Subjective Evaluation of Higher Dynamic Range Video

Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi

Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland

ABSTRACT

High dynamic range (HDR) imaging is able to capture a wide range of luminance values, closer to what the human eye can perceive. However, for capture and display technologies, it is important to answer the question on the significance of higher dynamic range for user preference. This paper answers this question by investigating the added value of higher dynamic range via a rigorous set of subjective experiments using paired comparison methodology. Video sequences at four different peak luminance levels were displayed side-by-side on a Dolby Research HDR RGB backlight dual modulation display (aka 'Pulsar'), which is capable of reliably displaying video content at $4000\ \mathrm{cd/m^2}$ peak luminance. The results of the subjective experiment demonstrate that the preference of an average viewer increases logarithmically with the increase in the maximum luminance level at which HDR content is displayed, with $4000\ \mathrm{cd/m^2}$ being the most attractive option.

Keywords: High dynamic range video, peak luminance, subjective evaluation, paired comparison, HDR monitor, Thurstone case V model.

1. INTRODUCTION

Several technological revolutions have impacted the television industry over the last decades, such as the shifts from black & white to color TV and standard to high definition TV. Nevertheless, considerable improvements may still be achieved along several orthogonal axes, including resolution, color, frame rate, contrast, and brightness. A few years ago, 3DTV was advertised as the future of television. However, due to the lower picture quality and the need to wear cumbersome glasses, 3DTV did not fulfill customer satisfaction and is loosing interest of research community and industry. Recently, the attention is turned to ultra high definition (UHD) TV but some believe that UHD TV is likely to repeat the fate of 3DTV. Another strong candidate for becoming a future TV technology is high dynamic range (HDR) imaging, which due to the recent advances in display technology is gaining a lot of interest from both researchers and practitioners.

HDR allows rendering a wider range of luminance values to better represent the details in both dark and bright areas, which is closer to what the human eye can perceive. An important question however is how significantly this technology impacts the viewers' quality of experience. Until now, most of the research in HDR imaging focused on the ability to capture, store, and display HDR content using conventional imaging technology. A significant amount of effort has been spent on designing algorithms, coined *tone-mapping operators*, for accurate reproduction of HDR content on conventional displays with limited dynamic range.[1–4] Several formats and compression techniques were also developed for HDR, mostly focusing on backward compatibility with popular imaging formats such as JPEG[5–8] and MPEG,[9,10] including the upcoming JPEG XT standard for HDR image compression. The tone-mapping operators and formats were studied and compared using different evaluation methodologies (subjective and objective) to determine the most suitable for different usage scenarios and displays.[11–17] However, since high quality HDR displays did not exist, the direct effect of HDR video technology on viewing experience was little studied, except for a few works[18,19] that relied on the limited first generation of HDR displays, such as BrightSide DR37-P monitor.

The recent breakthrough in the HDR capturing and displaying technologies allowed to bridge this gap and, hence, this paper evaluates the added value of higher dynamic range to user preference using high quality HDR video sequences and the latest professional HDR monitor. For this purpose, several HDR video sequences were displayed at four different peak luminance levels, including the maximal supported level of 4000 nit, in a side-by-side fashion on a professional reference HDR monitor, 'Pulsar' prototype developed by Dolby, renown for the accurate and reliable reproduction of color and luminance. The black level was held constant, so the luminance dynamic range was solely determined by

Further author information: (Send correspondence to Philippe Hanhart) E-mail: {firstname.lastname}@epfl.ch

the maximum luminance. The tested luminance levels reflect four levels of dynamic range that are typical for current and future consumer scenarios, given today's current displaying technology and latest advances in HDR displays. Two different evaluation methodologies were selected and compared in terms of accuracy and reliability in constructing a scale of perceptual preferences. The subjects participating in the evaluation experiment were naïve viewers. An adaptation of Thurstone Case V model was used in the analysis of the subjective scores to measure the subjective preference among different luminance levels at which the high dynamic range video sequences were displayed.

The remainder of the paper is organized as follows. Section 2 presents the details about the dataset and subjective evaluations. Section 3 discusses methods used for analysis of results, while Section 4 reports and analyzes subjective evaluation results. Section 5 concludes the paper.

2. METHODOLOGY

2.1 Dataset

Eight video sequences representing various levels of dynamic range and with different visual characteristics (see Fig. 1) were used in the experiments. Two additional sequences were used during a training session. Each video sequence was about 20 seconds long. All video clips, except one from "Star Trek: Into Darkness", were shot by professional film directors specifically for this experiment.

For each video content, four dynamic range levels were selected to represent several key use cases, as opposed to using uniform perceptual distances:

1) 4000 nit version, which was manually graded by professional colorists from the originally captured video. This value was determined by the HDR monitor used in the experiments (see Section 2.2) and the availability of professionally color graded content for this luminance range.
2) 1000 nit version, which was tone-mapped from manually graded 4000 nit version. This value represents some very high end consumer TVs.
3) 400 nit version, which was tone-mapped from manually graded 4000 nit version. This value is a good representation of the maximum luminance level of current high-quality consumer TVs.
4) 100 nit version, which was tone-mapped from manually graded 4000 nit version. This value is a commonly used maximum luminance level for reference monitors in a production environment.

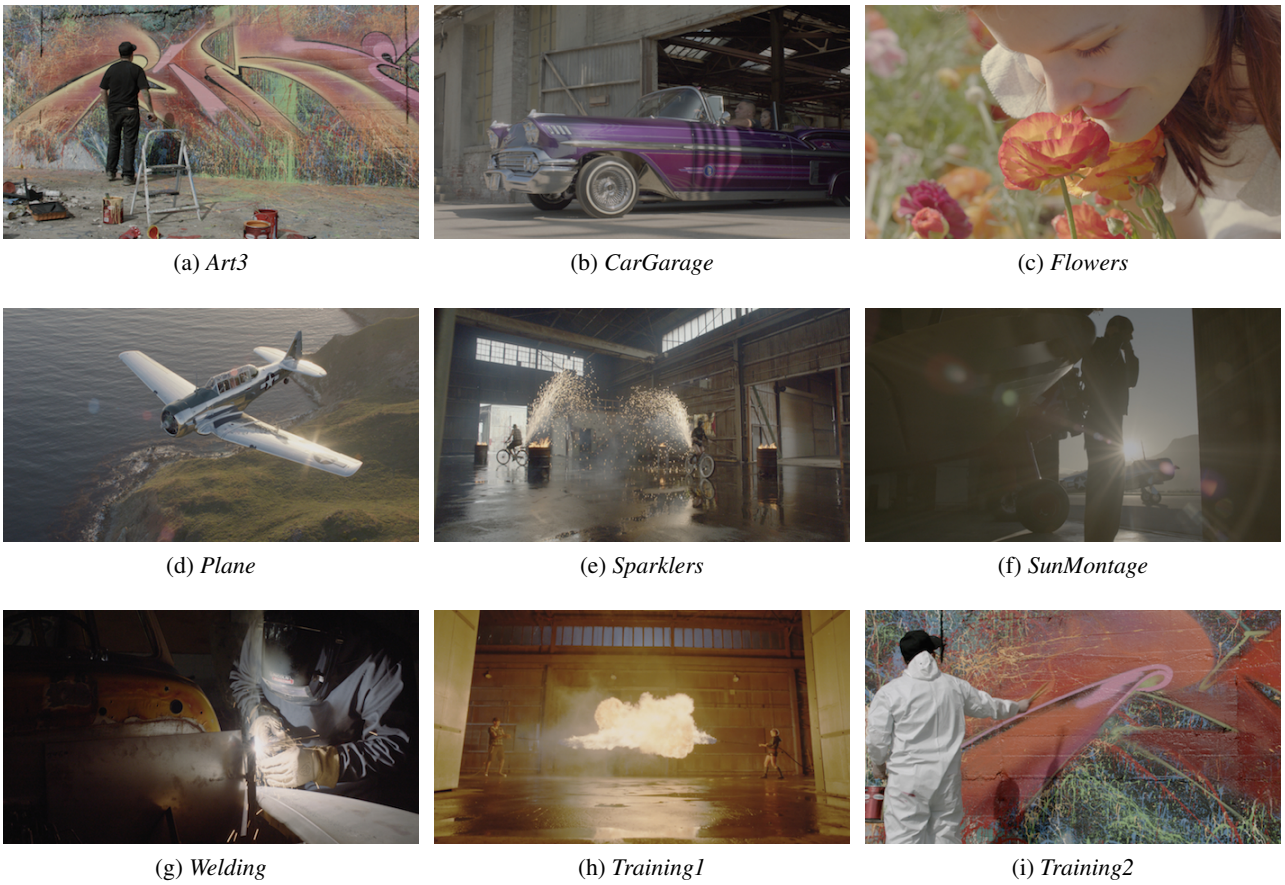
For tone-mapping, an automated proprietary tone-mapping algorithm was used. This algorithm was designed to preserve overall appearance to the input (graded) version and is not intended for image enhancement or to bias importance to specific image regions, as often occurs in a human-guided color grading. The original video sequences were uncompressed, with 12 bits per color, in a domain that has characteristics of gamma and log nonlinearities, as suited for HDR.[20] The combination of high bit-depth and uncompressed video is intended to remove secondary issues of dynamic range effects on needed bit-depth and compression algorithm parameters, since the study's aim was to isolate the question of dynamic range.

2.2 Test environment

The experiments were conducted at the MMSPG test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU-R.[21] The test room is equipped with a controlled lighting system of a 6500 K color temperature. The color of all background walls and curtains in the room is mid grey. The laboratory setup is intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors.

To display the test stimuli, a full HD (1920×1080 pixels) 42" Dolby Research HDR RGB backlight dual modulation display (aka 'Pulsar') was used. The monitor has the following specifications: full Rec. 709 color gamut, $4000\ \mathrm{cd/m^2}$ peak luminance, low black level ($0.005\ \mathrm{cd/m^2}$), 12 bits/color input with accurate and reliable reproduction of color and luminance. In the experiments, the luminance of the background behind the monitor was about $20\ \mathrm{cd/m^2}$. The ambient illumination did not directly reflect off of the display.

In every session, three subjects assessed the displayed test video content simultaneously. They were seated in one row perpendicular to the center of the monitor, at a distance of about 3 times the picture height, as suggested in recommendation ITU-R BT.2022.[22]



(a) *Art3* (b) *CarGarage* (c) *Flowers*

(d) *Plane* (e) *Sparklers* (f) *SunMontage*

(g) *Welding* (h) *Training1* (i) *Training2*

Figure 1: Representative frames of the sequences used in the experiments (a-g) (one additional sequence from "Star Trek: Into Darkness" is not included here due to Copyright) and training session (h-i). The 100 nit versions are shown, since typical displays and printers are unable to reproduce higher dynamic range images.

2.3 Test methodology

The video sequences were presented in pairs in side-by-side fashion to minimize visual working memory limitations. Since only one full HD 1920×1080 HDR monitor was available, each video was cropped to 950×1080 pixels with 20 pixels of black border separating the two sequences.

Subjects were asked to rate the overall quality of pairs of displayed video sequences. To select a score, subjects were instructed to consider such video characteristics as color rendition, quality of the reproduction of skin tones, details of shadows in the scene, contrast and the details of highlights, presentation of light sources appearing in the scene, etc. They were also asked to consider visual discomfort.

Two different subjective tests were performed using two different evaluation methodologies: full paired comparison and forced choice with hidden reference. In this paper, we compare these methodologies and discuss their advantages and drawbacks.

2.3.1 Full paired comparison

In this evaluation, subjects were asked to judge which video sequence in a pair ('left' or 'right') is preferred. The option 'same' was also included to avoid random preference selections. For each of the 8 contents, all the possible combinations of the 4 grades (100, 400, 1000, and 4000 nit) were considered, as well as an extra pair corresponding to 4000 nit vs 4000 nit, leading to a total of $8 \times \left[\binom{4}{2} + 1\right] = 56$ paired comparisons. The comparison of identical video content is useful for side analysis of observer performance and display symmetry.

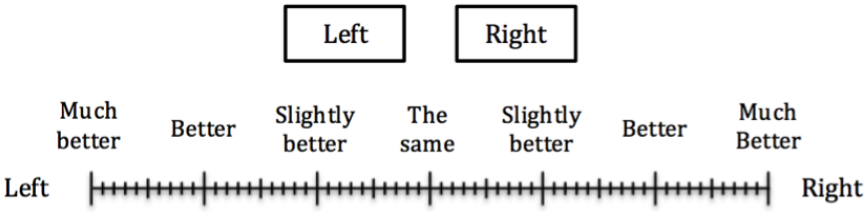


Figure 2: Forced choice scoring scales.

Table 1: Details of the two experiments.

	Full paired comparison	Forced choice
Number of sessions	2	1
Session length	15 minutes	17 minutes
Break length	10 minutes	N/A
Number of subjects (♂/♀)	21 (10/11)	20 (10/10)
Age range (average)	$18-33$ (25.8)	$18-31$ (24.1)

2.3.2 Forced choice with hidden reference

Similarly to full paired comparison, in forced choice evaluation, subjects were asked to judge which video sequence in a pair ('left' or 'right') is preferred, however, the option 'same' was not included and subjects were instructed to randomly select one option when both sequences appear equal.

Instead of considering all possible pairs, as it was done in the full paired comparison, the 4000 nit version was treated as a hidden reference. Therefore, only pairs with the 4000 nit version were evaluated. Again, an extra pair corresponding to 4000 nit vs 4000 nit was included for reliability checks. The hidden reference was shown in every pair, with a randomized position on the screen (either on the left or on the right).

Basically, the forced-choice preference is a binary scale that directly identifies which condition is preferred. Additionally, a horizontal preference scale was used in the evaluation to provide a finer comparison of the two conditions. Figure 2 depicts the scoring scales of forced choice methodology. This methodology was derived from a well-known video comparison methodology[21] with horizontal preference scale being a variant of the continuous comparison scale.

2.4 Test planning

Before the experiments, a consent form was handed to subjects for signature and oral instructions were provided to explain the evaluation task. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts respectively. A training session was organized using the selected video sequences (see Figure 1 for the screenshots) to allow subjects to familiarize with the assessment procedure.

In terms of session planning, the main differences between the two experiments using two methodologies are related to the different number of the evaluated stimuli and observers as summarized in Table 1. To reduce contextual effects, the order in which stimuli were displayed on the screen was randomized differently for each different group of subjects with the same video content never shown consecutively. In the full paired comparison experiment, the test material was also randomly distributed over two test sessions.

3. STATISTICAL ANALYSIS

A typical way to analyze and compare subjective evaluations is to compute mean opinion scores (MOS). However, from both subjective experiments, only preference scale scores can be used to compute MOS values (as the mean across the rates of the valid subjects) and corresponding 95% confidence intervals (CI) directly. Both forced choice and paired comparison methodologies do not provide such scores directly, and hence an estimation of MOS values needs to be computed instead. A variant of the Thurstone Case V model[23] is used in this paper for computation of these values.

Next sections describe in details the outlier detection and the methods for estimation of MOS values from the results of forced choice and paired comparison experiments. Also, four incomplete full paired comparison designs were identified

by taking incomplete subsets of scores from full paired comparison results. These four incomplete designs include pairs with one grade only (referred to as 'pairs with x nit only'). Analysis of these designs is useful for comparisons to single anchor methodologies, e.g., DSIS and DSCQS.[21] Additionally, one incomplete design was analyzed considering pairs with consecutive grades, i.e., 100 vs 400, 400 vs 1000, and 1000 vs 4000. As the result, incomplete designs include the evaluation results for 3 pairs of video content only, as opposed to 6 pairs forming the full paired comparison design.

3.1 Outlier detection

No outlier detection was performed for the scores from the full paired comparison evaluation, since there is no international recommendation or a commonly used outlier detection technique for paired comparison results.

However, the subjective results of the forced choice experiment were processed by first detecting and removing subjects whose preference scale scores appeared to deviate strongly from others. The outlier detection was performed according to the guidelines described in Section 2.3.1 of Annex 2 of recommendation ITU-R BT.500.[21] As the result, one outlier was detected and the corresponding scores were removed from the subjective results.

3.2 Full paired comparison

Before estimating MOS values for paired comparison results, the winning frequency w_{ij} and the tie frequency t_{ij} are computed from the obtained subjective ratings for each pair of stimuli i and j. Note that $t_{ij} = t_{ji}$ and $w_{ij} + w_{ji} + t_{ij} = N$, where N is the number of subjects. This can be done individually for each test video content or jointly over all contents.

3.2.1 Mean opinion scores estimation

The Thurstone Case V model[23] is used to convert the ratings from the ternary scale to continuous-scale quality score values, which are equivalent to mean opinion scores (MOS). For the description of this model, please refer to Ref. 24.

Using winning frequencies w_{ij} and the tie frequencies t_{ij}, count matrix C is constructed, without including the results for the 4000 nit vs 4000 nit pairs. Each element of the count matrix C_{ij} is computed using the following equation:

$$C_{ij} = w_{ij} + \frac{t_{ij}}{2} \tag{1}$$

Hence, C_{ij} represents the number of times stimulus i is preferred over stimulus j, where i and j are the row and column of the matrix. Ties are considered as being half way between the two preference options, i.e., they are distributed equally between C_{ij} and C_{ji}.

The Thurstone Case V model is used to estimate the quality score value μ_i for stimulus i. If only two stimuli, 'A' and 'B', are compared, the quality score values are given by

$$\mu_A - \mu_B = \Phi^{-1}\left(\frac{C_{AB}}{C_{AB} + C_{BA}}\right) \tag{2}$$

where $\Phi^{-1}(x)$ is the inverse cumulative distribution function of the standard normal.

If multiple stimuli are compared, then a maximum likelihood estimation of the quality score values should be performed to consider the interactions between the different pairs.[24] The log-likelihood function is

$$\mathcal{L}\left(\mu|C\right) = \sum_{i,j} C_{ij} \log\left[\Phi\left(\mu_i - \mu_j\right)\right] \tag{3}$$

To find the maximum likelihood solution quality scale values, one must solve

$$\arg\max_{\Delta\mu} \mathcal{L}\left(\Delta\mu|C, \mu\right) \quad \text{subject to} \quad \sum_i \mu_i = 0 \tag{4}$$

To help regularize the estimates, a prior of 1 was added on all the counts, meaning that *a priori* all choices are possible, which corresponds to Laplace smoothing.[24]

3.2.2 Confidence intervals estimation

Assuming that ties convey information about significant differences between two stimuli being compared, Lee *et al.*[25] have proposed a new analysis method to estimate confidence intervals. The original method was proposed for the Bradley-Terry-Luce model but is adapted to the Thurstone Case V model as follows. First, the lower and upper bounds of the count matrix of stimulus i, C_{ij}^{-} and C_{ij}^{+}, are computed as

$$C_{ij}^{-} = w_{ij} \quad C_{ij}^{+} = w_{ij} + t_{ij}, \tag{5}$$

assuming that the ties have been the preferences of stimulus j or i, respectively.

Then, the confidence interval $\left[\mu_i - \Delta\mu_i^{-}, \mu_i + \Delta\mu_i^{+}\right]$ related to the quality score value for stimulus i is estimated based on the Thurstone Case V model. If only two stimuli, 'A' and 'B', are compared, the lower and upper errors, $\Delta\mu^{-}$ and $\Delta\mu^{+}$, are given by

$$\begin{aligned}
\left(\mu_A - \Delta\mu_A^{-}\right) - \left(\mu_B + \Delta\mu_B^{+}\right) &= \Phi^{-1}\left(\frac{C_{AB}^{-}}{C_{AB}^{-} + C_{BA}^{+}}\right) \\
\left(\mu_A + \Delta\mu_A^{+}\right) - \left(\mu_B - \Delta\mu_B^{-}\right) &= \Phi^{-1}\left(\frac{C_{AB}^{+}}{C_{AB}^{+} + C_{BA}^{-}}\right)
\end{aligned} \tag{6}$$

where μ_A and μ_B are the quality score values for stimulus 'A' and 'B', respectively, estimated according to Section 3.2.1.

If multiple stimuli are compared, then a maximum likelihood estimation of the errors is performed. The log-likelihood function is

$$\mathcal{L}\left(\Delta\mu|C,\mu\right) = \sum_{i,j} C_{ij}^{-} \log\left\{\Phi\left[\left(\mu_i - \Delta\mu_i^{-}\right) - \left(\mu_j + \Delta\mu_j^{+}\right)\right]\right\} + \sum_{i,j} C_{ij}^{+} \log\left\{\Phi\left[\left(\mu_i + \Delta\mu_i^{+}\right) - \left(\mu_j - \Delta\mu_j^{-}\right)\right]\right\}, \tag{7}$$

where μ_i is the quality score values for stimulus i estimated according to Section 3.2.1. To find the maximum likelihood solution quality scale values, one must solve

$$\arg\max_{\Delta\mu} \mathcal{L}\left(\Delta\mu|C,\mu\right) \quad \text{subject to} \quad \Delta\mu_i \geq 0 \quad \forall i \tag{8}$$

The proof that the modified definitions of Thurstone's Law for the lower and upper counts (see Eq. (6)) yield the maximum likelihood solution (see Eq. (7)) for two stimuli is obtained by setting the partial derivatives of the objective to zero.

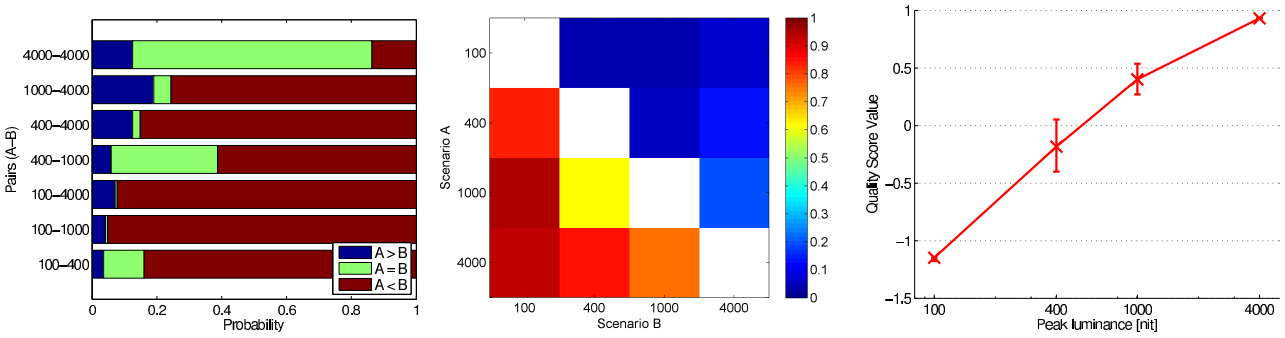
To help regularize the estimates, a prior of 1 was added on all the counts, meaning that *a priori* all choices are possible, which corresponds to Laplace smoothing.[24]

3.3 Forced choice and incomplete designs

For the forced choice with hidden reference experiment or for incomplete paired comparison designs, a different analysis has to be applied to consider the missing results for the pairs that were not evaluated.

Morrissey[26] and Gulliksen[27] have formulated an incomplete matrix solution of Thurstone's Law for the estimation of the quality scores from a subset of paired comparison data. The incomplete matrix solution is formulated as the least squares solution to a system of equations using only the valid data entries, i.e., missing data and pairs with 0 or 1 proportions are ignored. For a more detailed description of this model, please refer to Ref. 24.

Similar to the full paired comparison, the preference matrix is constructed from the winning frequencies. The matrix is incomplete, since it has no entries for the missing scores. The Morrisey-Gulliksen incomplete matrix solution is then used to convert the ratings from the binary/ternary scale to continuous-scale quality score values, based on Thurstone's case V model. When considering a subset of the full paired comparison experiment, ties are equally distributed. Again, a prior of 1 was added on all the valid counts to help regularize the estimates. However, the confidence intervals were not estimated for the incomplete matrix solutions, since there is no analysis method to estimate confidence intervals for an incomplete matrix. The approach to estimate CIs for full paired comparison might not work for incomplete matrix, since it might introduce uncertainty and make reliable estimation of CIs hard.



(a) Preference probabilities

(b) Preference probability of choosing "Scenario A" over "Scenario B"

(c) Quality scores

Figure 3: Full paired comparison: Overall results.

3.4 Correlation between between different designs and scales

To be compliant with the standard procedure for comparing estimated MOS values from different experiments,[28] a regression was fitted to each $\left[MOS^{ExpX}, MOS^{ExpY}\right]$ data set using cubic fitting. The Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SROCC) were computed to estimate accuracy and monotonicity. Since estimated MOS span different range values for different experiments, the computation of root-mean-square error is not possible and was not done. Also, since there was no estimation of confidence intervals for the incomplete paired comparison designs, the outlier ratio was not computed.

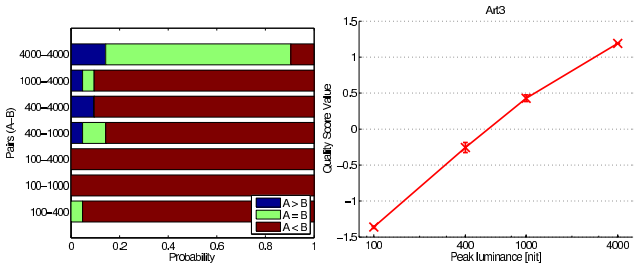
4. RESULTS

4.1 Full paired comparison

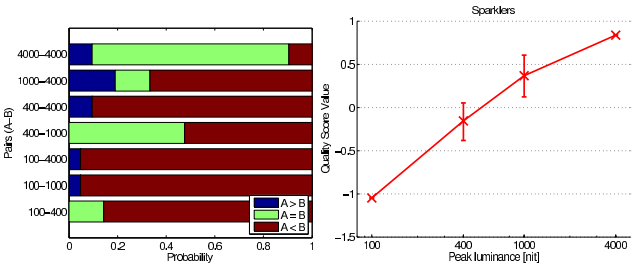
Figure 3 shows the preference probabilities, preference matrix, and estimated MOS values computed over all video contents for the full paired comparison experiment. Figure 4 illustrates the preference probabilities and estimated MOS values computed for selected individual contents. All figures of MOS values demonstrate that the score value increases with the increase in peak luminance. The quality score values tend to increase linearly (with log luminance), although they exhibit a slight concave shape for some contents, which indicates the existence of a saturation level. In most cases, the difference between individual grades is significant, as the confidence intervals do not overlap, except for content *Welding*, where about 70% of the subjects had no preference between the 400 and 1000 nit grades. For this content, 100 nit peak luminance was likely insufficient to represent the high contrast in the original (graded) image between the strong brightness created by the welding sparks and the dark garage, whereas the improvements in the 4000 nit peak luminance version were most likely related to the very strong luminance of the welding sparks.

In most paired comparisons, higher peak luminance was largely preferred and most ties occurred in the 400 nit versus 1000 nit pair. For example, for content *Sparklers*, 11 subjects preferred the 1000 nit grade, whereas the remaining 10 subjects did not express any preference between the two grades. As most ties occurred in the 400 nit versus 1000 nit pair, most of the uncertainty lies between those two grades, which is represented by generally wider confidence intervals associated with the 400 and 1000 nit grades. Note that the 400 and 1000 nit levels are, in fact, closer together on a log luminance scale. The log luminance scale is a likely candidate for a perceptually uniform scale based on Weber's law. So having closer ratings, as well as more ties for these levels, is entirely expected based on a Weber's law model.

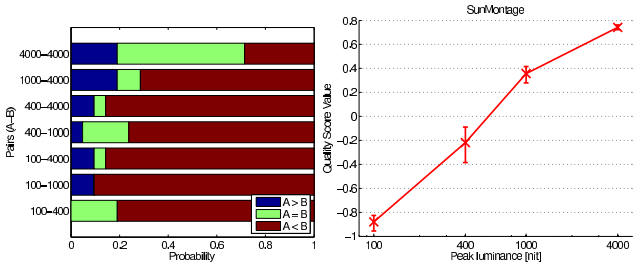
As stated previously, the dynamic range was explored by fixing the black level, and increasing the peak luminance level. Thus, there is some conflation of dynamic range and brightness. Trying to isolate brightness alone, with constant dynamic range, would elevate the black level, and this is already known to cause reductions in preference.[29] On the other hand, trying to isolate dynamic range alone, without changes in brightness, means that the black level must change. It is already known that black level improvements lower than $0.005\ \mathrm{cd/m^2}$ do not lead to preference improvements, except for specific imagery and viewed in total dark environments. The method chosen in this study of fixing the black level and changing the dynamic range via changes in brightness is most relevant to current display technology and viewing conditions.



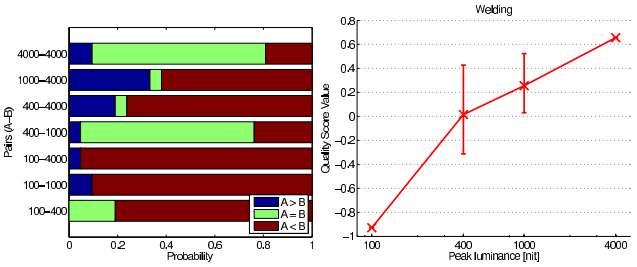
(a) Results for content *Art3*



(b) Results for content *Sparklers*

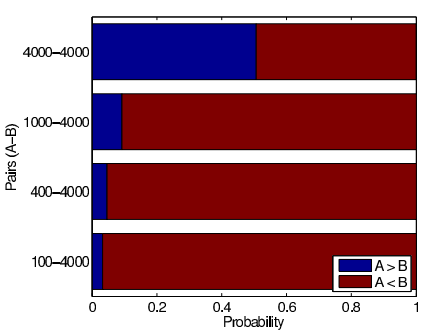


(c) Results for content *SunMontage*

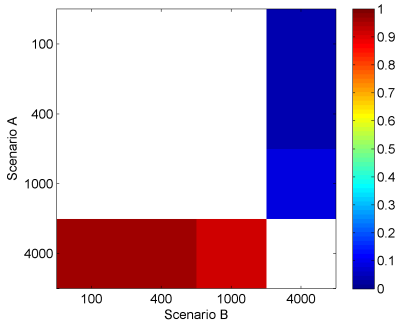


(d) Results for content *Welding*

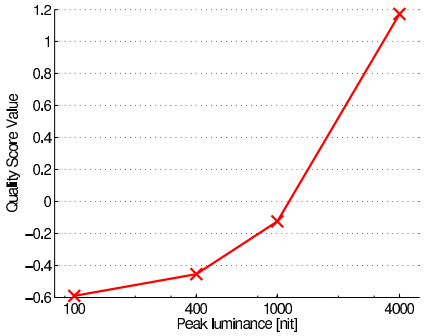
Figure 4: Full paired comparison: Examples of results for individual contents.



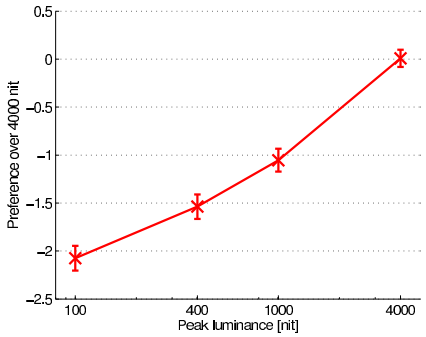
(a) Preference probabilities



(b) Preference probability of choosing "Scenario A" over "Scenario B"

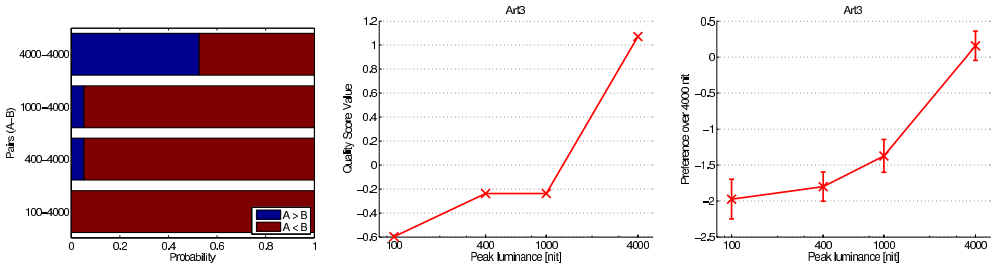


(c) Quality scores

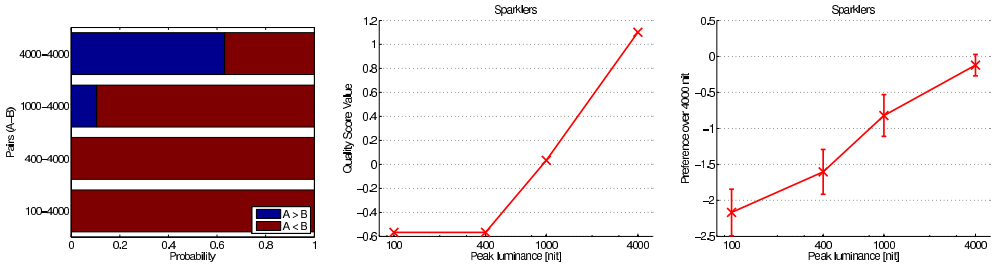


(d) Preference scale scores

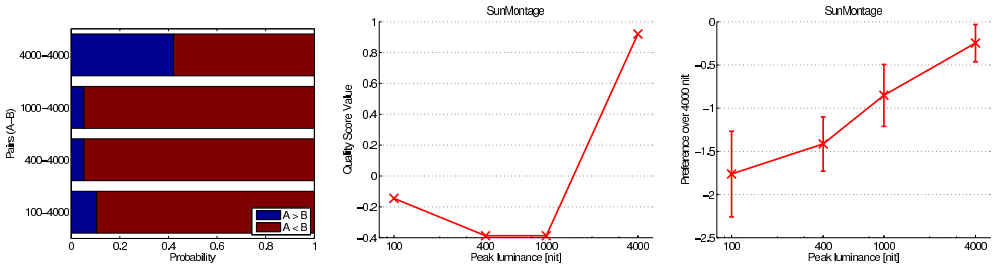
Figure 5: Forced choice: Overall results.



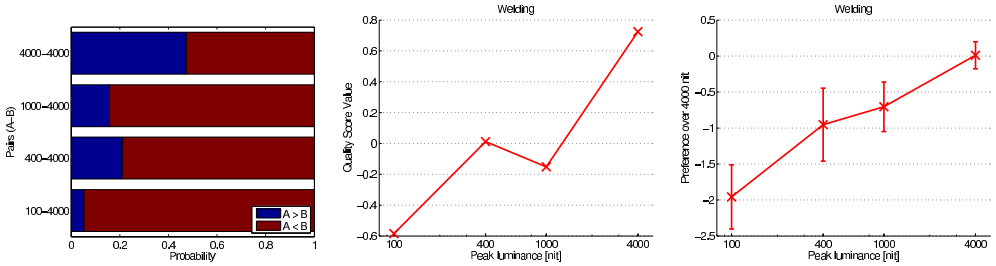
(a) Results for content *Art3*



(b) Results for content *Sparklers*



(c) Results for content *SunMontage*



(d) Results for content *Welding*

Figure 6: Forced choice: Examples of results for individual contents.

4.2 Forced choice with hidden reference

Figure 5 depicts the preference probabilities, preference matrix, estimated MOS values, and preference scale scores computed over all contents for the forced choice with hidden reference experiment. Figure 6 illustrates the preference probabilities, estimated MOS values, and preference scale scores computed for some individual contents. Regarding the plots for the preference scale scores, values -3, -2, and -1 on the y-axis indicate that the hidden reference (4000 nit grade) was judged as 'much better', 'better', and 'slightly better', respectively, than the tested peak luminance level on the x-axis, whereas 0 indicates no preference.

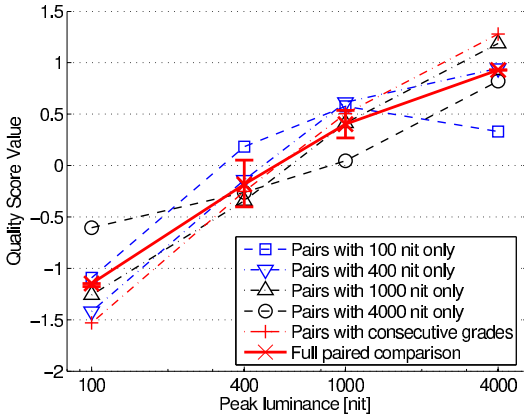


Figure 7: Full paired comparison: Comparison between different designs.

Again, all figures of MOS values demonstrate that the score value increases with the increase in peak luminance. The overall results for the preference scale tend to increase linearly (with log luminance), although they exhibit a slight convexe shape. However, when considering all contents, the quality score values tend to increase exponentially (with log luminance). These results indicate that preference would increase linearly as peak luminance increases, which is contradictory with the findings of the full paired comparison experiment. Specific properties of the human visual system, e.g., the Hunt effect,[30] may explain this behavior in some particular cases. However, for extremely bright monitors (well above $4000\ \mathrm{cd/m^2}$), it is expected that visual discomfort might severely impact the overall QoE.

When considering the overall results for the preference scale, the difference between individual grades is significant, as the confidence intervals do not overlap. However, when considering individual results, the difference between two consecutive peak luminance levels is not significant, as the confidence intervals overlap in most cases, except between 1000 and 4000 nit grades. In general, the confidence intervals are wider when compared to the results for the full paired comparison experiment. However, the estimation of confidence intervals is different between the two experiments. Therefore, one should not conclude that the paired comparison methodology necessarily produces more precise results, although it is easier for the subjects to indicate their preference on a ternary scale than on a continuous scale, as it is difficult to have a clear, unambiguous, and commonly agreed definition of the different levels of the rating scale.

In general, results obtained for the forced choice scale are comparable to those of the corresponding pairs in the full paired comparison experiment. However, the quality scores values estimated from the incomplete design are quite different from those estimated from the full design, as most relations between the different grades were not evaluated. Because of the hidden reference in all pairs, it is impossible to estimate relative scores between the other grades from the binary scale. The preference scale might better represent the relative differences between the other grades, but the hidden reference could act as an upper anchor and influence the score difference between the other grades.

4.3 Comparison between different designs and scales

As stated previously, the quality score values tend to increase exponentially (with log luminance) for the forced choice experiment (see Fig. 5 (c)). However, the scores from the preference scale tend to increase linearly (with log luminance), although they exhibit a slight convexe shape (see Fig. 5 (d)). Figure 7 depicts the difference between the results obtained over all contents for different designs estimated from the full paired comparison. When considering pairs with 4000 nit only, the quality score values exhibit a convex shape, whereas they exhibit a concave shape when considering pairs with 100 nit only. However, the quality score values tend to increase linearly (with log luminance) when considering all pairs, which is somehow a combination of the different trends observed when considering one specific grade as a hidden reference.

These results suggest that complex mechanisms of the human visual system are involved when comparing different brightness levels and that deeper analysis is required to understand what are the significant factors impacting subjects' preference. However, to consider all these factors and their interactions, a full paired comparison design reveals more information than incomplete designs in the estimation of quality score values. Nevertheless, the full paired comparison

Table 2: Pearson linear correlation coefficient.

		Full paired comparison						Forced choice	
		Full	100 nit ref	400 nit ref	1000 nit ref	4000 nit ref	Consecutive	4000 nit ref	Pref scale
Full PC	Full	-	0.9383	0.9852	0.9870	0.9324	0.9917	0.8987	0.9428
	100 nit ref	0.8592	-	0.8912	0.7573	0.5800	0.8338	0.4329	0.6606
	400 nit ref	0.9779	0.9303	-	0.9469	0.8345	0.9718	0.7913	0.8841
	1000 nit ref	0.9823	0.8902	0.9598	-	0.9214	0.9811	0.8924	0.9418
	4000 nit ref	0.8867	0.6611	0.8065	0.8916	-	0.8616	0.9431	0.9328
	Consecutive	0.9928	0.9106	0.9876	0.9891	0.9099	-	0.8651	0.9280
FC	4000 nit ref	0.7951	0.5255	0.6951	0.8418	0.9154	0.7761	-	0.9075
	Pref scale	0.9341	0.8100	0.9117	0.9359	0.9404	0.9214	0.9539	-

Table 3: Spearman rank order correlation coefficient.

		Full paired comparison						Forced choice	
		Full	100 nit ref	400 nit ref	1000 nit ref	4000 nit ref	Consecutive	4000 nit ref	Pref scale
Full PC	Full	-	0.6658	0.9692	0.9930	0.8831	0.9883	0.7980	0.9186
	100 nit ref	0.6673	-	0.6716	0.6601	0.4374	0.6617	0.4399	0.5597
	400 nit ref	0.9692	0.6716	-	0.9593	0.8189	0.9692	0.7177	0.8699
	1000 nit ref	0.9930	0.6601	0.9593	-	0.8721	0.9846	0.8050	0.9146
	4000 nit ref	0.8831	0.4361	0.8216	0.8721	-	0.8622	0.8015	0.8795
	Consecutive	0.9883	0.6586	0.9692	0.9846	0.8622	-	0.7527	0.9014
FC	4000 nit ref	0.7980	0.4382	0.7270	0.8050	0.8015	0.7597	-	0.8258
	Pref scale	0.9186	0.5561	0.8699	0.9146	0.8768	0.9014	0.8303	-

methodology requires more time as more pairs need to be assessed as the number of test conditions increases. To overcome this drawback, carefully designed limited set of pairs can be considered, e.g., using pairs with consecutive grades.

To further investigate the correlation between different designs and scales, the Pearson and Spearman correlation coefficients were computed according to Sec. 3.4. As the mapping of MOS^{ExpX} to MOS^{ExpY} yields slightly different results when compared to mapping of MOS^{ExpY} to MOS^{ExpX}, both mappings are considered and results are reported for both cases. In the following, a value $v(i, j)$ on row i and column j is computed considering mapping of MOS^{Expi} to MOS^{Expj}.

Tables 2 and 3 show Pearson and Spearman correlation coefficients. The tables demonstrate that there is a strong correlation between results of the full paired comparison experiment and results estimated considering pairs with 1000 nit only and pairs with consecutive grades, as the correlation indexes are above 0.98 in both cases. These results show that considering only pairs with consecutive grades could be an alternative to the full paired comparison. On the other hand, considering incomplete designs with pairs of 100 or 4000 nit only have the lowest correlation with the full design. Considering only pairs with 4000 nit, the two experiment produced quite similar results according to the Pearson correlation coefficient (0.91-0.94). However, the Spearman correlation coefficient is slightly lower (0.8). The difference is probably due to individual preferences, as the pool of subjects was different between the two experiments, and the non-linear process involved in the Thurstone Case V model to convert the preference scores to continuous quality scores.

5. CONCLUSION

This paper investigates the added value of higher dynamic range to user preference using forced choice with hidden reference and paired comparison methodologies. Subjective tests were conducted to evaluate the preference between video sequences at four different peak luminance levels, which were displayed side-by-side on a professional reference HDR monitor. The analysis of the subjective results demonstrates that the increase in maximum luminance level at which higher dynamic range video is displayed is preferred by average viewers, with a steady increase in preference as the maximum

luminance increases. The results show a significant increase in the perceptual experience when viewing HDR content at $4000\ \mathrm{cd/m^2}$ peak luminance compared to the current standards in TV and cinema.

The analysis of the results from different subjective experiments demonstrate that a full paired comparison provides more detailed information about viewing preferences. Hence, this methodology is preferred when there is a reasonable number of pairs. In cases when the number of stimuli is too large for full paired comparison, a limited subset of pairs can be considered instead.

ACKNOWLEDGMENTS

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1), EC funded Network of Excellence VideoSense, Swiss SERI project Compression and Evaluation of High Dynamic Range Image and Video, COST IC1005 The digital capture, storage, transmission and display of real-world lighting HDRi, and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET. The authors would like to thank Dolby Laboratories Inc. staff, and in particular Suzanne Farrell, Sherif Gallab, and Timo Kunkel for providing Dolby Research HDR RGB backlight dual modulation display (aka 'Pulsar') and HDR video sequences.

REFERENCES

[1] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., "Photographic Tone Reproduction for Digital Images," *ACM Transactions on Graphics* **21**, 267–276 (July 2002).

[2] Drago, F., Myszkowski, K., Annen, T., and Chiba, N., "Adaptive Logarithmic Mapping For Displaying High Contrast Scenes," *Computer Graphics Forum* **22**(3), 419–426 (2003).

[3] Kuang, J., Johnson, G. M., and Fairchild, M. D., "iCAM06: A Refined Image Appearance Model for HDR Image Rendering," *Journal of Visual Communication and Image Representation* **18**, 406–414 (October 2007).

[4] Mantiuk, R., Daly, S., and Kerofsky, L., "Display Adaptive Tone Mapping," *ACM Transactions on Graphics* **27**, 68:1–68:10 (August 2008).

[5] Ward, G. and Simmons, M., "JPEG-HDR: a backwards-compatible, high dynamic range extension to JPEG," in [*ACM SIGGRAPH 2005 Courses*], *SIGGRAPH'05*, ACM, New York, NY, USA (August 2005).

[6] Liu, J., Hassan, F., and Carletta, J., "Embedding high dynamic range tone mapping in JPEG compression," *Proc. SPIE* **8655**, 86550B–86550B–7 (2013).

[7] Korshunov, P. and Ebrahimi, T., "Context-dependent JPEG backward-compatible high-dynamic range image compression," *Optical Engineering* **52**, 102006–1–102006–11 (September 2013).

[8] Richter, T., "Backwards Compatible Coding of High Dynamic Range Images with JPEG," in [*Data Compression Conference*], 153–160 (March 2013).

[9] Mantiuk, R., Efremov, A., Myszkowski, K., and Seidel, H.-P., "Backward compatible high dynamic range MPEG video compression," in [*ACM SIGGRAPH 2006 Papers*], *SIGGRAPH'06*, 713–723, ACM, New York, NY, USA (August 2006).

[10] Mantiuk, R., Krawczyk, G., Myszkowski, K., and Seidel, H.-P., "High Dynamic Range Image and Video Compression - Fidelity Matching Human Visual Performance," 9–12, IEEE International Conference on Image Processing (2007).

[11] Yoshida, A., Blanz, V., Myszkowski, K., and Seidel, H.-P., "Perceptual Evaluation of Tone Mapping Operators with Real-World Scenes," **5666**, 192–203, Human Vision and Electronic Imaging X, IS&T/SPIE's 17th Annual Symposium on Electronic Imaging (2005).

[12] Chen, M., Qiu, G., Chen, Z., and Wang, C., "JPEG Compatible Coding of High Dynamic Range Imagery using Tone Mapping Operators," in [*25th Picture Coding Symposium (PCS)*], **1**, 22–28, Curran Associates, Inc., Red Hook, NY, USA (April 2006).

[13] Kuang, J., Yamaguchi, H., Liu, C., Johnson, G. M., and Fairchild, M. D., "Evaluating HDR rendering algorithms," *ACM Transactions on Applied Perception* **4**, 9:1–9:27 (July 2007).

[14] Čadík, M., Wimmer, M., Neumann, L., and Artusi, A., "Evaluation of HDR tone mapping methods using essential perceptual attributes," *Computers & Graphics* **32**(3), 330–349 (2008).

[15] Annighöfer, B., Tajbakhsh, T., and Grigat, R.-R., "Prediction of results from subjective evaluation of real-time-capable tone-mapping operators applied to limited high-dynamic-range images," *Journal of Electronic Imaging* **19**, 011015–011015–12 (Januray 2010).

[16] Mai, Z., Doutre, C., Nasiopoulos, P., and Ward, R., "Subjective evaluation of tone-mapping methods on 3D images," in [*17th International Conference on Digital Signal Processing*], 1–6 (July 2011).

[17] Korshunov, P. and Ebrahimi, T., "A JPEG backward-compatible HDR image compression," in [*SPIE Optics and Photonics*], *Applications of Digital Image Processing XXXV* **8499**, 84990J–84990J–12 (August 2012).

[18] Ledda, P., Chalmers, A., Troscianko, T., and Seetzen, H., "Evaluation of tone mapping operators using a High Dynamic Range display," in [*ACM SIGGRAPH 2005 Papers*], *SIGGRAPH'05*, 640–648, ACM, New York, NY, USA (August 2005).

[19] Akyüz, A. O., Fleming, R., Riecke, B. E., Reinhard, E., and Bülthoff, H. H., "Do HDR Displays Support LDR Content?: A Psychophysical Evaluation," *ACM Transactions on Graphics* **26** (July 2007).

[20] Miller, S., Nezamabadi, M., and Daly, S., "Perceptual Signal Coding for More Efficient Usage of Bit Codes," *SMPTE Motion Imaging Journal* **122**(4), 52–59 (2013).

[21] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures." International Telecommunication Union (January 2012).

[22] ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays." International Telecommunication Union (August 2012).

[23] Thurstone, L. L., "A law of comparative judgment.," *Psychological review* **34**(4), 273 (1927).

[24] Tsukida, K. and Gupta, M. R., "How to analyze paired comparison data," Tech. Rep. UWEETR-2011-0004, Department of Electrical Engineering, University of Washington, Seattle, Washington, USA (2011).

[25] Lee, J.-S., Goldmann, L., and Ebrahimi, T., "A New Analysis Method for Paired Comparison and Its Application to 3D Quality Assessment," in [*Proceedings of the 19th ACM International Conference on Multimedia*], 1281–1284 (November 2011).

[26] Morrissey, J. H., "New Method for the Assignment of Psychometric Scale Values from Incomplete Paired Comparisons," *Journal of the Optical Society of America* **45**, 373–378 (May 1955).

[27] Gulliksen, H., "A least squares solution for paired comparisons with incomplete data," *Psychometrika* **21**(2), 125–134 (1956).

[28] ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models." International Telecommunication Union (July 2012).

[29] Seetzen, H., Li, H., Ye, L., Heidrich, W., Whitehead, L., and Ward, G., "Observations of Luminance, Contrast and Amplitude Resolution of Displays," *SID Symposium Digest of Technical Papers* **37**(1), 1229–1233 (2006).

[30] Hunt, R. W. G., "Light and Dark Adaptation and the Perception of Color," *Journal of the Optical Society of America* **42**, 190–199 (March 1952).