance green in Figure 15. This inconsistency may be due to metameric deviation from the standard observer or due to the low number of observers for this point.

A metric used to define tolerances should be consistent across all colours. ΔE_{00} performs inconsistently for the white test points in figures 6-8. The ellipse size relative to the JND test points for the 0.1 cd/m^2 case is substantially smaller than the 760 cd/m^2 case, whereas ΔIC_TC_P remain consistent throughout the range. ΔIC_TC_P does, however, under-predict the median white differences, but it is close to within the mean absolute error of the data. Magenta is another case where the performance of ΔE_{00} suffers. As the luminance increases in Figures 18-20, the JND's get smaller, but the ΔE_{00} metric gets larger.

The red test points in Figures 9-11 have the strongest deviation between the performance of ΔE_{00} and ΔIC_TC_P. The long and narrow ellipses of ΔE_{00} extend well beyond the limits of the axes shown here. This means that ΔE_{00} will report vastly different measurements from the red primary towards green versus the red primary towards blue. This may make calibration both difficult and inconsistent because slight perceptual variations in primary location will yield substantially different numerical results.

We can compare these results to an existing tolerance in practice today: the EBU monitor specification. The requirement specifies a tolerance of 4Δu*v* (3) for a grade 1 monitor. Therefore, the delta prediction is a circle when plotted in u'v'. The blue 25 cd/m^2 point has been repeated in Figure 27, now including the EBU grade 1 monitor requirement. As the requirement is not colour specific, this same circle tolerance (shape and size) would also apply to other colours and luminance levels. If the JND results were uniform across all colours and luminance levels (i.e., had the same length), then a circle would be an appropriate representation. Because this is not the case, however (except perhaps for white), specifying colour tolerances in u*v* may not be appropriate for HDR/WCG displays.

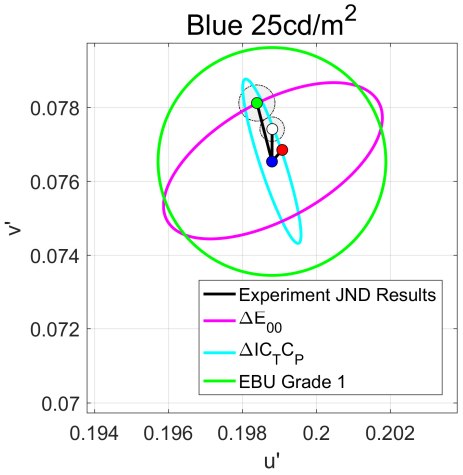

Figure 27

Table 2 shows the objective results of the categorized colours. As anticipated, we see that in all categories ΔIC_TC_P produces less error than ΔE_{00} as was reflected in the individual plots. It is surprising, that even for colours within the typical operating range of ΔE_{00} and within normal use for SDR imagery, ΔIC_TC_P still outperforms ΔE_{00}. As expected, the most significant differences in performance occur for colours in the 1000 cd/m^2 and saturated categories. This means that the ΔIC_TC_P metric better matches human colour difference perception than ΔE_{00}, especially for HDR/WCG colours.

CONCLUSION

In this paper, we presented a ground-truth colour difference dataset that encompassed HDR and WCG colours. We analysed the performance of ΔIC_TC_P and ΔE_{00} for the purpose of defining a calibration colour tolerance. The goal of a tolerance guideline is to align with perceptual colour differences. Because it better aligns with the ground truth subjective data, we therefore suggest that ΔIC_TC_P be used as the colour difference metric for HDR and WCG displays. We found that 1/720 ΔIC_TC_P was, on average, equivalent to a JND. If perceptual equivalency is the goal of calibration, then a tolerance of 1/720 ΔIC_TC_P should be used. In the future, this experiment should be repeated to increase the number of participants per sample point.

REFERENCES

1. Rec. ITU-R BT.2100. Image parameter values for high dynamic range television for use in production and international programme exchange. July 2016.
2. Pytlarz, J., Pieri, E., Atkins, R. 2017. Objectively Evaluating High-Dynamic-Range and Wide-Color-Gamut Color Differences. SMPTE Motion Imaging Journal. March 2017, pp. 27 to 32.
3. EBU. User Requirements for Video Monitors in Television Production. 3320, Oct, 2014.
4. Hunt, R. W. G., 2004. The Reproduction of Colour. John Wiley & Sons, Ltd: Chichester.
5. ISO/CIE 11664-6:2014. Colorimetry – Part 6: CIEDE2000 Colour-difference-formula. CIE International Commission on Illumination. Feb 2014.
6. MacAdam, D. L.,1942. Visual Sensitivities to Color Difference in Daylight. Journal of the Optical Society of America. 32, May 1942, pp. 247-274.
7. Hou, B., 2010. Extending the RIT-DuPont suprathreshold data set: Weighted individual discrimination pair data and new chroma dependency visual data. RIT. Nov 2010.
8. Hillis, J. M. et al, 2015. Closing in on Rec. 2020 – how close is close enough? The Society for Information Display. 2015.
9. Jäkel, F. et al, 2006. Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. Journal of Vision. 6. Pp. 1307-1322.
10. Smith, V. et al, 1999. Chromatic Contrast Discrimination: Data and Prediction for Stimuli Varying in L and M Cone Excitation. University of Chicago.
11. Watson, A., Pelli, D. 1983. Quest: A Bayesian Adaptive Psychometric Method. Perception & Psychophysics. 33 (2), pp. 113-120.
12. Rinner, O., Gegenfurtner, K. 2000. Time Course of Chromatic Adaptation for Color Appearance and Discrimination. Vision Research 40. pp. 1813-1826.