Research & Development White Paper

WHP 342

January 2019

A Brightness Measure for High Dynamic Range Television

K. C. Noland and M. Pindoria

BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper WHP 342

A Brightness Measure for High Dynamic Range Television

Katy C. Noland and Manish Pindoria

Abstract

Now that standards for a complete high dynamic range (HDR) television ecosystem are complete, the industry is taking its first steps in HDR production. HDR is associated not only with a greater dynamic range, but also brighter screens than conventional television, so the potential arises for unwanted, uncomfortable brightness jumps at programme junctions and channel changes. To ensure a degree of consistency between programmes, some production guidelines for HDR brightness are required. In this paper, we summarise tests showing that the mean displayed pixel luminance is a good predictor of subjective brightness. We then explore viewer tolerance to brightness shifts of different sizes, and propose a potential normal operating range for the mean display luminance of 10–80 cd/m^2, extending to 5–160 cd/m^2 for special creative effect.

This paper was originally published in the IBC Conference, Amsterdam, The Netherlands, September 2017. It has been updated to reflect developments in HDR standards since the original publication.

Additional key words: HDR; UHD; brightness; perception; vision; human visual system

White Papers are distributed freely on request.

Authorisation of the Head of Standards or Chief Scientist is required for publication.

BBC Research & Development
White Paper WHP 342

A Brightness Measure for High Dynamic Range Television

Katy C. Noland and Manish Pindoria

1 Introduction

High dynamic range (HDR) television is now standardised in ITU-R BT.2100 [1], and programme makers are starting to create HDR content. With the brighter screens and greater dynamic range associated with HDR, there comes the potential for sudden uncomfortable brightness jumps at programme junctions, breaks for advertisements, or when switching between channels. A need for production guidelines for HDR television has been identified, just as guidelines were required for audio loudness [2]. The guidelines should describe suitable brightness levels for HDR programmes, to prevent unwanted, uncomfortable brightness jumps.

Conventions that map specific scene luminance levels to appropriate signal levels are widely used in standard dynamic range (SDR) television, although they are not officially standardised. They include "diffuse white" or an ice hockey rink at around 90 % signal level, "flesh tones" at around 50–70 % signal level, and grass or an 18 % reflectance card at 50 % signal level. Conventional SDR displays are generally too dim to allow good quality pictures to be produced using only part of the signal range, so these best-practice conventions have developed to ensure that full use is made of the available dynamic range. This has the secondary effect of maintaining a degree of brightness consistency between programmes, which means that large jumps in brightness do not usually occur. There has not therefore been a specific need for brightness guidelines for SDR.

For HDR, displays are generally brighter, so there can be more flexibility in the signal range used for individual scenes. Reference levels to ensure good quality pictures will still be needed for HDR, but they are likely to be more flexible than those for SDR. A reference level for graphics and "diffuse white" of 75 % signal level for Hybrid Log-Gamma or 58 % signal level for Perceptual Quantization [3, table 1]. This represents a nominal level for "diffuse white" in a typical scene, with some variation around this level expected to allow artistic freedom.

However, reference levels for specific objects and the overall brightness are separate issues: a nominal "diffuse white" level tells us little about the overall scene brightness. The greater flexibility of HDR means that there is a greater chance of uncomfortable brightness jumps occurring scene-to-scene. Some brightness guidelines are therefore also required.

Before production guidelines for brightness can be defined, it is necessary to be able to measure brightness. It is a subjective quantity, so subjective test methodology must be used. In this paper we summarise an experiment to find the perceived brightness of a set of HDR images. We use the results to evaluate a range of potential objective brightness metrics, and show that the mean displayed pixel luminance performs very well. It is also simple enough to implement easily in real-time brightness monitoring systems.

Having established an objective method to measure brightness, our main contribution in this paper is an investigation of viewer tolerance to sudden brightness changes. The test results suggest a range of HDR brightness levels that would prevent viewer discomfort in normal television programming.

2 Related Work

There is a large body of work on brightness perception and adaptation, but none to date that specifically addresses the question in the context of HDR television. The human visual system is sensitive to changes in viewing conditions, so previous experiments may or may not apply to television viewing.

One possible approach to mapping luminance to brightness is to count just noticeable differences (JNDs) in brightness. A logarithmic relationship is implied by Weber and Fechner's work on JNDs [4, p. 136], or a more precise characterisation of JNDs is provided by the Perceptual Quantization inverse electro-optic transfer function (EOTF^{-1}) [1]. However, Stevens [5] contested the use of JNDs for luminance differences above the threshold of perceptibility, and proposed instead a power law relationship. Bartleson and Breneman [6] proposed a modified log relationship, which has been shown to have the same form as Stevens' power law when the parameters are chosen appropriately for the viewing conditions [7]. The CIE 1976 lightness measure [8] also has the form of a power law, but is normalised to a reference white so is unlikely to apply to absolute brightness.

For television viewing it is also necessary to have a method of accounting for the contribution of different parts of the image to the overall brightness. Bauer [9] showed that Stevens' power law holds for estimates of the mean brightness of a set of test patches, but for the current purpose we require a measure of the overall impression of brightness rather than a viewers estimate of the mean. This could be achieved with a measure of the eye's adaptation level. Suggested models of adaptation level include perceived "middle gray" [10], the average luminance [11, p. 1], and a weighted average luminance where the area at the centre of the field of view has the highest weighting [12].

Adaptation is not instantaneous, and complete adaptation to darkness can take many minutes [13]. However, adaptation to a lesser difference in luminance is much faster, and can pass largely unnoticed [11, p. 1]. Brightness guidelines should aim to keep the adaptation level within the range of comfortable sudden changes.

3 Measuring Brightness

Our first experiment aimed to find a suitable objective metric for the overall perceived brightness of HDR images [14, 15]. First we conducted subjective tests to obtain ground truth brightness values for a set of HDR images, then we evaluated a range of potential objective metrics using the correlation of their output with the subjective test results. Here we provide an overview of the tests, since they are critical to our main contribution on viewer tolerance to brightness jumps of different sizes.

3.1 Subjective Test

Test subjects were asked to adjust the brightness of a grey flat frame until it matched the perceived overall brightness of a test image. The luminance of the grey slate is known, and can be used as a numerical value that is representative of the brightness of the image. Fifteen HDR test images were used, stored as Hybrid Log-Gamma HDR, and shown at four peak display luminances: 500, 1000, 2000 and 4000 cd/m^2, with an appropriate gamma function for the peak luminance applied [1]. This is simply a method of increasing the range of brightnesses used in the test, and does not limit the applicability of the results to any particular approach to display adaptation. Twenty subjects completed the test.

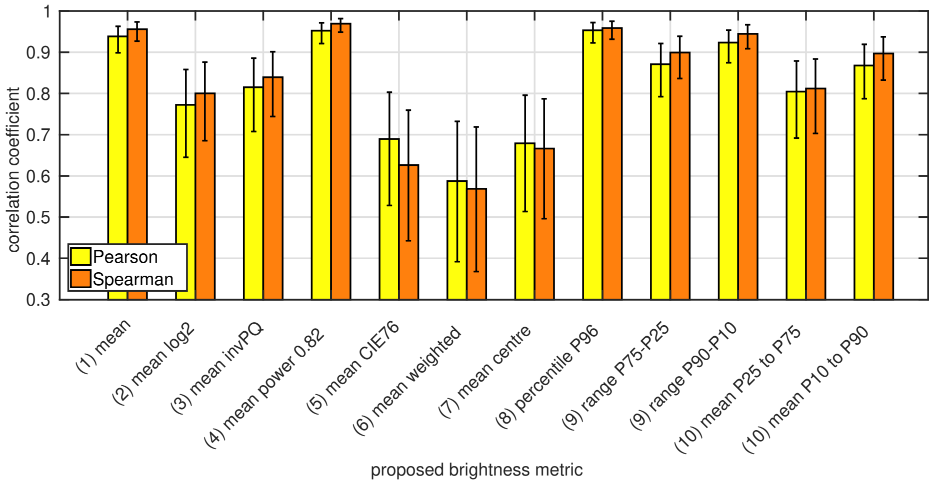


Figure 1: Correlation coefficients for all proposed metrics, with 95 % confidence intervals.

3.2 Objective Metrics

We implemented ten classes of objective metric that estimate the overall brightness from the displayed pixel luminance values of our test images. For full details, see [14, 15].

1. **Mean display luminance**
2. **Mean $\log_2$(display luminance)**
3. **Mean PQ inverse EOTF of display luminance**
4. **Mean display luminance raised to a power**
5. **Mean CIE 1976 lightness** using the displayed light values corresponding to 75 % signal level as the reference white.
6. **Weighted mean display luminance**, a simplified version of Moon and Spencer's weighting function [12].
7. **Mean of values in central quarter of screen**
8. **Percentiles** from P_{10} to P_{100}.
9. **Percentile ranges** $P_{75} - P_{25}$ and $P_{90} - P_{10}$.
10. **Mean of values within a percentile ranges** P_{25} to P_{75} and P_{10} to P_{90}.

The mean selected levels of the grey flat frames, for the 60 test images (15 images at 4 peak luminance levels), were used as ground truth brightness values to evaluate the models. The base-2 logarithm of the ground truth and of the output of each model was calculated, to improve the perceptual uniformity of the error space, then Pearson's correlation coefficient and Spearman's rank correlation coefficient were found. These are summarised in Figure 1.

The best performing metrics are the 96th percentile of the displayed pixel luminance values, the displayed pixel luminances raised to a power of 0.82 before calculating the mean, and the mean displayed pixel luminance values calculated directly. The performance differences between these three metrics are very small, and their confidence intervals overlap, so the simplest method, direct calculation of the mean, is preferred for real-time applications. The high correlation of 0.94 suggests that this simple metric will be an effective basis for a brightness monitor.

4 Measuring Tolerance to Brightness Shifts

Having established a method of measuring the brightness, it is desirable to understand viewer tolerance of brightness jumps. We conducted an experiment to find the size of brightness jump that can be tolerated.

4.1 Experiment

To simulate a programme junction or channel change, subjects were presented with an HDR image for 10 seconds, followed by a transition to a second HDR image with another mean displayed luminance (brightness). Subjects were asked to rate the change in overall brightness using a five-grade impairment scale [16]:

1. imperceptible
2. perceptible, but not annoying
3. slightly annoying
4. annoying
5. very annoying

This was repeated for a range of brightness transitions, including dark to bright and bright to dark.

The test images were shown on a SIM2 HDR47E display using its calibrated LogLUV mode. Three adjustable LED lights illuminated the wall behind the screen such that the light reflected off the wall measured $5\,\mathrm{cd/m^2}$. We used a fixed peak display luminance, since this is representative of typical viewing. It was set to $1000\,\mathrm{cd/m^2}$. The black level was set to $0.014\,\mathrm{cd/m^2}$, using a PLUGE-style test image [17] specifically designed for HDR. Subjects were seated at a distance of 3.2 times the screen height (3.2 H).

As a basis for our test images, we used 12 images from Mark Fairchild's HDR Photographic Survey [18], supplemented by 2 images created by BBC R&D. The raw images had been converted to BT.2100/BT.2020 colour primaries [1], and scaled to look aesthetically pleasing (as judged by a small number of expert viewers) on an HDR display, which is equivalent to adjusting the camera iris. The images had a resolution of 1920×1080 pixels to match the maximum resolution of the display.

In order to create a set of images with specific mean displayed luminance values, further processing was required. It was not important for the images to be beautifully graded, we only required images of reasonably good quality that covered a range of brightnesses. Hence we used a simple method to adjust the mean displayed luminance values whilst maintaining the signal range.

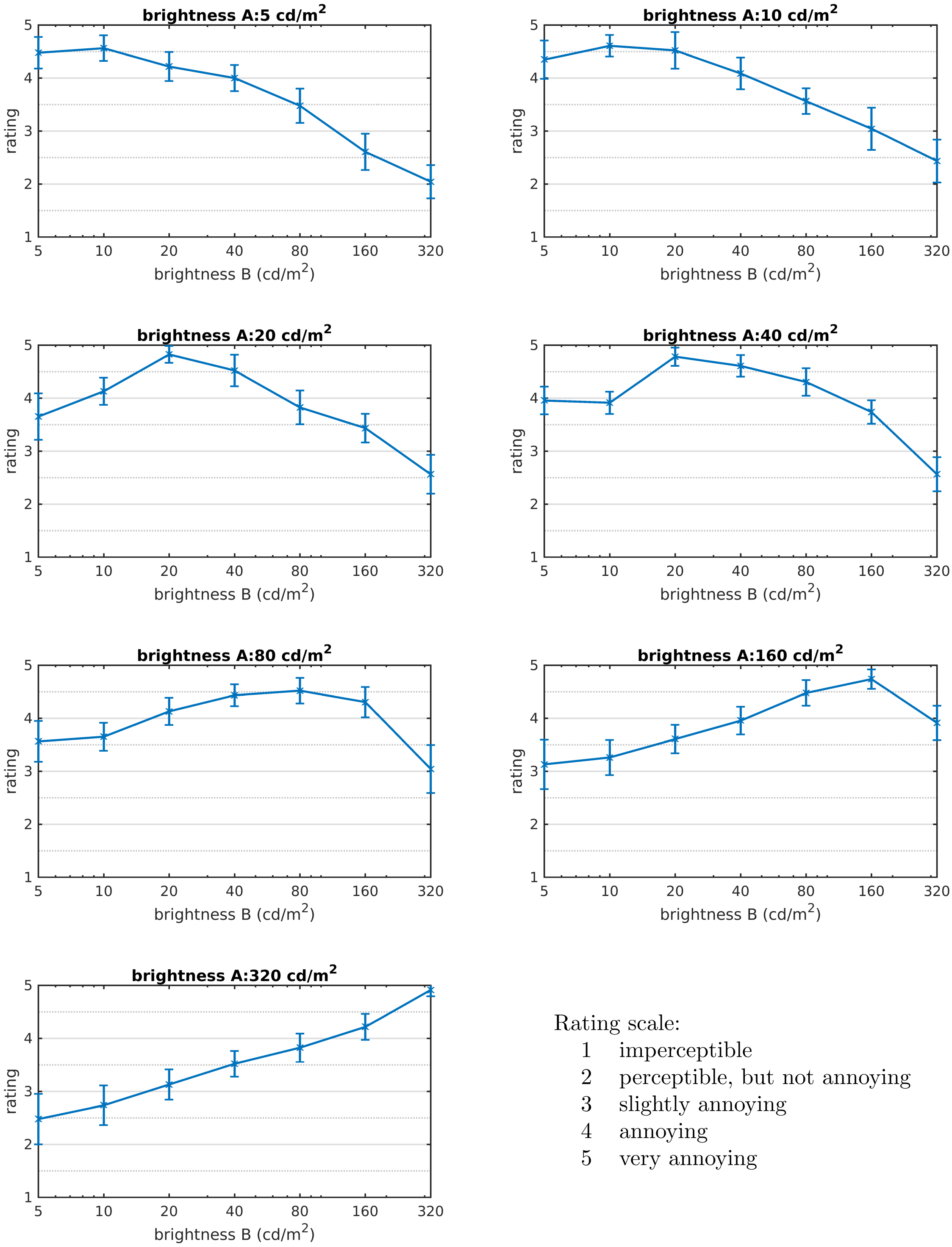
To increase the mean displayed luminance of an image, a gain higher than 1 was applied up to a threshold of 70 % signal level, above which a gain of less than 1 compressed the highlights to within the signal range. To decrease the mean displayed luminance, the same approach was applied, but with a gain of less than one below 70 % signal level, and a gain higher than 1 above the threshold to maintain highlights. Precise gains were manually tuned for each image to achieve the desired mean luminance. This produced images of the desired brightness that were reasonably aesthetically pleasing. Subjects were specifically instructed to ignore image quality and concentrate on brightness when making their judgements. The test images are shown in Figure 2.

We created two sets of seven test images with mean displayed luminance values of 5, 10, 20, 40, 80, 160 and $320\,\mathrm{cd/m^2}$ on a $1000\,\mathrm{cd/m^2}$ display, and tested transitions with all combinations of brightnesses, from bright to dark and dark to bright. This gave a total of 49 combinations, which were presented in a different random order for each subject. The images making up each transition were chosen at random from the pool of two with the same brightness. Subjects were presented with a grey frame indicating the test number for 3 seconds, then image A was displayed for 10 seconds, followed by image B for 10 seconds. Two training examples were provided before the test, and three "dummy" presentations were included at the start of the test that were not included in the results. Subjects were screened for normal visual acuity, after which twenty-three subjects completed the test.

5 Results

The mean ratings are shown in Figure 3. If we regard a rating of 3.5 as the threshold for an annoying change in brightness (halfway between "perceptible, but not annoying" and "slightly annoying")

Figure 2: Test images at specified mean display luminance values on a 1000 cd/m^2 display. The two images marked with an asterisk were created by BBC R&D, all others are from [18].



Rating scale:

1 imperceptible
2 perceptible, but not annoying
3 slightly annoying
4 annoying
5 very annoying

Figure 3: Mean ratings and 95 % confidence intervals for transitions from brightness A to B.

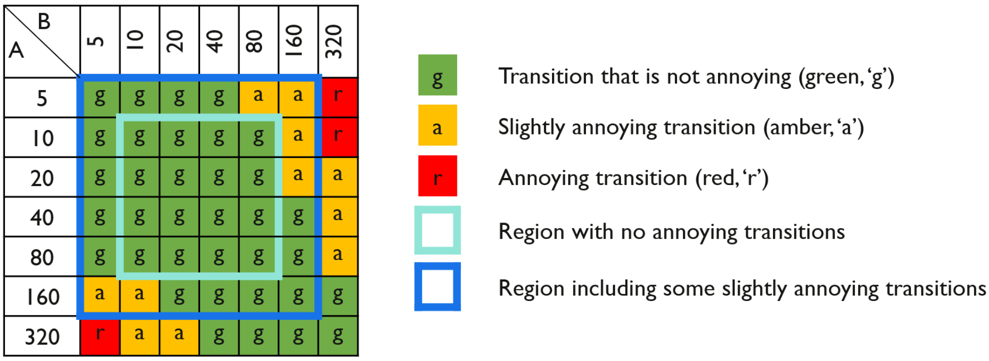


Figure 4: Transitions from brightness A (cd/m^2) to brightness B (cd/m^2) categorised by level of annoyance.

it can be seen that a transition to a scene that is around 3 stops brighter is not annoying for the darkest values of brightness A, decreasing to 1 stop when brightness A is $80\,cd/m^2$. This suggests that the brightest images were uncomfortably bright, regardless of the preceding image. With an agreed reference level for graphics at $203\,cd/m^2$ on a $1000\,cd/m^2$ display, it is not surprising that a mean image brightness of $320\,cd/m^2$ felt too bright, as it would not normally be desirable for the mean display brightness to be higher than the graphics level.

When the transition is from bright to dark, subjects appear to have a slightly higher tolerance to brightness jumps. There were no annoying downward transitions from 5, 10, 20, 40 or $80\,cd/m^2$, and downward jumps of 3 stops from 160 or $320\,cd/m^2$ were considered acceptable. That implies that it is slightly more annoying to increase than to decrease the brightness. However, for the purposes of production guidelines, a transition can occur in either direction, for example when switching between two channels, so the more restrictive values must be used.

One subject commented that it was more annoying to change from dark to bright than bright to dark. Another said that large transitions from dark to light were uncomfortable, whereas large transitions from light to dark resulted in a delay while the image "appeared", but both effects were considered annoying. Three subjects also mentioned that the context was important, so, for example, a transition from dark to bright was considered less annoying if it was a transition from a night scene to a day scene, because a large jump in brightness would be expected in this case. Large jumps were more annoying when both scenes looked as if they ought to have the same brightness.

If the rating threshold is relaxed to 2.5, a larger increase or decrease in brightness of 4-5 stops can be permitted. Figure 4 colour-codes the transitions according to thresholds of 2.5 (slightly annoying, amber) and 3.5 (not annoying, green). Based on these results, we might consider a normal brightness range of around 10–$80\,cd/m^2$ mean display luminance for most long form content, extending to 5-$160\,cd/m^2$ for special creative effect.

6 Conclusion

With a view to creating production guidelines for the brightness of HDR television programming, we have studied the perception of brightness. We conducted subjective tests to obtain values for the perceived brightness of a set of HDR images, and evaluated several potential objective brightness metrics using their correlation with our subjective brightness scores. We found that the 96th percentile of the mean pixel display luminance, mean pixel display luminance raised to a power of 0.82, and the mean pixel display luminance calculated directly were all good predictors of the subjective brightness. The simplest metric, calculating the mean display luminance directly, is preferred for real-time applications.

We then conducted subjective tests to investigate viewer tolerance to brightness shifts. Suggested mean display luminance levels for normal operation are in the range 10–$80\,cd/m^2$, extending to 5–$160\,cd/m^2$ for particular creative effects. Before finalising these ranges, further testing should be conducted in different viewing environments, on different displays, and with a wider range of material. It will also be necessary to map the display luminance values back to signal levels if brightness monitors are to operate as part of a broadcast signal chain.

Future brightness production guidelines should not prohibit brightness jumps within programmes that are wanted for artistic effect, such as explosions or black-outs; the aim is rather to maintain some consistency of the normal operating level between programmes. Further work is required to determine appropriate guidelines regarding the amount of time for which normal operating ranges can be exceeded for artistic effect.

Acknowledgments

We would like to thank Andrew Cotton, Tim Borer, Simon Thompson, and all the test subjects.

References

[1] ITU-R, "Image Parameter Values for High Dynamic Range Television for Use in Production and International Programme Exchange," Recommendation ITU-R BT.2100-2, July 2018.

[2] ——, "Algorithms to Measure Audio Programme Loudness and True Peak Audio Level," Recommendation ITU-R BS.1770-4, October 2015.

[3] ——, "Operational Practices in HDR Television Production," Report ITU-R BT.2408-1, April 2018.

[4] G. T. Fechner, *Elemente der Psychophysik.* Breitkopf und Haertel, 1860.

[5] S. S. Stevens, "On the Psychophysical Law," *The Psychological Review*, vol. 64, no. 3, pp. 153–181, 1957.

[6] C. J. Bartleson and E. J. Breneman, "Brightness Perception in Complex Fields," *J. Opt. Soc. Am.*, vol. 57, no. 7, pp. 953–957, Jul 1967.

[7] K. F. Choi, "Relationship Between Bartleson and Breneman's Brightness vs. Luminance Equation and Stevens' Power Law," in *IS&T's 47th Annual Conference/ICPS*, 1994, pp. 433–436.

[8] Commission Internationale de l'Eclairage, "CIE 1976 L*u*v* Colour Space," 2014. [Online]. Available: http://eilv.cie.co.at/term/159

[9] B. Bauer, "Does Stevens' Power Law For Brightness Extend to Perceptual Brightness Averaging?" *The Psychological Record*, vol. 59, pp. 171–186, 2009.

[10] C. J. Bartleson, "Optimum Image Tone Reproduction," *Journal of the SMPTE*, vol. 84, no. 8, pp. 613–618, August 1975.

[11] H. B. Barlow, *Dark and Light Adaptation: Psychophysics.* Berlin, Heidelberg: Springer Berlin Heidelberg, 1972, pp. 1–28.

[12] P. Moon and D. E. Spencer, "The Visual Effect of Non-Uniform Surrounds," *J. Opt. Soc. Am.*, vol. 35, no. 3, pp. 233–248, 1945.

[13] W. S. Stiles, "Adaptation, Chromatic Adaptation, Colour Transformation," *Anales Real Soc. Espan. Fis. Quim.*, vol. 57, Series A, pp. 149–175, 1961.

[14] K. C. Noland, M. Pindoria, and A. Cotton, "Modelling Brightness Perception for High Dynamic Range Television," in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017.

[15] ——, “Modelling Brightness Perception for High Dynamic Range Television,” BBC Research & Development White Paper WHP 341, 2019.

[16] ITU-R, “Methodology for the Subjective Assessment of the Quality of Television Pictures,” Recommendation ITU-R BT.500-13, January 2012.

[17] ——, “Specifications of PLUGE Test Signals and Alignment Procedures for Setting of Brightness and Contrast of Displays,” Recommendation ITU-R BT.814-4, July 2018.

[18] M. Fairchild, “The HDR Photographic Survey,” 2006–07. [Online]. Available: http://rit-mcsl.org/fairchild/HDR.html