Evaluating luminance uniformity metrics using online experiments

Belal Abboushi, Lia Irvin, Eduardo Rodriguez-Feo Bermudez, Michael Royer
Pacific Northwest National Laboratory, Portland, OR, U.S.A.

Corresponding Author: Belal Abboushi, Pacific Northwest National Laboratory, 620 SW, 5th Ave, Suite 810, Portland, OR 97204, USA. E-mail: Belal.Abboushi@pnnl.gov

This is an archival copy of an article published in *LEUKOS*. Please cite as:

Belal Abboushi, Lia Irvin, Eduardo Rodriguez-Feo Bermudez & Michael Royer (2023) Evaluating Luminance Uniformity Metrics Using Online Experiments, *LEUKOS*, 19:3, 308-323, DOI: 10.1080/15502724.2022.2133964

Evaluating luminance uniformity metrics using online experiments

Abstract

Luminance uniformity of luminaires is an important design aspect that can affect perceived discomfort glare, luminaire efficiency, and visual satisfaction. There is, however, a lack of studies that evaluated the performance of different luminance uniformity metrics. This article presents results of two studies where luminance patterns were presented via online questionnaires and subjective ratings of uniformity were collected. Study 1 examined the performance of a uniformity metric based on the human visual system (U_{HVS}) using a priori hypotheses, whereas Study 2 compared U_{HVS} to four other metrics: Max:Min, Avg:Min, entropy uniformity (EU), and coefficient of variation (CV) using correlations and non-linear models. Of the metrics evaluated, U_{HVS} performed best for predicting perceived luminance uniformity. In situations where a tradeoff between metric calculation simplicity and performance is acceptable, the use of CV is recommended.

Keywords: luminaire luminance uniformity, luminance patterns, uniformity metrics.

1. Introduction

Luminaire luminance uniformity (LU) is a characteristic that describes the evenness of luminance across the luminous aperture. In LED luminaires, LU can be influenced by several design factors such as the luminous intensity and photometric distribution of LEDs, the distance between the LED array and optical material, the spacing between LEDs, and the type of optical material (Tashiro *et al.* 2015).

Previous studies showed that uniform luminaires were perceived as less glaring compared to non-uniform luminaires at the same mean source luminance or illuminance at the eye from the source (Tashiro *et al.* 2015; Yang *et al.* 2017a; CIE 2019). To address differences in perceived glare between uniform and non-uniform sources, previous studies proposed different approaches such as accounting for maximum source luminance (Bullough and Hickcox 2012), adding a contrast term to discomfort glare models (Yang *et al.* 2017b), or considering the size and mean luminance for luminous areas above a certain luminance threshold (Kohko *et al.* 2015; CIE 2019).

In addition to potential effects on discomfort glare perception, luminaire LU may affect acceptance or satisfaction ratings. A field study in offices found differences in visual satisfaction ratings by luminaire type; luminaires with prismatic diffusers were associated with lower visual satisfaction compared to luminaires with direct/indirect distribution, which might be related to luminaire LU (Park *et al.* 2021). Another study found that uniform luminaires with maximum luminance of about 10 kcd/m^2 had higher acceptance ratings than non-uniform luminaires with small bright spots and a maximum luminance of about 300 kcd/m^2 (Geerdinck *et al.* 2014). It is unclear, however, if differences in acceptance ratings were due to differences in LU and/or maximum luminance.

Design decisions that alter luminaire LU, such as optical material selection, may impact luminaire efficiency (Gago-Calderon *et al.* 2018; Rozowicz *et al.* 2016; Tashiro *et al.* 2015). For instance, placing an optical material with low transmittance farther away from an LED array could improve LU but is also likely to reduce the overall luminous flux from the luminaire, depending on the scattering behavior of the optical material. Hence, the tradeoffs between luminaire LU and efficiency should be investigated to balance the energy-benefit relationship.

Given the potential effects of LU on perceived discomfort glare, visual satisfaction, and luminaire efficiency, it is important to quantify LU in a way that closely matches human perception. A LU metric can help lighting manufacturers, designers, and users make informed decisions that balance LU and efficiency. Examples include selecting a luminaire with higher efficiency while delivering an acceptable level of uniformity or selecting a luminaire with higher uniformity without compromising efficiency.

1.1 Uniformity metrics

Several metrics have been proposed to quantify LU. The commonly used maximum-to-minimum luminance ratio (Max:Min) and average-to-minimum luminance ratio (Avg:Min) (Ngai 2000; CIE 2020) rely on extreme points, which make them sensitive to changes in luminance image resolution (CIE 2019; Irvin *et al.* 2020). Another potential limitation to these two metrics is that they may not distinguish between luminance patterns with different gradients and rates of change.

A third metric is the coefficient of variation (CV) (Armstrong 1990), which is the ratio of the standard deviation (σ) to the mean ($\bar{x}$) as shown in (1). This means that the entire luminous area is sampled, producing a potentially more stable metric that is less likely to be affected by photometric measurement errors or other anomalies, compared to Max:Min and Avg:Min. Yang *et al*. (2017b) added CV to the unified glare rating (UGR) equation to account for differences in uniformity that can affect perceived discomfort glare. For CV, a lower value implies a more uniform pattern.

$$CV = \frac{\sigma}{\bar{x}} \qquad (1)$$

The recently proposed entropy uniformity (EU) was shown to be exponentially related to perceived uniformity (Yao *et al.* 2017). As shown in (2), EU uses the ratio of i^{th} luminance value to total luminance (p_i) which is calculated for all n luminance points. EU values range between 0 and 1, where EU is equal to 1 when the luminous surface is completely uniform, and 0 when completely non-uniform. EU values range from 0 to 1 with 1 being a very uniform pattern. Yao *et al.* showed that the ranks of EU had a high goodness of fit with perceived uniformity scores (R^2 = 0.96), which was higher than that for CV (R^2 = 0.91).

$$EU = \frac{1}{n} \cdot exp\left(-\sum p_i \, ln(p_i)\right) \qquad (2)$$

The previously discussed metrics Max:Min, Avg:Min, CV, and EU are statistical and do not incorporate a term that accounts for how the human eye processes different contrast levels. The ability of the human eye to perceive complex patterns can be addressed by accounting for the spatial frequency of patterns and related contrast sensitivity (Ashdown 1996). The contrast sensitivity function (CSF) relates the visibility of a spatial pattern to its size and contrast (Dorr et al. 2017). Given that perceived uniformity describes the perceived evenness in luminance, quantifying perceived contrast provides insight into the detectability of luminance variations.

Simonson *et al.* (2003) investigated correlations between CV and preference ratings for different luminance patterns produced by MR16 lamps. These lamps include a multi-faceted ellipsoidal reflector and a small quartz-halogen lamp. They calculated CV based on two luminance data sets. The first data set was from an unprocessed beam image captured with a camera, which was transformed to produce the second data set. The transformation included applying a Fast Fourier Transform, applying a contrast sensitivity function (CSF), and inverting the Fast Fourier Transform. The reason for applying CSF was to produce images that are more representative of how the eye might interpret the patterns. This allowed the correlation coefficients for CV to be improved, compared to CV calculated using the first data set.

Moreno (2010) proposed a metric called uniformity based on the human visual system (U_{HVS}). In (3), non-uniformity based on the human visual system (NU_{HVS}) is calculated by summing the Fourier transform of the luminance pattern $F(\omega_n)$ weighted by the human visual contrast sensitivity function $CSF(\omega_n)$. This is then divided by the quantity of the constant C added to the sum of the Fourier transform. $CSF(\omega_n)$ describes the human eye's sensitivity to luminance contrast as a function of spatial frequency (Barton 1992). Sensitivity increases up to about three cycles (spatial wavelengths) per degree in the visual field and then decreases slowly to ten cycles per degree, where there is little sensitivity to luminance contrast (Moreno 2010). This means that although the presence of high frequencies indicates a less uniform pattern; past a certain frequency, the human eye is less sensitive and therefore less able to discern these photometric differences.

$$NU_{HVS} = \frac{\sum_n F(\omega_n) CSF(\omega_n)}{C + \sum_n F(\omega_n)} \qquad (3)$$

In (4), U_{HVS} is calculated based on NU_{HVS} and CV along with constants k, α, and β. U_{HVS} values range from 0 to 1, with 1 being a very uniform pattern.

$$U_{HVS} = \frac{1}{1 + k \cdot CV^{\alpha} \cdot NU_{HVS}^{\beta}} \qquad (4)$$

The relationship between uniformity metrics and perceived uniformity can be illustrated using data from Yao *et al.* (2017) as shown in Fig. 1. This figure leads to three main observations. First, it shows that under the conditions of Yao *et al.'s* experiment, U_{HVS}, EU, and CV had non-linear relationships with perceived uniformity scores. Second, given that U_{HVS} was the only metric that accounted for contrast sensitivity, it is unclear if that provided U_{HVS} any advantage in distinguishing between patterns with different levels of perceived LU, as compared to other metrics. Yao *et al*. suggested that U_{HVS} performed just as well as EU in quantifying uniformity, but this was based on linear regression models between metric ranking rather than absolute metric values. Lastly, Figure 1 shows some variation in mean uniformity scores for patterns within 0.01 U_{HVS}. Hence, it is also unclear if patterns with similar U_{HVS} values were similarly perceived.

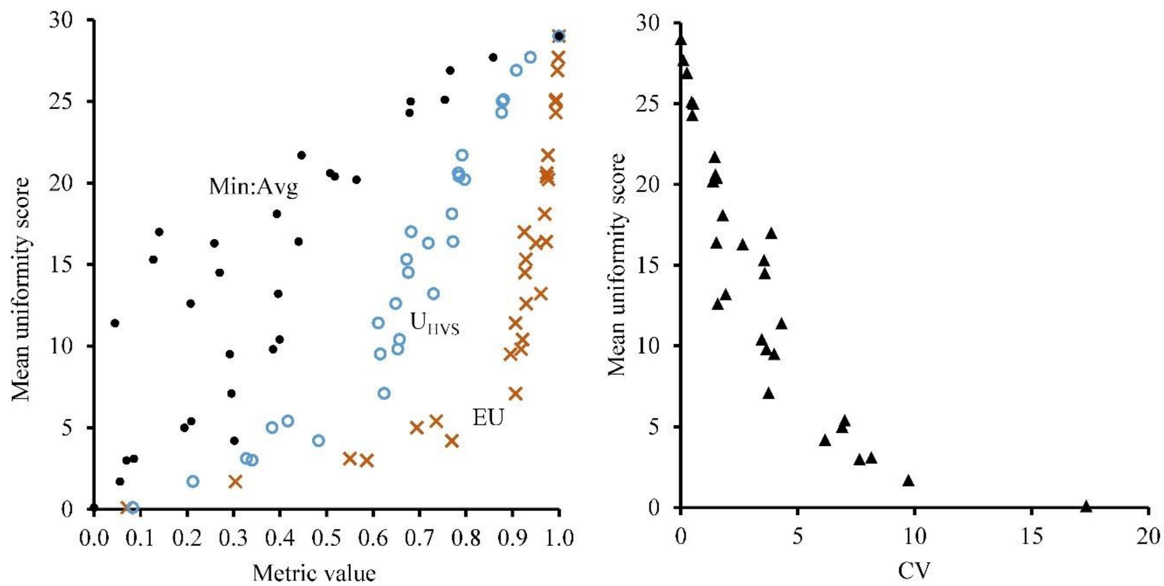

Fig. 1: The graph on the left shows a plot of U_{HVS}, EU, and Min:Avg versus uniformity scores using data published in (Yao *et al.* 2017). The graph on the right shows a plot between CV and uniformity scores.

1.2 Overview of studies

The performance of a uniformity metric can be judged based on its ability to predict and correlate with perceived uniformity ratings. This article presents two studies. In Study 1, simulated patterns with specific differences in U_{HVS} were used to test the ability of U_{HVS} to predict perceived uniformity ratings. In Study 2, the predictions and correlations of U_{HVS} were compared to four other metrics (EU, CV, Max:Min, and Avg:Min) using photographed luminance patterns. Like Study 1, Study 2 also presented the patterns on a computer screen but used a larger number of patterns and a wider range of LU than Study 1. Both studies are presented together in this article to provide a more complete evaluation of the performance of U_{HVS} compared to other LU metrics. The focus on U_{HVS} in these two studies was motivated by reported improvements to LU predictions when incorporating the contrast sensitivity function (Yao *et al.* 2017; Simonson *et al.* 2003; Moreno 2010).

2. Study 1: Evaluating the Ability of U_{HVS} to Match Perceived Uniformity Ratings

This study examined perceived uniformity ratings for patterns with similar U_{HVS} values and patterns with larger differences in U_{HVS}. The analysis and results presented in this article are revised from those previously reported (Abboushi *et al.* 2021). Given that the sensitivity level of U_{HVS} has not been investigated—meaning that it is unclear how the magnitude of differences in U_{HVS} relate to the magnitude of differences in perceived uniformity—we hypothesized that 1) differences smaller than 1% (0.01 U_{HVS}) would not lead to significantly different uniformity ratings; and 2) patterns with a difference in U_{HVS} larger than 1% (0.01 U_{HVS}) would receive significantly different uniformity ratings.

2.1 Method

2.1.1 Stimuli

Eight grayscale luminance patterns were created by keeping the same mean luminance and manipulating the number of modeled point sources, the distance between these sources, and the distance between the sources and modeled diffusing material (Fig. 2). Figure 2 shows the eight simulated patterns and corresponding U_{HVS} metric values. For U_{HVS} calculations, default constant values of $k = 5$, $\alpha = 1$, $\beta = 0.5$, and $C = 1\text{x}10^{-7}$ were used (Moreno 2010).

The patterns were generated such that 1) three pairs consisted of patterns with similar U_{HVS} values within 0.01 (A-B, C-D, E-F); and 2) several patterns could be compared to investigate three levels of differences in U_{HVS}: 0.02, 0.07, and 0.14. Two pairs consisted of patterns with a 0.02 difference in U_{HVS} (E-H and F-H), four pairs had patterns with a 0.07 difference in U_{HVS} (A-C, A-D, C-E, C-F), and two pairs had patterns with a 0.14 difference in U_{HVS} (A-F and B-E).

The patterns were simulated in Python3, primarily using the NumPy and matplotlib libraries (Harris et al. 2020; Hunter 2007). In the simulation, two two-dimensional arrays were created representing LED point sources and the diffusing material. Assuming a cosine distribution from each point source, the vector intensity was calculated at each receiving point on the diffusing plane. Vector intensities that landed outside the diffusing plane, such as those reaching side surfaces, were assumed to reflect inwards towards the diffusing plane with 80% Lambertian reflectance. Lastly, a linear grayscale was applied to the values of the diffusing plane to generate the patterns.

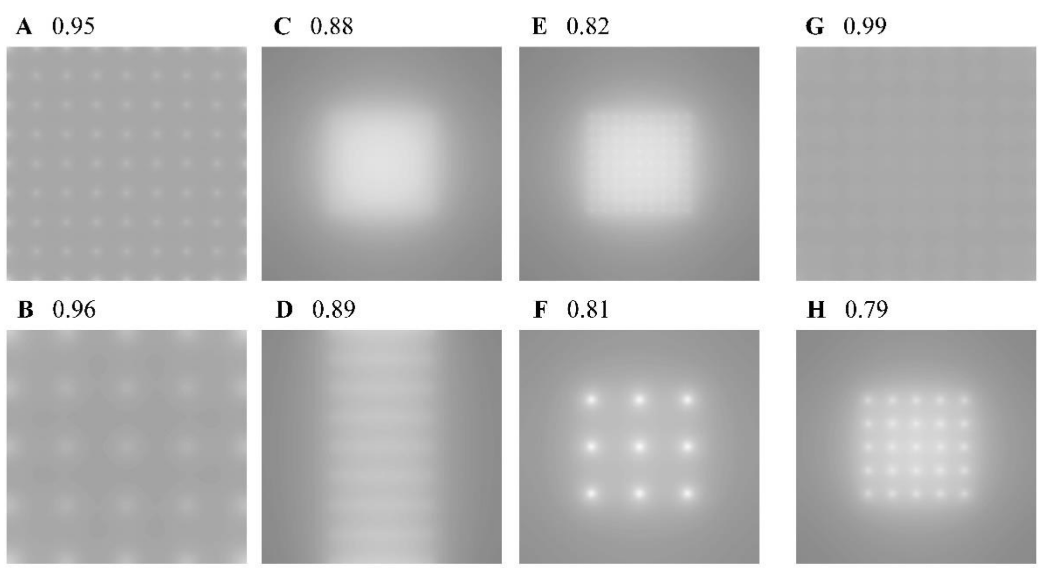

Fig. 2: The eight patterns used in the experiment with corresponding U_{HVS} values.

To calculate U_{HVS}, the ends of the grayscale values (black = 0, white = 255) were assumed to linearly map to 0 and 310 cd/m^2, respectively. These luminance values were based on measurements taken using a calibrated Konica Minolta LS-160 luminance meter placed 0.61 m away from a computer screen (Dell Precision 7730). These measurements were taken within two years of meter calibration (one-point calibration and within tolerance).

2.1.2 Procedure

While conducting online experiments has benefits, it also comes with limitations. For example, different computer screens and internet browsers might have different contrast and brightness settings, ambient illumination may vary among participants' rooms, and computer screen size and resolution cannot be controlled. The procedure used in this study included steps aimed to address and document some of these variables. The perceived uniformity responses were collected using the online platform SurveyMonkey. Duplicate responses from the same participant were prevented without collecting any personally identifiable information. The procedure consisted of the following steps:

(1) To reduce variability in screen sizes and viewing distance, after completing the digital consent form, participants were asked to view the questionnaire on the native laptop or PC screen, not to view the questionnaire using phones or tablets, and to sit an arm's length away from the computer screen in a comfortable position.

(2) Questions were asked about the computer screen make, internet browser, age group, and vision condition. Vision condition was included to eliminate any potential impacts by excluding participants that needed corrective lenses but weren't wearing them, as well as those with a visual disability that could not be corrected.

(3) Participants were not instructed to adjust screen contrast or brightness; instead, a procedure was used to ensure that participants could discern between different gradient levels, similar to the procedure used in previous studies (Villa *et al.* 2013; Sprow *et al.* 2009). Two grayscale gradients (Fig. 3) were shown—one with black background and another with white background—and participants were asked to click on the darkest/brightest bar that they could distinguish from the black/white background.

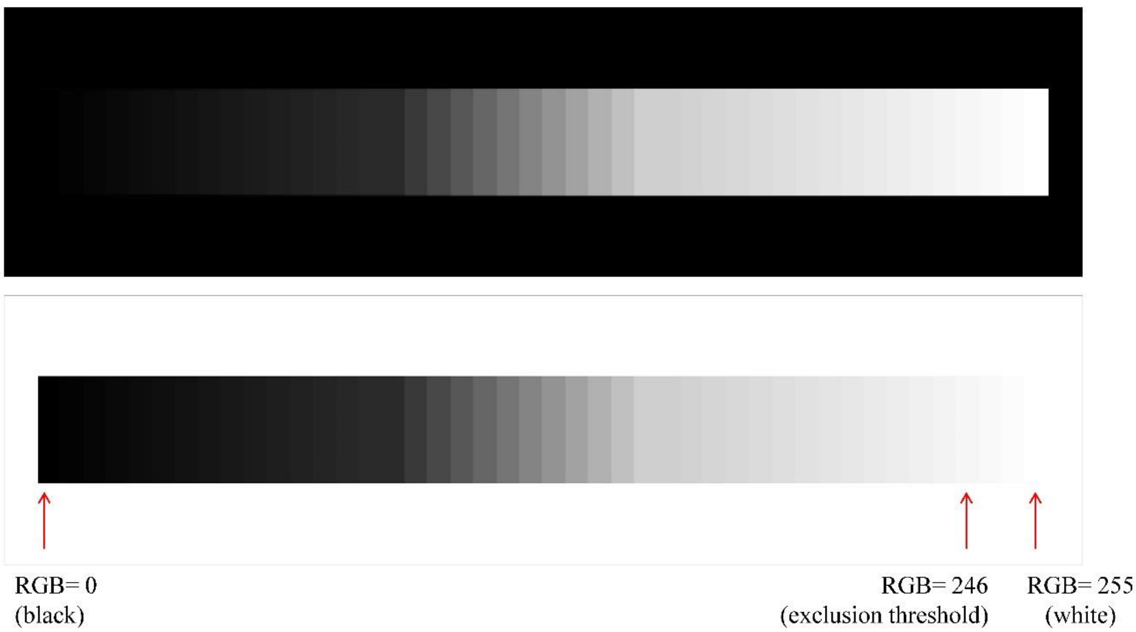

Fig. 3: The gray bars with a black background (top image) and white background (bottom image) that were used to check gradient discernment. The red arrows highlight reference RGB values.

(4) To ensure a consistent viewing size of the patterns across participants, participants were asked to adjust the viewing size (zoom) settings of their internet browser. The questionnaire showed a picture of a driver's license card and asked participants to hold their own license card against the screen while adjusting their browser viewing size to match the size of their actual card. When the browser viewing size was adjusted, the pattern size was approximately 8.4 by 8.1 cm (3.3 by 3.2 inches).

(5) As a pre-trial demonstration, participants were shown a picture of an office space with a luminaire, and informed that in this study they will be presented with side-by-side images of light patterns that could occur on a lighting fixture and asked to select the image that looks more uniform. Lastly, they were provided with a definition of uniformity as “the consistency/evenness of color across the face of the fixture”, and were shown examples of a very uniform and a very non-uniform pattern (see Fig. S6 in supplement 2).

(6) To collect uniformity ratings, a two-alternative forced-choice procedure was used by pairing each pattern with every other pattern resulting in 28 combinations. Additionally, eight null conditions were included by pairing each pattern with itself. Participants were asked to assess the uniformity of the resultant 36 comparisons, responding to the prompt: “Please look at the two light patterns and click on the one that looks more uniform.” The order of pairs was randomized to address order bias, and the left/right position of patterns was counterbalanced across participants to account for potential left/right bias.

2.1.3 Participants

To determine an appropriate sample size, a priori calculations were conducted using G*Power software (Faul *et al.* 2007). Assuming a medium Cohen’s *D* effect size of 0.3, a power of 0.8, and a two-tailed paired *t*-test, a minimum sample size of 90 was required. The sampling frame for this study was restricted to office employees working in one firm to improve the homogeneity of computer screens and laptop make. Participants were recruited using internal social media and information exchange sites. No compensation was provided for participation. This study was approved by the institutional review board at the Pacific Northwest National Laboratory (IRB No: 2020-21). At the beginning of the questionnaire, a consent form was shown to participants and they were asked to click the ‘Next’ button on the questionnaire if they consent to participate.

Responses from 118 participants were collected. Participants with incomplete responses (n = 8), those that needed corrective lenses but were not wearing them (n = 8), those with a visual disability that could not be corrected (n = 1), those that were not able to adjust their screen setting (n = 3), and those that could not distinguish at least the bar with RGB = 246 from the white background (n = 4, see Fig. 3) were excluded. The reason the RGB = 246 was assumed as an exclusion threshold is because responses below that value were outliers (*i.e.*, values that lie beyond the 75^{th} percentile + 1.5 x interquartile range). These criteria resulted in 94 responses that were included in the analysis. Median duration for completing this study was approximately 8 minutes.

Of the 94 participants whose data were included, 66 needed corrective lenses and were wearing them while completing the questionnaire. The majority of participants used a Dell computer screen (n = 62) and the Google Chrome browser (n = 83). The rest of participants used Hewlett Packard (n = 14), Apple Macintosh (n= 8), or other screen makes (n = 10). Few participants used the Firefox (n = 9) or Internet Explorer (n = 2) browsers. Participants’ ages were distributed across different age groups such that 17 participants were within the 18–29 years of age group, 21 were 30–39 years of age, 22 were 40–49 years of age, 20 were 50–59 years of age, 12 were 60–69 years of age, and two were 70–79 years of age. This study did not explore potential effects of computer screen make, internet browser, corrective lens use, or age.

2.2 Results

The mean number of times each pattern was selected as being more uniform are shown in Fig. 4. The number of times each pattern was selected as more uniform is ordinal data, so a Friedman Rank Sum test was used to test whether there were any significant differences between the patterns. The test assumptions of ordinal data and randomized presentation order were met. This test showed a significant difference in uniformity ratings among the eight patterns $X^2(7) = 427.95$, $p < 0.01$. Post hoc comparisons of patterns with similar U_{HVS} (A-B, C-D, E-F) and patterns with different U_{HVS} (A-C, C-E) were conducted using the Wilcoxon Signed-Rank test. Testing five comparisons required adjusting α, using the Holm's method (Holm 1979). In the results below, Holm's corrected α levels were used.

To address the first hypothesis, three pairs with patterns similar in U_{HVS} (A-B, C-D, and E-F) were tested. The Wilcoxon Signed-Rank test showed that uniformity ratings for pattern B were significantly higher than A ($p< 0.01$) and ratings for E were significantly higher than F ($p< 0.01$). There was not a significant difference between patterns C and D. These results do not support the first hypothesis expecting patterns that had similar U_{HVS} values (± 0.01) to not receive different uniformity ratings.

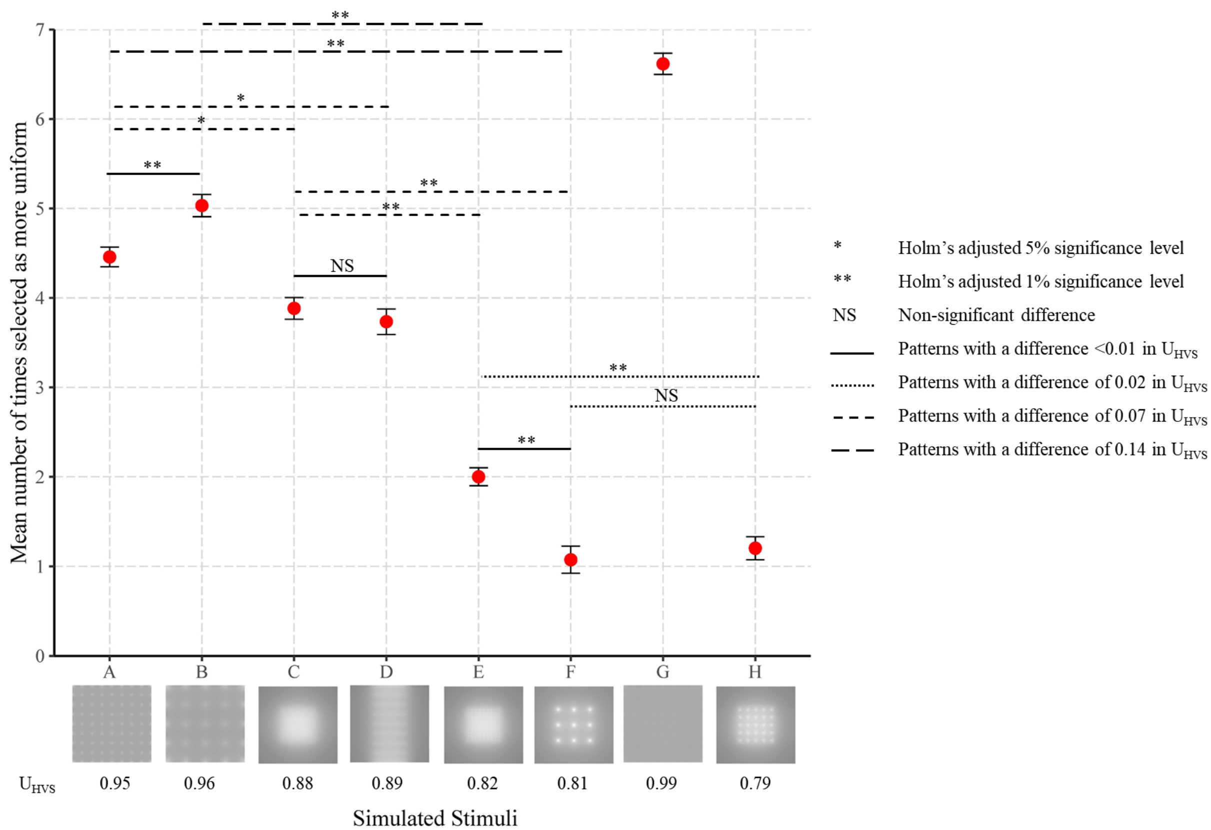

Fig. 4: Perceived uniformity ratings for the eight simulated patterns. The bars show 95% confidence intervals. Self-pairs (null conditions) were not included in this analysis, hence the maximum number of times a pattern can be selected as more uniform was seven.

For the second hypothesis, we evaluated whether differences in U_{HVS} at three levels (0.02, 0.07, 0.14) resulted in significantly different ratings. This includes two pairs with a 0.02 difference in U_{HVS} (E-H and F-H), four pairs with a 0.07 difference in U_{HVS} (A-C, A-D, C-E, C-F), and two pairs with a 0.14 difference in U_{HVS} (A-F and B-E). With 0.07 or 0.14 difference in U_{HVS}, we consistently found significant differences in uniformity ratings. A 0.02 difference in U_{HVS} had mixed results such that one pair received significantly different ratings whereas the other pair did not. These results do not fully support the second hypothesis where we expected that a difference >0.01 in U_{HVS} would consistently result in different ratings. This hypothesis was only true for patterns with a difference of 0.07 or 0.14 in U_{HVS}.

The analysis of null condition pairs (*i.e.*, each pattern paired with itself) examined the percentage of times the pattern on the left and right were selected. Wilcoxon Signed-Rank tests showed no significant differences in any of the null comparisons. There was also not a significant difference in overall left/right choices as indicated by a Friedman test, indicating no significant left/right position bias in the responses.

3. Study 2: Examining Correlations and Predictability of Metrics

This study compared the performance of U_{HVS} to four other metrics (Max:Min, Avg:Min, EU, and CV) using a larger number of patterns, compared to Study 1. This evaluation consisted of examining correlations between these metrics and perceived uniformity ratings, and metric performance using non-linear models.

3.1 Method

3.1.1 Stimuli

The stimuli used in this experiment were images of a 0.6 m by 0.6 m (2 ft by 2 ft) luminaire aperture produced using six different optical materials placed at a distance that ranged from 1.3 to 6.4 cm (0.5 to 2.5 inches) from a 20 by 20 LED array. The aperture of the luminaire was photographed using a Canon 5D Mark II 24 mm DSLR camera with a 17-40 mm lens to create high dynamic range (HDR) images. HDR images capture a wider range of luminance compared to an individual image at a certain shutter speed. This technique ensures that images include the full range of luminance variations. To create HDR images, shutter speeds were varied from 1/3200 seconds to 8 seconds while keeping the aperture at f/11 and ISO at 100. For each stimulus, 15 to 16 images were selected such that in the shortest exposure image every pixel had no RGB values above 228; whereas the longest exposure image was selected such that every pixel had no RGB values below 250. The RGB value of 228 allowed for the highest luminance pixels to be captured while avoiding oversaturation and blooming (Pierson *et al.* 2021). The RGB value of 250 was chosen because it allowed for the lowest luminance pixels to be the only pixels under saturation, ensuring that minimum values were not lost (Irvin *et al.* 2020).

The selected images were combined using *raw2hdr* Radiance tool to produce one HDR image for each material and distance combination (Ward 2011). The generated HDR image in standard RGB (sRGB) is converted to luminance using (5) (Inanici 2006).

$$L = 179 * (0.2127 * R + 0.7152 * G + 0.0722 * B) \qquad (5)$$

Each HDRI was calibrated using a spot luminance measurement, then cropped and tone mapped using a Reinhard02 tone mapper (key value = 0.18, Phi = 1) using Luminance HDR 2.5.1 software (Comida and Anastasia 2017), and converted to grayscale images because the effect of hue on perceived LU was outside the scope of the current investigation. Each grayscale image was a 1056 by 1028 pixel matrix.

To calculate LU metrics, the next step was to convert the grayscale images from values between 0 and 255 to corresponding luminance values as displayed on a computer screen. To account for variability in computer screens among participants, the grayscale to luminance relationship was characterized for three computer screens at two levels of screen brightness (50% and 100%). The three computers were a Dell Precision 5540, a Dell Latitude 7480, and an HP EliteBook x360 1040 G6. These three computer screens were selected based on availability and based on the results of Study 1 where Dell and HP computer screens were the most widely used within the sample. Study 2 used the same sampling frame and recruitment method as Study 1.

We measured the luminance of 21 grayscale levels equally distributed from black (grayscale = 0) to white (grayscale = 255) using the same luminance meter and procedure described in 2.1.1. These luminance measurements were then used to fit three-degree polynomial models as shown in Fig. 5, which were used to convert the grayscale matrices to luminance matrices.

While the grayscale-luminance curve derived from three computers at two brightness levels may not be representative of all participants' computer screens, these data provided an approximation of the uniformity metrics and confirmed that relative differences in uniformity metrics were largely consistent. For example, the HP EliteBook screen at 100% consistently yielded higher U_{HVS} values than Dell Latitude (see supplement 1).

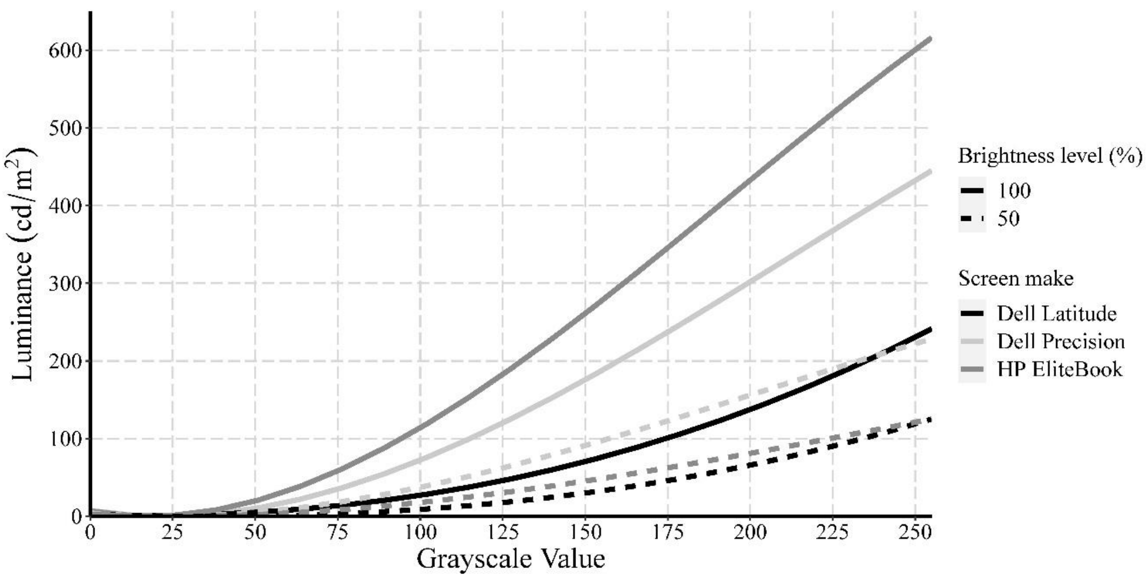

Fig. 5: Fitted lines of three-degree polynomial equations used to convert grayscale values (0 to 255) to luminance [cd/m^2] for the three computer screens at 50% and 100% screen brightness level.

Figure 6 shows the 26 patterns used in Study 2. In this figure, the letters I through N refer to different optical materials, whereas the numbers 1 through 5 represent different distances between the optical material and the LED array ranging from 1.3 to 6.4 cm (0.5 to 2.5 inches). For optical materials M and N, the appearance of the aperture did not visually differ as a function of distance, so not all combinations were presented. The optical material type and the distance were varied to create luminance patterns that differed in LU. Unlike Study 1 where mean luminance was kept the same for all patterns, the imaging procedure in Study 2 did not allow for controlling mean luminance, which varied from 136 cd/m^2 to 285 cd/m^2. Table 1 shows mean metric values for each pattern and mean luminance.

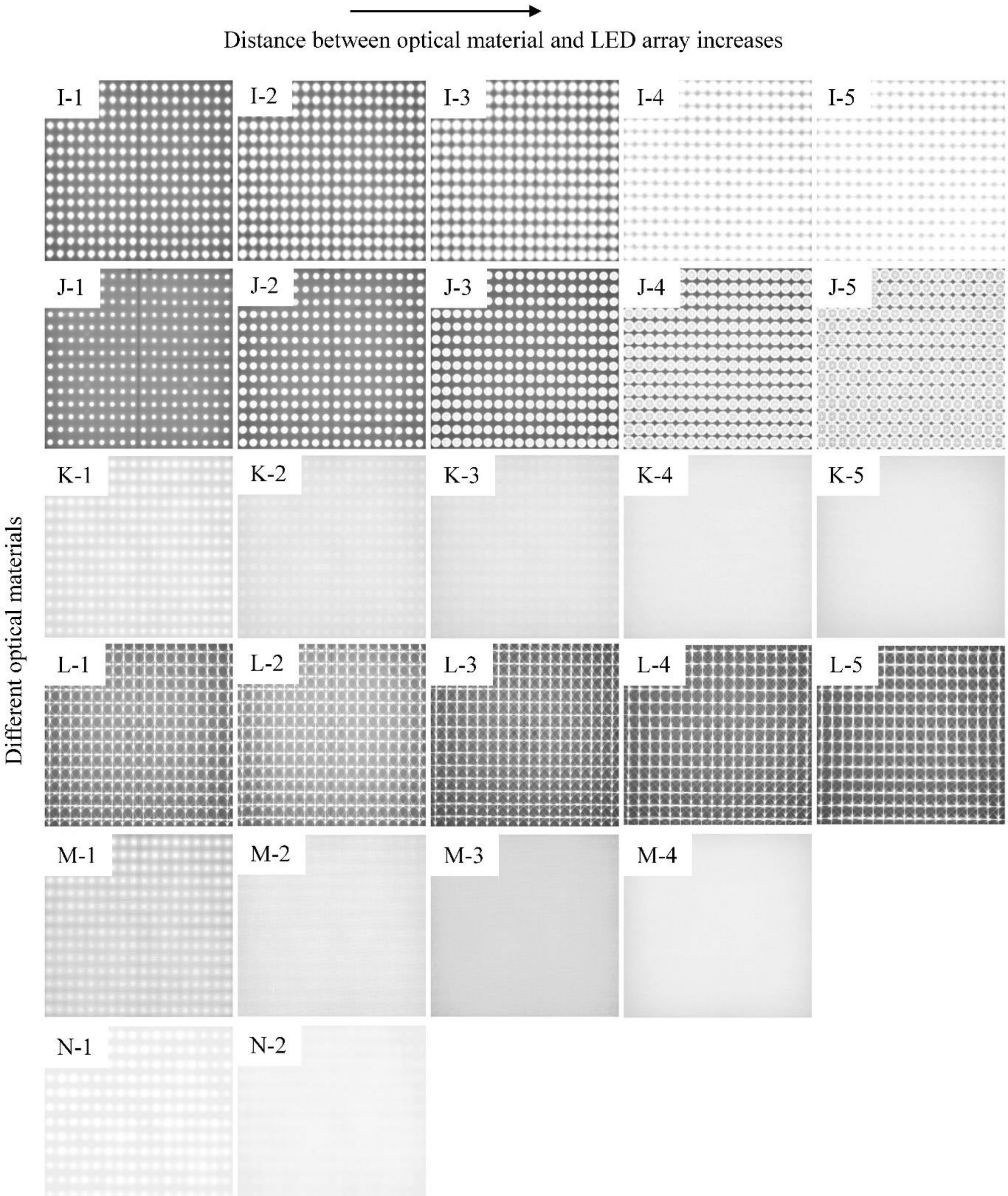

Fig. 6: The 26 luminance patterns used in Study 2. Six optical materials were used labeled I, J, K, L, M, and N; the numbers represent different distances from the LED array for the same material ranging from 1.3 to 6.4 cm (0.5 to 2.5 inches).

Table 1: Luminance uniformity metrics and mean luminance for each pattern, shown as the mean for the six combinations of three screen types x two brightness levels. For each metric, cells with a black or gray shading represent the pattern with lowest or highest uniformity, respectively.

Pattern	Max:Min	Avg:Min	CV	U_{HVS}	EU	Mean luminance (cd/m^2)
I-1	24.80	12.02	0.60	0.38	0.84	157
I-2	22.52	12.98	0.53	0.41	0.86	182
I-3	16.03	10.74	0.43	0.46	0.90	208
I-4	2.79	2.61	0.12	0.77	0.99	280
I-5	2.14	2.04	0.09	0.82	1.00	285
J-1	202.77	82.17	0.55	0.44	0.88	145
J-2	38.92	19.21	0.60	0.43	0.84	160
J-3	41.57	22.81	0.63	0.43	0.81	174
J-4	21.51	15.72	0.39	0.59	0.91	225
J-5	18.07	14.34	0.28	0.69	0.95	243
K-1	2.92	2.48	0.09	0.88	1.00	258
K-2	1.64	1.37	0.07	0.90	1.00	253
K-3	1.77	1.45	0.07	0.90	1.00	249
K-4	1.61	1.38	0.06	0.91	1.00	259
K-5	1.68	1.41	0.07	0.88	1.00	254
L-1	21.51	10.59	0.52	0.43	0.88	160
L-2	13.33	7.79	0.39	0.50	0.92	186
L-3	44.51	18.25	0.67	0.37	0.81	137
L-4	47.81	19.53	0.69	0.36	0.80	136
L-5	28.97	12.04	0.73	0.35	0.78	136
M-1	2.49	1.91	0.13	0.81	0.99	236
M-2	1.63	1.37	0.07	0.90	1.00	256
M-3	2.16	1.60	0.10	0.85	1.00	228
M-4	1.66	1.43	0.06	0.90	1.00	261
N-1	1.20	1.12	0.05	0.89	1.00	279
N-2	1.19	1.10	0.03	0.93	1.00	276

3.1.2 Procedure

The procedure used in this study is the same as that used in Study 1 and described in section 2.1.2 with two exceptions. First, in the pre-trial demonstration, photographed images of a very uniform and a very non-uniform pattern were shown (see Fig. S7 in supplement 2). These were used for relevance to the photographed patterns presented in Study 2. Second, the 26 patterns were each individually presented and participants were asked to use an on-screen slider (controlled by the mouse) to rate the uniformity of each pattern on a scale from 0 (very non-uniform) to 100 (very uniform). The prompt was: “Please rate the uniformity of the image below." The slider starting position was set to start from the middle (value of 50) and could be moved in either direction in 5-point increments. The order by which the patterns were presented was randomized across participants.

3.1.3 Participants

A priori analysis indicated that 44 participants would provide a power of 0.95 for correlations assuming $\alpha = 0.05$, correlation coefficient of hypothesis = 0.7, and correlation coefficient of null hypothesis = 0.3. For recruitment, this study used a similar recruitment method as Study 1. A total of 69 responses were collected. Responses were excluded if they were incomplete ($n = 9$), if the participants needed glasses but were not wearing them ($n = 4$), had a vision condition that could not be corrected with lenses ($n = 1$), were not able to adjust their browser to match the size of driver's license ($n = 1$), or could not distinguish at least RGB = 246 from RGB = 255 ($n = 5$) using the gray bar procedure described in 2.1.2. This resulted in 49 responses that were included in the analysis. Median duration for completing this study was approximately 9 minutes.

Forty-five participants were less than 60 years old. Seven participants were 18–29 years of age, ten were 30–39 years of age, ten were 40–49 years of age, 18 were 50–59 years of age, three were 60–69 years of age, and one was 70–79 years of age. Dell was the most common type of computer screen ($n = 29$), followed by HP ($n = 9$), Mac ($n = 6$), and other types ($n = 5$). The majority of participants used a Google Chrome browser ($n = 46$) and the remaining participants used Firefox ($n = 3$). Regarding vision conditions, 41 participants needed vision correction and were wearing their glasses, and eight did not need vision correction.

3.2 Results

Mean uniformity ratings ranged from 33 to 86 as shown in Fig. 7. The patterns with the lowest uniformity ratings were for patterns L-1 to L-5 whereas the highest uniformity ratings were for patterns K-4, K-5, M-3, and M-4. Out of the 26 patterns, 14 patterns had a uniformity rating between 33 and 43. A visual inspection of the results suggests that patterns with higher uniformity ratings tended to have lower variability compared to lower uniformity patterns, as shown by the smaller 95% confidence intervals.

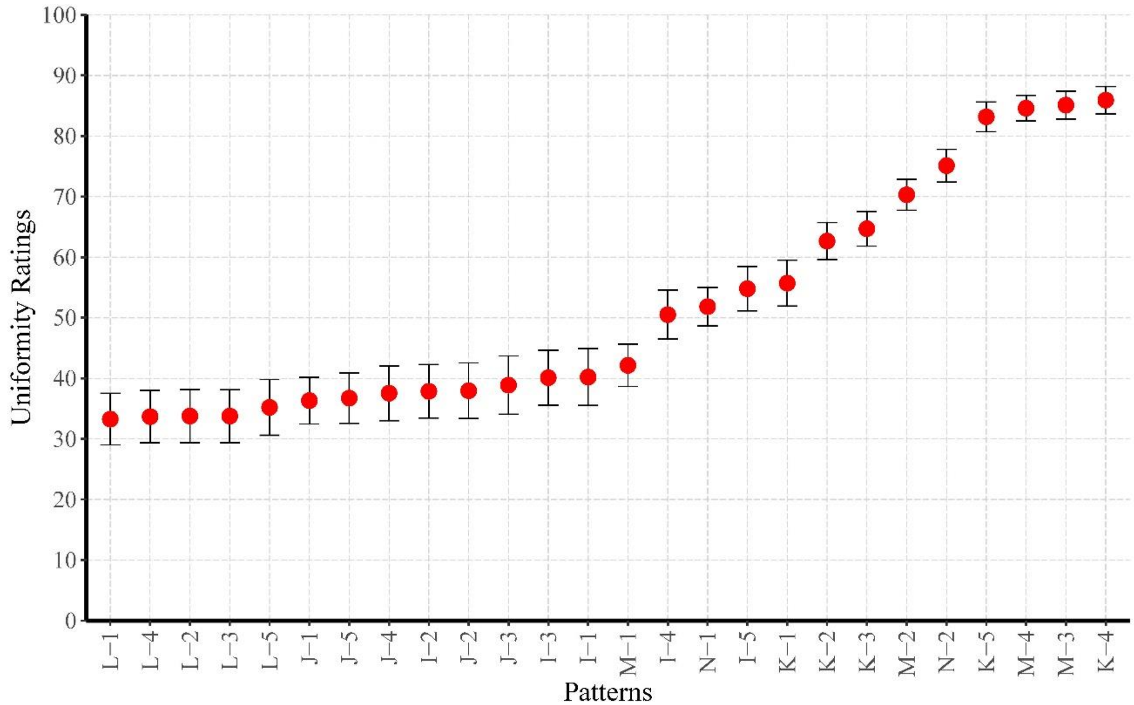

Fig. 7: Mean uniformity ratings with 95% confidence interval on a scale 0 (very non-uniform)-100 (very uniform) sorted from lowest to highest uniform patterns. See Fig. 6 for the patterns.

Initial boxplots showed that pattern J-1 was anomalous because its Max:Min and Avg:Min were beyond the whisker: 75^{th} percentile + 1.5 x interquartile range. Including J-1 in the regression models may affect the fit line, so it was not included in the analysis for both correlations and regression models. Because five metrics were examined, it was necessary to adjust the *p* value to reduce chances of type I error. Like Study 1, Holm's correction, which uses a gradual adjustment to the *p* value, was used.

Spearman correlations and regression models between the logarithm of mean ratings and mean metric values were calculated. Spearman correlations were used because the linearity assumption of Pearson's correlations was not met. The correlation coefficients were significant with a large effect size for the five metrics (Cohen 1988). The lowest coefficient was -0.79 for Avg:Min and the highest coefficient was -0.85 for 1-U_{HVS} as shown in Table 2. CV and 1-EU had the same coefficient of -0.83. Note that U_{HVS} and EU were reversed (1-U_{HVS} and 1-EU) to allow for a power function to be consistently applied to all metrics, as subsequently described. This transformation also maintains smaller metric values being more uniform across the five metrics.

To examine the relationship between the metrics and uniformity ratings, initial plots showed nonlinear relationships that could be represented using power regression models. Regression assumptions that residuals have a mean of zero, normality of residuals, and homoscedasticity were evaluated and met. Regression analyses were conducted using the logarithm of the mean ratings to reduce heteroscedasticity.

Figure 8 shows scatter plots and regression fit lines for each metric. Given that R^2 is not a recommended performance indicator with non-linear models (Kvalseth 1983), Akaike information criteria (AIC)—which can be used for relative comparisons between the nonlinear models (Spiess and Neumeyer 2010)—was used instead. A smaller AIC value indicates a better model, such that a difference in AIC less than 2 suggests that both models are similar, a difference between 4 and 7 indicates that the model with lower AIC is considerably better, and a difference greater than 10 suggests that there is essentially no support for the model with higher AIC (Fabozzi *et al.* 2014).

Table 2 shows AIC values for the power regression models. 1-U_{HVS} had the lowest AIC value with differences in AIC larger than 2 from other metrics, indicating best performance. On the other hand, Max:Min had the highest AIC value with differences larger than 2 compared to 1-U_{HVS}, 1-EU, and CV, and a difference smaller than 2 compared to Avg:Min. These results suggest that Max:Min performed similar to Avg:Min, and that neither of these ratio metrics performed as well as 1-$U_{HVS,}$ 1-EU, or CV.

Table 2: Spearman correlations, regression models, and AIC of regression models for the five uniformity metrics. The regression models were between each uniformity metric and the logarithm of the mean ratings.

	Uniformity metrics				
	1-U_{HVS}	**Max:Min**	**Avg:Min**	**1-EU**	**CV**
Spearman correlation	-0.85**	-0.82**	-0.79**	-0.83**	-0.83**
Regression model	$1.47x^{-0.1}$	$1.87x^{-0.06}$	$1.87x^{-0.07}$	$1.45x^{-0.04}$	$1.49x^{-0.07}$
AIC of regression model	-61.2	-55.1	-55.9	-58.2	-57.4

** represents significance at the 1% level using Holm's- corrected *p* value.

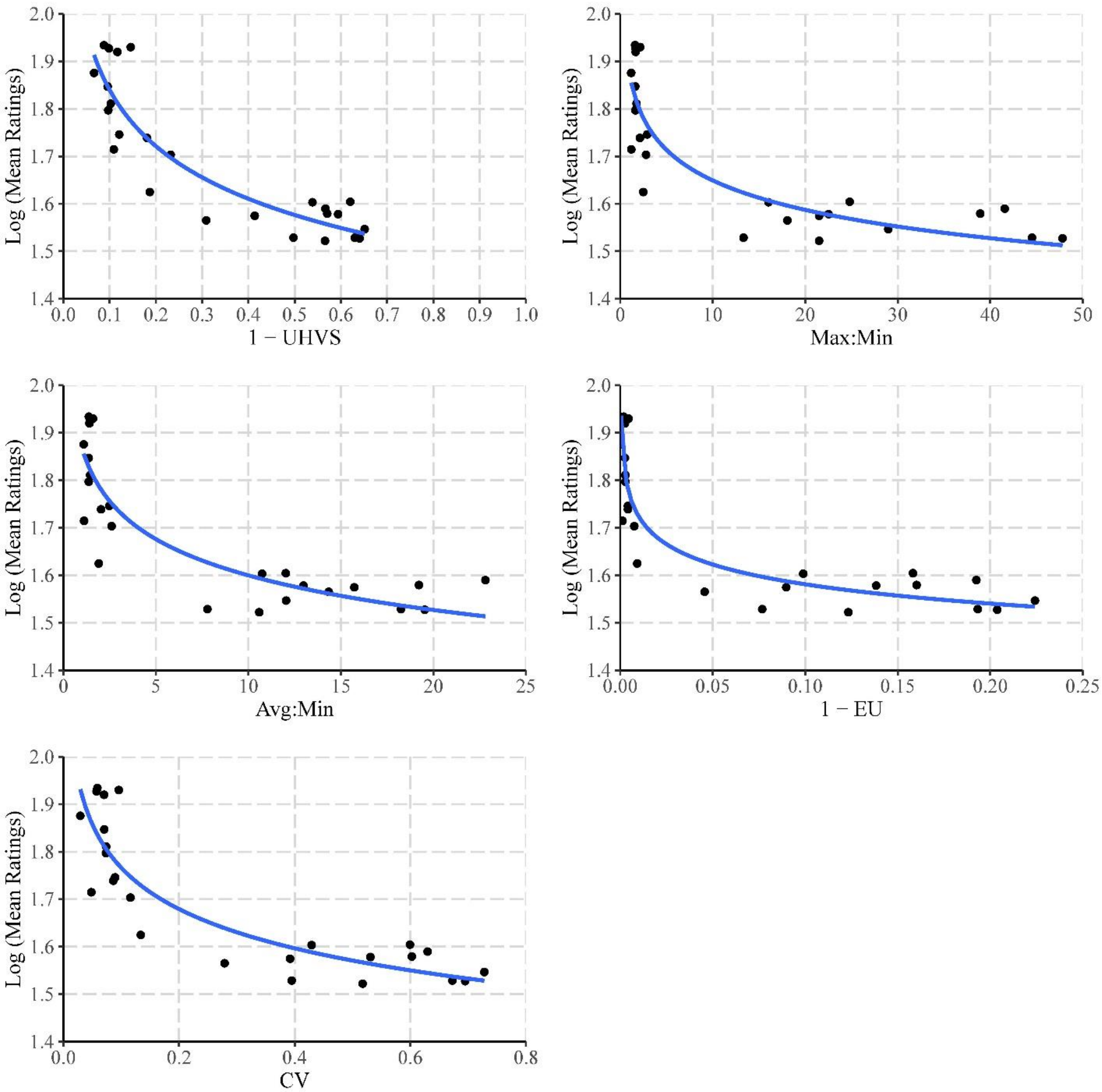

Fig. 8: Scatterplots showing the relationship between different metrics and Log(mean ratings) from Study 2. The continuous blue lines represent power regression fits. For each metric, the left and right side of the graph represent high and low uniformity, respectively.

4. Discussion

In Study 1, pairs of patterns with a U_{HVS} difference of 0.07 or 0.14 consistently received significantly different uniformity ratings. A 0.02 difference in U_{HVS} did not consistently indicate differences in uniformity ratings. On the other hand, two of the three comparisons between patterns with U_{HVS} values within 0.01 (A-B and E-F) received significantly different uniformity ratings. This means that larger differences in U_{HVS} seemed to more reliably predict differences in perceived uniformity, however some patterns with small differences in U_{HVS} were perceived as being different in LU. This reveals an underlying weakness in U_{HVS} that warrants further exploration in future studies.

In Study 2, the Spearman correlations were highest and similar for U_{HVS}, EU, and CV. This is in line with results using data from Yao *et al*.'s (2017) Table 2 where significant Spearman correlation coefficients of 0.98, 0.97, and 0.95 were found for U_{HVS}, EU, and CV, respectively.

The regression models included the reversed U_{HVS} (1-U_{HVS}), the reversed EU (1-EU), Max:Min, Avg:Min, and CV. As mentioned earlier, U_{HVS} and EU were reversed so their values align with the rest of the models: smaller values represent higher uniformity. The differences in AIC values between the models indicated that some metrics performed similarly, such as the pairs CV and 1-EU, CV and Avg:Min, and Max:Min and Avg:Min. On the other hand, 1-EU performed better than Max:Min and Avg:Min. The finding that 1-U_{HVS} was better than CV is in line with results from a previous study that found a slightly higher R^2 for U_{HVS} ranks than CV ranks (Yao *et al.* 2017). We found difference greater than 2 in AIC between 1-U_{HVS} and 1-EU, suggesting a better performance for 1-U_{HVS}. This finding does not align with a result from Yao *et al.* that U_{HVS} and EU were similar in their performance. It is important to note that the linear regression models in Yao *et al.* used the ranking of metrics, not absolute metric values. The use of metric ranks, rather than absolute values, reduces information about differences in metric values; regardless of the difference in metric values two patterns might receive the same ranks.

Fig. 8 shows that all metrics had a non-linear relationship with the logarithm of mean ratings. These non-linear relationships persisted even without applying the logarithm. This non-linearity was also shown for U_{HVS}, EU, and CV using data from Yao et al. (2017) as shown earlier in Fig. 1.

The plots in Fig. 8 also show variability in ratings in the most uniform side of the five metrics. For example, patterns M2 and M4 had a U_{HVS} value of 0.9, but M4 had a higher mean rating (85) than M2 (70). This variability in ratings in the most uniform side of U_{HVS} is in line with that observed in Study 1 for U_{HVS}; some patterns with a similar U_{HVS} (pairs A-B and E-F) received different ratings. This issue was most notable for EU; 13 patterns had EU values between 0.99 and 1, but these patterns had mean ratings that ranged between 42 and 86.

While 1-U_{HVS} had the lowest AIC—indicating best performance—calculating U_{HVS} requires making assumptions about contrast sensitivity, pattern size, and adaptation luminance. Moreno (2010) assumed the adaptation luminance was the mean pattern luminance. Study 1 showed that patterns with a difference of 0.07 or 0.14 in U_{HVS} were consistently perceived differently; but patterns with a difference smaller than 0.01 still received different uniformity ratings. CV, on the other hand, is simpler to calculate because it does not require making detailed assumptions but had higher AIC values compared to U_{HVS}. Results of Study 2 support the use of U_{HVS} to quantify LU, though further work in needed to improve its predictability. In situations where metric calculation simplicity is required, CV can be used.

U_{HVS} or CV can be used in practice to evaluate luminance uniformity. For instance, to evaluate luminance uniformity of a 61 x 61 cm (2 x 2 feet) luminaire from specific viewing points in a room. Because 1-U_{HVS} had the lowest AIC, it means that it can provide better predictions compared to CV. However, the complexity of calculating U_{HVS} may limit its use in practice. The simplicity of a uniformity metric is important if the metric was to be incorporated into discomfort glare models as done using CV in a previous study (Yang *et al.* 2017b).

5. Limitations

The results of the two studies have to be interpreted considering the following limitations. First, in both studies, subjects viewed the luminance patterns on their own computer screens under potentially different lighting conditions. Maximum luminance was likely limited to between 100 cd/m^2 and 600 cd/m^2 compared to likely higher luminance levels from a luminaire. Another inherent limitation in online studies is the variation in computer screen make and model, brightness and contrast settings, resolution, and ambient lighting conditions across participants.

Second, in Study 1, only U_{HVS} was investigated for its ability to indicate similarity or differences in uniformity ratings. This is because the patterns had to be generated with specific metric values.

Third, the patterns used in both studies represent a perpendicular viewing condition. In buildings, the luminaire luminous areas are likely viewed at an angle. This might bring into play other factors such as the angle of view, position of luminance pattern within field of view, and the texture of optical materials. These factors were not explored in the two studies presented.

6. Conclusion

In this article, we explored the performance of different luminance uniformity metrics using simulated and photographed luminance patterns presented via online questionnaires. The metrics that had lowest AIC and highest correlations with mean ratings were $1\text{-}U_{HVS}$, 1-EU, and CV. To predict perceived luminance uniformity, $1\text{-}U_{HVS}$ or CV can be used. The EU metric values congregated within a small range (0.99-1) and this metric did not perform better than CV, hence its use is not recommended.

Future laboratory studies are warranted to investigate the issue found in Study 1: determining if U_{HVS} can be used to infer similarity or differences in perceived uniformity between two patterns. There is also a need for studies to quantify the extent to which the limited luminance range and maximum luminance of the patterns presented on a computer screen affects perceived uniformity ratings compared to viewing luminaire luminance patterns. The U_{HVS} and CV metrics would benefit from further improvements to their predictions for patterns on most uniform side (U_{HVS} or $CV < 0.2$).

Acknowledgements

We would like to acknowledge the help of Sarah Safranek for her help with pattern preparation for Study 2, and Robert Davis for reviewing an earlier draft of this article.

Funding

This work was supported by the U.S. Department of Energy's Lighting R&D Program, part of the Building Technologies Office within the Office of Energy Efficiency and Renewable Energy (EERE).

Disclosure statement

The authors report there are no competing interests to declare

References

Abboushi B, Bermudez ER-F, Irvin L, Feagin B, Royer M, Davis R. 2021. Examining perceptual luminance uniformity of simulated luminaire patterns. In: Kim JK, Krames MR, Strassburg M, editors. Light-Emitting Devices, Materials, and Applications XXV. Vol. 11706. SPIE. p. 115–124. https://doi.org/10.1117/12.2577901.

Armstrong JD. 1990. A new measure of uniformity for lighting installations. J Illum Eng Soc. 19(2):84–89. doi:10.1080/00994480.1990.10747967.

Ashdown I. 1996. Luminance gradients: Photometric analysis and perceptual reproduction. J Illum Eng Soc. 25(1):69–82. doi:10.1080/00994480.1996.10748135.

Barten PGJ. 1992. Physical model for the contrast sensitivity of the human eye. Hum Vision, Vis Process Digit Disp III. 1666(August 1992):57. doi:10.1117/12.135956.

Bullough J, Sweater Hickcox K. 2012. Interactions among Light Source Luminance, Illuminance and Size on Discomfort Glare. SAE Int J Passeng Cars - Mech Syst. 5(1):199–202. doi:https://doi.org/10.4271/2012-01-0269.

CIE. 2019. "Discomfort Caused by Glare from Luminaires with a Non-Uniform Source Luminance (CIE 232:2019)." Vienna (Austria). CIE. https://doi.org/10.25039/TR.232.2019.

CIE 2020. Luminance uniformity. ILV: International Lighting Vocabulary CIE S 017/E:2020. 2nd ed. Vienna,https://cie.co.at/e-ilv

Comida and Anastasia. 2017. "Luminance HDR 2.5.1 – HDR imaging workflow application". http://qtpfsgui.sourceforge.net/

Cohen J. 1988. Statistical power analysis for the behavioural sciences. 2nd ed. Mahwah: Lawrence Erlbaum Associates, publishers.

Dorr M, Lesmes LA, Elze T, Wang H, Lu Z-L, Bex PJ. 2017. Evaluation of the precision of contrast sensitivity function assessment on a tablet device. Sci Rep. 7(1):46706. doi:10.1038/srep46706. https://doi.org/10.1038/srep46706.

Fabozzi F, Focardi S, Rachev S, Arshanapalli B. 2014. Appendix E: Model Selection Criterion: AIC and BIC. In: The Basics of Financial Econometrics. John Wiley & Sons, Ltd. p. 399–403. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118856406.app5.

Faul F, Erdfelder E, Lang AG, Buchner A. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 39(2):175–191. doi:10.3758/BF03193146.

Gago-Calderón A, Hermoso-Orzáez MJ, De Andres-Diaz JR, Redrado-Salvatierra G. 2018. Evaluation of uniformity and glare improvement with low energy efficiency losses in street lighting LED luminaires using laser-sintered polyamide-based diffuse covers. Energies. 11(4). doi:10.3390/en11040816.

Geerdinck LM, Van Gheluwe JR, Vissenberg MCJM. 2014. Discomfort glare perception of non-uniform light sources in an office setting. J Environ Psychol. doi:10.1016/j.jenvp.2014.04.002.

Hu Y, Luo MR, Yang Y. 2015. A study on lighting uniformity for LED smart lighting system. 2015 12th China Int Forum Solid State Light SSLCHINA 2015.:127–130. doi:10.1109/SSLCHINA.2015.7360705.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. Nature. 585(7825):357–362. doi:10.1038/s41586-020-2649-2. https://doi.org/10.1038/s41586-020-2649-2.

Hunter JD. 2007. Matplotlib: A 2D graphics environment. Comput Sci \& Eng. 9(3):90–95. doi:10.1109/MCSE.2007.55.

Irvin L, Feagin B, Bermudez ER-F, Royer M. 2020. Developing High Dynamic Range Imaging Procedures for Luminance Uniformity Measurement. In: Proceedings of the 2020 Illuminating Engineering Society Annual Conference. Virtual: IES.

Inanici M. 2006. Evaluation of high dynamic range photography as a luminance data acquisition system. Light Res Technol. 38(2):123–136. doi:10.1191/1365782806li164oa.

Kohko S, Ayama M, Iwata M, Kyoto N, Toyota T. 2015. Study on evaluation of LED lighting glare in pedestrian zones. J Light Vis Environ. 39:15–25. doi:10.2150/jlve.IEIJ150000566.

Kvalseth T. 1983. Note on the R2 measure of goodness of fit for nonlinear models. Bull Psychon Soc. 21(1):79–80.

Moreno I. 2010. Illumination uniformity assessment based on human vision. Opt Lett. doi:10.1364/ol.35.004030.

Ngai PY. 2000. The relationship between luminance uniformity and brightness perception. J Illum Eng Soc. 29(1):41–50. doi:10.1080/00994480.2000.10748480.

Park J, Loftness V, Aziz A, Wang TH. 2021. Strategies to achieve optimum visual quality for maximum occupant satisfaction: Field study findings in office buildings. Build Environ. 195(November 2020):107458. doi:10.1016/j.buildenv.2020.107458.

Pierson C, Cauwerts C, Bodart M, Wienold J. 2021. Tutorial: Luminance Maps for Daylighting Studies from High Dynamic Range Photography. LEUKOS. 17(2):140–169. doi:10.1080/15502724.2019.1684319. https://doi.org/10.1080/15502724.2019.1684319.

Rozowicz A, Lesko M, Wachta H. 2016. The technical possibilities of losses reduction in the LED optical systems. Proc 2016 IEEE Light Conf Visegr Countries, Lumen V4 2016. doi:10.1109/LUMENV.2016.7745534.

Simonson K, Narendran N, Boyce P, Bierman A. 2003. Development of a metric to quantify beam quality of reflectorized lamps. J Illum Eng Soc. 32(1):63–72. doi:10.1080/00994480.2003.10748405.

Spiess A-N, Neumeyer N. 2010. An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC Pharmacol. 10(1):6. doi:10.1186/1471-2210-10-6.

Sprow I, Zolliker P, Baranczuk Z, Stamm T, Zolliker P. 2009. Web-based psychometric evaluation of image quality. In: Farnand SP, Gaykema F, editors. Image Quality and System Performance VI. Vol. 7242. SPIE. p. 95–106. https://doi.org/10.1117/12.805313.

Tashiro T, Kawanobe S, Kimura-Minoda T, Kohko S, Ishikawa T, Ayama M. 2015. Discomfort glare for white LED light sources with different spatial arrangements. Light Res Technol. 47(3):316–337. doi:10.1177/1477153514532122.

Villa C, Labayrade R. 2013. Validation of an online protocol for assessing the luminous environment. Light Res Technol. doi:10.1177/1477153512450452.

Ward G. 2011. raw2hdr. http://www.anyhere.com/gward/pickup/raw2hdr.tgz.

Yang Y, Ronnier M, Ma SN, Liu XY. 2017a. Assessing glare. Part 1: Comparing uniform and non-uniform LED luminaires. Light Res Technol. 49(2):195–210. doi:10.1177/1477153515607396.

Yang Y, Luo MR, Ma SN. 2017b. Assessing glare. Part 2: Modifying Unified Glare Rating for uniform and non-uniform LED luminaires. Light Res Technol. 49(6):727–742. doi:10.1177/1477153516642622.

Yao Q, Zhong B, Shi Y, Ju J. 2017. Evaluation of Several Different Types of Uniformity Metrics and Their Correlation with Subjective Perceptions. LEUKOS - J Illum Eng Soc North Am. doi:10.1080/15502724.2016.1172488.