Murdoch *et al. Journal of Solid State Lighting* (2015) 2:12
DOI 10.1186/s40539-015-0029-6

Journal of Solid State Lighting
a SpringerOpen Journal

RESEARCH Open Access

Towards perceptual accuracy in 3D visualizations of illuminated indoor environments

Michael J. Murdoch*, Mariska G. M. Stokkermans and Marc Lambooij

* Correspondence:
michael.murdoch@mail.rit.edu
Philips Research Europe, High Tech Campus 34, 5656AE Eindhoven, The Netherlands

Abstract

Through a series of experiments, we have measured the extent to which 3D visualizations of a variety of lighting conditions in an indoor environment can accurately convey primary perceptual attributes. Our goal was to build and rigorously test perceptually accurate visual simulation tooling, which can be valuable in the design, development, and control of complex digital solid-state lighting systems. The experiments included assessments of lighting-related perceptual attributes in a real-world environment and a variety of virtual presentations. Iteratively improving choices in modeling, light simulation, tonemapping, and display led to a robust and honest visualization pipeline that provides a perceptual match of the real world for most perceptual attributes and that is nearly equivalent in perceptual performance to photography. One persistently difficult attribute is scene brightness, as observers consistently overestimate the brightness of dimmed scenes in virtual presentations. In this paper we explain the experimental 3D visualization pipeline variables that were addressed, the perceptual attributes that were measured, and the statistical methods that were applied to evaluate our success.

Keywords: 3D visualization, Simulation, Rendering, Lighting, Atmosphere, Perception, Perceptual accuracy

Background

Human-centered illuminated environments are developed through creative steps, including optics design, luminaire architecture, lighting design, control system optimization, and scene authoring, as well as essential communication with clients and/or suppliers at each stage. All of these steps can become clearer and more concrete with a trustworthy visual preview of the resulting light distribution, in the context of a detailed scene, including the compound effects of multiple light fixtures. Yet, while 3D visualization is ever more common in architectural interior design, it typically remains an artistic tool to convey impressions rather than a simulation tool to convey reality. There is a need now for accurate visual simulation tools designed for lighting – indeed this is true regardless of the lighting technology employed, but it is especially valuable given the design freedom in terms of form factors and light distribution offered by digitally-controlled solid-state lighting. As designers and users are discovering new ways to use flexible light sources, the conceptual value of virtual prototyping becomes clear and reliance on simulation and preview will only increase.

The goal of this paper is to summarize research conducted to assemble and evaluate a 3D visualization pipeline optimized for perceptual accuracy in the presentation of lit indoor environments. A series of experiments was conducted which show that properly-prepared virtual, on-screen presentations of rendered 3D visualizations of a lit environment can result in observers' assessments of relevant perceptual attributes which correspond closely to the assessments made of a similar environment in the real world.

Prior work

Computer graphics has made continuous progress in accuracy and efficiency since the advent in the 1970s of raytracing [1, 2], which essentially simulates the transport and material interactions of photons or rays of light in a three-dimensional virtual scene. The concept is simple, but practical reality is slow and memory-intensive, so while for decades people have understood the problem to be solved – the "rendering equation" [3] – making usable visualization and rendering systems has typically entailed taking practical shortcuts with the lighting simulation. Because of this, lighting has historically been a specialized application of visualization and simulation. RADIANCE [4] for many years stood alone as a lighting-oriented visual simulation tool, used especially successfully for daylighting and fenestration applications. Recently, progress in both processing speed and algorithm efficiency has made physically-based lighting simulation much more mainstream, with the software market including a variety of efficient offline rendering software packages (Indigo, V-Ray, Corona, Octane) and real-time graphics engines for gaming and virtual-reality presentations (Unreal, Unity, Brigade) relying on a mix of rendering algorithms including path tracing and photon mapping. Presently, it is possible to efficiently, physically simulate lighting in indoor illuminated environments at a level of quality where concepts of photorealism and perceptual accuracy can be discussed and measured.

Beyond the physical accuracy of a simulation engine, many other aspects of image creation and presentation affect a viewer's perceptual impression of a scene, notably aspects of display dynamic range and image tone compression. The real world and accurate simulations thereof typically include many log units of intensity range from the darkest shadow to the brightest light source or specular reflection. Such scenes are often described as high dynamic range (HDR). The real world is HDR and the human visual system is very good at adapting to different intensity levels both within a scene and over time. However, electronic displays, prints, and other media are unable to fully show HDR intensity ranges. With these media, compression of image intensities is required, regardless of whether the image source is general photography or a lighting simulation. Tonemapping operators (TMOs) are algorithms or systems that accomplish intensity range compression: some are designed to mimic human adaptation; others simply behave like camera autoexposure systems.

A number of researchers have assessed the fidelity of virtual representations of real spaces. Drago and Myszkowski made direct comparisons between a real architectural interior viewed through an aperture and renderings shown on a display [5]. They found that the perceived fidelity of photographs was best, nearly approached by renderings carefully crafted by an artist to match the scene. Realizing the importance of display dynamic range, they noted that the limitations in displayed highlight and shadow regions were the weakest aspects of their images. In research published by Villa et al., the effects of light simulation engine, TMO, and post-processing corrections were studied using renderings of artificial lighting in interior scenes [6, 7]. Overall they found a good match between real and virtual scenes, though they observed that TMO had a strong effect on the perceived realism of displayed images. Recently, Schielke measured good correlations between real and virtual scenes for brand image studies that included visual characteristics spanning lighting (brightness, uniformity, etc.) and marketing (price, style, attractiveness, etc.). His paper does not emphasize the details of image preparation and tonemapping, but apparently he excluded high-luminance light sources and specular reflections in the presentations while using relatively low-luminance projection and web-based uncalibrated displays [8]. Other authors have conducted visual simulation-based lighting research. Newsham et al. used a simulated office space in an experiment that employed a genetic algorithm to optimize the attractiveness of the luminance distribution on scene surfaces [9].

We have undertaken an in-depth study of the effects of relevant simulation creation and image presentation variables on perceptual accuracy, comparing perceptual results acquired in a real environment with those acquired using virtual environments. Some parts of this corpus have been presented previously [10–13]. The present paper brings this whole set of experiments together with new insights and includes a meta-analysis of the importance of all 3D visualization pipeline variables studied. It is a detailed extension of the summary presented at the SID/IES Special Lighting Track, SID 2015 [14].

Methods

The methods employed in this research included direct assessment of perceptual attributes of artificially illuminated indoor scenes, both real-world baseline environments and rendered virtual presentations. Through a series of experiments using virtual presentations, key components of a 3D visualization pipeline were varied in order to determine their influence on perceptual accuracy – the similarity between the perception of the virtual stimuli and the real-world stimuli – in order to create an optimal pipeline. The perceptual attributes of interest, the procedures for the real-world and virtual experiments, the creation of virtual stimuli, and the statistical methods are described below.

Perceptual attributes

There are many perceptual attributes that are relevant to lighting research. Of primary interest to us are those such as brightness and uniformity, which relate closely to physical quantities, and atmosphere metrics such as coziness and tenseness, which are affective evaluations of an environment that are strongly influenced by lighting changes. Atmosphere terms and assessments were originally outlined by Vogels [15], and their relationship to lighting characteristics were studied by Vogels, Seuntiens, and others [16, 17]. Other, secondary perceptual attributes relevant to lighting may include color rendition, glare, and appropriateness for a task or space, but these were considered out of scope for this stage of research.

In the present experiments, ten primary perceptual attributes were studied: Overall Pleasantness, Overall Brightness, Overall Diffuseness, Contrast, Uniformity, Shadow

Visibility, Coziness, Liveliness, Tenseness, and Detachment (Businesslike). This set of attributes can be thought of as three distinct groups. One is overall impressions – Pleasantness and Brightness; a second is perceived atmosphere – Coziness, Liveliness, Tenseness, and Detachment – terms that Vogels found to be orthogonal descriptors via factor analysis. The third group has to do with the uniformity of light distribution – Diffuseness, Contrast, Uniformity, and Shadow Visibility. These distribution attributes presumably correlate to some extent, but without enough a priori knowledge to select a representative single term or subset they were all used. Our assertion is that if observers' responses to these perceptual attributes in virtual presentations match those in analogous real-world environments, then we have accomplished perceptual accuracy. Our research track has focused on quantifying how the perceptual accuracy is affected by the choices made in the creation and presentation of visualizations of lighting scenes.

Baseline real-world experiment

A baseline experiment was conducted in a real-world artificially illuminated environment in order to uncover ground truth lighting perception data over a range of lighting conditions. The real environment employed was our Light Lab, which is an otherwise-typical office room with a flexible, computer-controlled lighting system including 60x60cm variable color temperature fluorescent luminaires, halogen spot lights oriented as downlights, halogen and RGB LED spot lights oriented toward one wall, and RGB LED grazing luminaires at the bottom of the opposing wall. A plan view of the Light Lab can be seen in Fig. 1. For our series of experiments, we defined 15 distinct lighting conditions using different groupings of fluorescent lights and two groups of halogen lights, downlights (large spots) and wall-oriented spotlights (small spots), at different intensity levels, details of which are given in Table 1. A representative subset of these conditions is shown in the rendered images in Fig. 2.

In the baseline experiment, 28 observers viewed all 15 lighting conditions in random order from a seated position near one wall. Actually, the experiment was run in two separate events, with 12 and 16 observers in each. We found no statistical differences between these populations so the data are combined as if done in a single experiment. Using a MATLAB GUI questionnaire on a laptop screen, observers were asked to assess the ten perceptual attributes. The questionnaire said "Please rate the scene on the

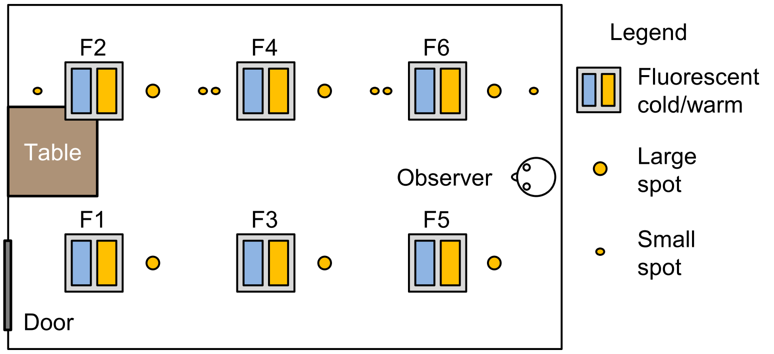

Fig. 1 Light Lab layout. This top-view of the Light Lab shows the locations of 6 fluorescent luminaires, 6 large spotlights/downlights, and 6 small wall-oriented spotlights. The observer sat at the right end of the room facing the table and door

Table 1 Light Lab conditions

			Luminaire intensity (%)							
Category	Condition	Color temp	F1	F2	F3	F4	F5	F6	Sm. spot	Lg. spot
Warm	1	2700K	-	-	100	100	-	-	-	-
	2		100	-	-	100	100	-	-	-
	3		100	100	-	-	100	100	-	-
	4		100	100	100	100	100	100	-	-
Cool	5	6500K	-	-	100	100	-	-	-	-
	6		100	-	-	100	100	-	-	-
	7		100	100	-	-	100	100	-	-
	8		100	100	100	100	100	100	-	-
Dimmed	9	2700K	50	50	50	50	50	50	-	-
	10		5	5	5	5	5	5	-	-
	11	6500K	50	50	50	50	50	50	-	-
	12		5	5	5	5	5	5	-	-
Spots	13	3000K	-	-	-	-	-	-	100	-
	14		-	-	-	-	-	-	-	100
	15		-	-	-	-	-	-	-	5

Luminaire color temperature and intensity settings for the 15 light conditions which fit into categories of warm, cool, dimmed, and spot. Light fixtures used are the variable color temperature fluorescent fixtures (F1-F6), wall-oriented halogen spotlights (Sm. spots), and halogen downlights (Lg. spots)

following aspects:" then listed each perceptual attribute alongside a 7-point scale using radio buttons. The leftmost button was labeled "Low," the middle button "Neutral," and the rightmost button "High." Mean opinion scores (MOS) for each of the ten perceptual attributes over each of the 15 light conditions are illustrated in Fig. 3. Rather than to try to explain the trends visible, this data-rich figure is provided simply to illustrate that this set of relatively simple light conditions indeed provides a strong effect on all of the perceptual attributes being measured (in fact light condition has a significant effect on every attribute according to analysis of variance (ANOVA) results). The MOS

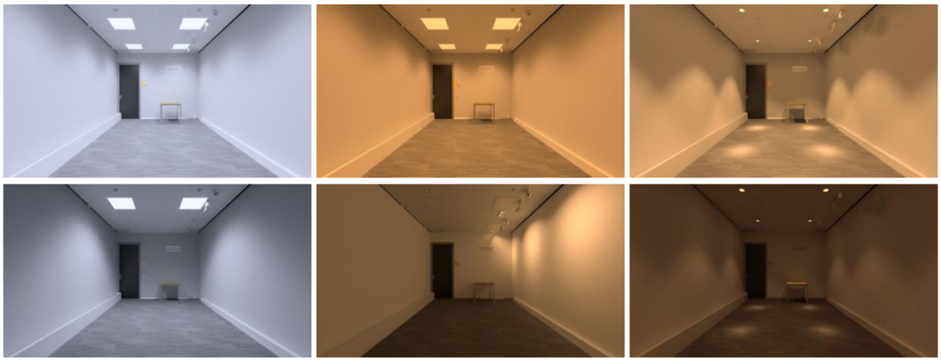

Fig. 2 Light Lab images. Shown are rendered images of six of the fifteen light conditions. *Top row*, L-R are conditions 8 (6 cool 100 %), 4 (6 warm 100 %), 14 (large spots 100 %); *bottom row*, L-R are 5 (2 cool), 13 (small spots), and 15 (large spots 5 %)

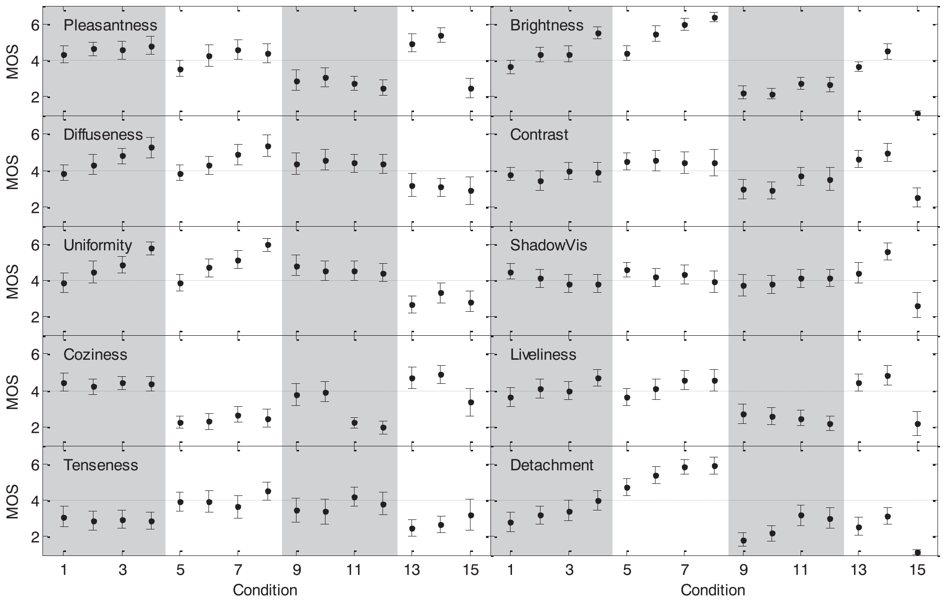

Fig. 3 Baseline MOS. Ten separate plots are shown, each showing the mean opinion scores (MOS) over all 28 observers versus lighting condition for each perceptual attribute in the real-room baseline experiment, shown with 95 % confidence intervals. Shaded regions indicate condition categories: warm (conditions 1–4), cool (5–8), dimmed (9–12), and spots (13–15)

illustrated comprise the ground truth that we aim to match with virtual presentations created through our 3D visualization pipeline.

Virtual presentation experiments

The baseline experiment was followed by a series of experiments that mimicked the baseline experiment but used only virtual stimuli – rendered visualizations of the Light Lab in the various lighting conditions presented on displays. In all of the virtual experiments, observers viewed randomized, displayed images of the virtual Light Lab and used the same second-screen MATLAB GUI that was used in the baseline experiment, thus making assessments of each of the perceptual attributes on a 7-point scale. Each separate experiment included about 20 observers who each assessed either 45 or 60 image stimuli: the same 15 lighting conditions through three or four different presentations. Our 3D visualization pipeline, explained in the following section, was tested in parts over time, not with a full-factorial experimental design, resulting in iterative improvements along the way. Full details of the presentations, observers, etc., are explained below and summarized in Table 2.

3D visualization pipeline

The series of steps used in creating 3D visualizations, or renderings, can be thought of as a production or graphics pipeline, as shown in Fig. 4. In the first block of our 3D visualization pipeline, a scene is created including models of its geometric, material, and lighting system characteristics. Geometry may originate from an architectural drawing or 3D model, be inferred from a laser scan of a real space, or be manually built up from graphics primitives. Material models include a model of the bi-directional reflectance distribution function (BRDF) of surfaces in the scene, generally with texture

Table 2 Virtual presentations

Presentation	Experiment	N observers	Model	Renderer	TMO	Display	View
A	1	25	Orig	Ind	Rein	Ph42	Static
B					iCAM		
C				LT	Rein		
D					iCAM		
E	2	24	Orig	Ind	Rein	Ph42	Static
F					Rein	3D	
G					Rein	HB	
H					Lin	HB	
I	3*	16	Orig	Ind	Rein	Ph42	Static
J	4	17	Orig	Ind	Rein	NEC46	Static
K			Photo	-	-		
L	5	12	Update	Ind	Rein	NEC46	Static
M	6	24**	Update	Ind	Rein2	Laptop	Static
N							Pano
O						NEC46	Static
P							Pano
Q						Proj	Static
R							Pano
S	7	19	Update	Ind	Rein2	NEC46	Static
T			V-Ray	V-Ray	Rein2	NEC46	Static

Each row represents a different virtual presentation (letter codes), each of which is a combination of 3D visualization pipeline components including model, renderer, TMO, display, and view. Also noted are the experimental groupings and number of observers for each presentation

*Experiment 3 also included within-subject assessment of the real-world environment

**24 total observers, but with balanced incomplete block design, each of 6 presentations was viewed by 16 observers

and bump maps, which may be taken from available libraries or captured with photography and reflectance measurements. Lighting details include models of the geometry, technology, light distribution, color, and control limitations of light sources in the scene, where light distribution can be specified via measured photometry or simulated by including optical elements in the geometric model. The formats and limitations of material and lighting models are necessarily linked to the renderer in the next block.

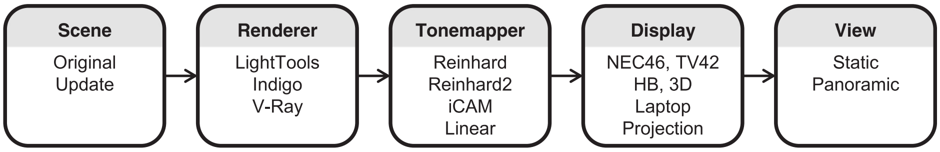

Fig. 4 3D visualization pipeline. This flow diagram is a framework for describing the steps involved in creating and using rendered visual simulations of lighting environments. A scene is created by describing its geometric, material, and lighting system characteristics; a rendering engine models light transport in the scene and a virtual camera and TMO result in an image; the image is reproduced on a display or print; then it is viewed in a specific viewing condition for a specific purpose

In the Renderer block, the scene is rendered, in the computer graphics sense, by a rendering engine that simulates physical light transport in the scene and results in a high-dynamic range (HDR) simulation of the scene from the viewpoint of a virtual camera. The third block, Tonemapper, executes rendering in the photographic sense, where a tonemapping operator (TMO) renders the scene-referred image into a displayable image by compressing and limiting the dynamic range of scene intensities. In the Display block, the image is reproduced on a specific display, including display-specific image processing and color management, and finally in the View block, the visualization is presented to human eyes in a specific viewing condition for a specific purpose. Differences in purpose may conceptually lead to different requirements for perceptual accuracy: for two examples, using a calibrated display in a controlled lab for visual threshold testing of uniformity differences, or using a conference room projector for an optics design review. We undertook the goal to understand how choices of components in each step of the 3D visualization pipeline affect the perceptual accuracy of the resulting images.

Pipeline variations

In the following subsections, the experimental variations are presented in relation to the visualization pipeline, rather than as a chronology of experiments. All of the presentations, main variables, experiment numbers (chronological numbers), and numbers of participants are listed in the following Table 2. Parenthetical abbreviations are explained in the following subsections and are consistently used in the results tables following.

Most experiments discussed herein were approved by the Internal Committee for Biomedical Experiments (ICBE) of Philips Research, though a few early studies predate the ICBE's involvement in non-medical studies. In every case, observers were given an informed consent sheet explaining their task, summarizing the research goals, promising confidentiality of their personal information, and clarifying that they were free to leave the test at any time. Observers were all employees (including student interns) of Philips Research, and none were lighting experts. They were all tested for normal color and spatial vision. There were a few observers who participated in more than one experiment, but in general there was not much overlap. Thus, we treated all intra-experiment presentations as within-observer, and all inter-experiment presentations as between-observer in our analyses.

Scene, renderer, and photograph

An obvious starting point is in the 3D scene model itself, and a comparison between renderings and photographs of the real-world scene. The 3D scene and the simulation of light transport within carry huge leverage over the later pipeline components. Several of our experimental variations included changes in the 3D model and the selection of light simulation rendering engine. In one experiment (see [11]) we included photographs of the 15 lighting conditions as a variation (Photo), with the hypothesis that a photo of the real situation would provide an upper limit to the accuracy attainable via synthetic renderings. Photographs were taken at a fixed exposure setting, but manually adjusted in final brightness as discussed in the following section Tonemapper.

The model of the Light Lab was built and improved over time. An early experiment, described in [10], included the original 3D scene model (Orig), built manually based on physical measurements made in the real Light Lab. This experiment included a comparison of two renderers: LightTools (LT), an optics design and simulation package, and Indigo Renderer (Ind), a path-tracing-based general-purpose rendering engine. For both we kept the scene materials as simple as possible, with diffuse Lambertian wall materials and semi-gloss painted elements modeled with Phong BRDF models. Texture maps used for the carpet, wall sign, tabletop, and ceiling tile models were created from photographs of real scene elements. A later variation involved the scene model ported to V-Ray renderer, a highly-efficient general-purpose engine popular in architectural visualization among other applications, to test if its high computational efficiency comes at an accuracy cost relative to the strictly physically-based path-tracing approach of Indigo.

In all renderings, the fluorescent luminaires' light distribution was modeled using measured goniometry in IES format, despite the IES model assuming a point source and our scene clearly including a 60 × 60 cm extended source. We accepted this discrepancy because any error it would cause would be primarily on objects very close to the light sources, and we have none. In the LT and Ind renderings, the spot lights' distributions were modeled by simulating, or ray-tracing, the reflective luminaire optics in the rendering, while in V-Ray these were also modeled using IES data. In some later experiments, updates of the model's luminaires and wall reflectance values were made based on improved measurements of scene reflectances, and the carpet was replaced (with one similar in lightness but with a different pattern) both in the real lab and virtual model (Update) (see [11]).

Tonemapper

The tonemapping operator (TMO) is a critical element which compresses the tonal range of an inherently HDR render output to a displayable image. We selected two TMOs from literature to use in our research and created a third. One is the TMO of Reinhard et al. (Rein) [18], which we found very robust and flexible in early testing, and which Villa [7] also concluded was a particularly good performer. Based on the photographic techniques of American photographer Ansel Adams, it provides an adjustable key parameter for overall image intensity, much like a camera exposure compensation control. The TMO employs logarithmic tone compression, which is similar to the power-law "gamma" compression familiar in photographic systems for lower exposure values but more aggressively compressive at higher exposure values. This allows a very wide dynamic range to be compressed, which keeps highlights from clipping but occasionally results in noticeably low-contrast highlights. We found the Reinhard TMO's key parameter always required separate manual adjustments for each lighting condition, despite our testing of the methods described in Reinhard's own addendum to automate it [19]. In the creation of experiment stimuli, we accomplished this visually, walking from the real-world Light Lab to the display lab and choosing the key parameter level that matched best in our opinion. For the last two experiments updates in the Reinhard TMO key parameter settings were made based on our observations that dimmed lighting conditions were over-estimated in brightness (Rein2).

The second TMO – iCAM06 (iCAM), presented by Kuang, Johnson, and Fairchild – uses a perceptual approach, accounting for visual adaptation through image color appearance modeling [20]. Its success in adapting to both the luminance level and the chromaticity of the virtual scene also turned out to be its weakness for our application, as described later in the Discussion. Additionally we employed simple linear reproduction, basically a scale factor applied to scene luminance values, as a TMO in one unique high-brightness display presentation (Lin). An experiment that employed Rein and iCAM was described in [9], and the Lin TMO was used in the experiment presented in [10].

Related to tonemapping, we did adjust the overall image intensity of the photographs, using the Brightness modifier in Adobe Lightroom, to match the intensity of the corresponding Reinhard-tonemapped renderings. Thus both the tonemapped renderings and the brightness-adjusted photographs were manually made to visually match the corresponding real world lighting conditions, according to the authors' consensus.

Display type and size

Several types and sizes of displays were used in the experiments, all viewed in a display lab with dimmed indirect lighting on the wall behind the display and no direct light on the display surface. Most of the variations involved one of two 1920×1080 TV-sized LCDs: a 42-inch Philips 42PFL9703 with a peak white of 230 cd/m^2 (Ph42) and a 46-inch NEC P462 with peak of 278 cd/m^2 (NEC46), a commercial display with robust calibration options. One or both of these displays was used in every experiment. In all cases, displays were thoroughly measured and images were processed separately for each display to account for its specific tone and color characteristics, resulting in device-specific RGB images used in the experiments.

In some variations, advanced displays were employed, such as a 47-in. 3D stereo passive-polarized LCD TV (a prototype similar to Philips 47PFL7696) of 397 cd/m^2 without polarized glasses (3D), and a 42-inch high-brightness VHBLCD display of 1800 cd/m^2 (HB). In a later experiment (Experiment 6 in Table 2) exploring image size, a small-screen HP 8460p laptop with 14-inch 1366×768 LCD with peak white of 217 cd/m^2 (Laptop) and a large 138-inch diagonal projection from a Sanyo PLC-ZM5000L 3-LCD projector with 1920×1200 pixels and a peak white of 180 cd/m^2 (Proj) were used. Further detail on this experiment can be found in [12].

Field of view and interactivity

Presentation was varied in field of view (FOV) and interactivity. The FOV is the width in degrees of a displayed image from the point of view of the observer. Perfect perspective is attained when the FOV of the [virtual] camera matches the FOV of the display from the viewer's position. In most of our experiments, we employed static images with a camera FOV slightly wider than our displayed FOV (Static), which means each image looks noticeably "wide-angle" with converging lines and exaggerated depth. We chose this imperfection because matched FOV was possible only with tradeoffs: either by sitting very close to the 46-inch display, which had the side-effect of making its pixels visible, or by narrowing the FOV of the camera, which meant that the ceiling in the Light Lab was not visible on the screen – hardly appropriate for evaluating lighting settings.

Because in the real room baseline experiment the observer was free to look around the room with his or her head from the fixed, seated viewing position, we adopted a similar viewing mode in some presentations. Starting with a cube-map (six cube-face renderings from a single viewpoint covering the full spherical view) panoramic rendering of the room, an interactive viewing mode was implemented which allowed the observer to look around the room, keeping a natural FOV and mimicking the head movement which was possible in the real room (Pano). This can be thought of as an intermediate between a static view of a scene and a walk-through, where the latter would require a real-time graphics engine to render new viewpoints as needed. Our Pano presentation takes advantage of the quality of offline rendering but still gives the observer some freedom to look around. Details of the experiment with FOV and interactive panorama viewing are described in [12].

Statistical analysis

Analyses were performed within each experiment to uncover the effects of the variables under study, some aspects of which were presented in earlier papers, but they do not easily assess accuracy across experiments. In this paper we present a meta-analysis that was performed over all experiments to assess the perceptual accuracy of all virtual presentations. The meta-analysis employed linear mixed models (LMM) [21], a mean comparison statistical hypothesis test similar to analysis of variance (ANOVA) but which uses maximum likelihood rather than least-squares fitting and which is robust to our combination of mixed between- and within-subject and incomplete designs. LMM provides an estimate of statistical significance for multiple independent variables' effects on a single dependent variable. Thus, separate LMMs were computed using SPSS software for each perceptual attribute with a model including lighting condition and virtual presentation as fixed factors and observer as a random factor. In all analyses we set our significance level to 0.05. Our goal remains to draw a conclusion about a match between the perception of the real environment and the virtual presentations, and we realize that both LMM and ANOVA are not completely suitable for this – they provide evidence by which a null hypothesis (that is, the assertion that there is no difference) can be rejected, but they do not specifically prove that a null hypothesis is true.

With this in mind, we additionally look at Cohen's d effect size to assess the relative magnitude of the differences observed [22]. Cohen's d effect size is calculated by dividing the mean difference of two conditions by the standard deviation (STD). Rules of thumb show that an effect size of 0.2 is considered to be small, an effect size of 0.5 to be moderate, and an effect size > 0.8 is large. Corresponding to a small Cohen's d effect size, we calculated a threshold in DMOS by multiplying 0.2 by the average STD over all attributes and light conditions in the baseline experiment. We then define a match as the combination of no significant difference according to the LMM and a DMOS difference smaller than our Cohen's d-based threshold.

Results and discussion

The results from each experimental pipeline variation included in the experiments were aggregated and analyzed using LMM. Note that for the LMMs we merged the data from repeated presentations of the same pipeline variations: specifically the group A, E and I and the pair O and S. Statistical analyses on the pre-merged presentations revealed very few differences, a single significantly different attribute in each case. We judged this as representing random, rather than structural variation, and treated the merged data sets as if they originated from single experimental presentations.

Table 3 Results overview

						Marginal mean difference with baseline										
Presentation	Model	Renderer	TMO	Display	View	Pleasantness	Brightness	Diffuseness	Contrast	Uniformity	ShadowVis	Coziness	Liveliness	Tenseness	Detachment	Avg. Abs.
C	Orig	LT	Rein	Ph42	Static	**1.04**	0.20	−0.13	**1.18**	0.08	**1.55**	**0.57**	**0.50**	0.02	**0.78**	0.61
D	Orig	LT	iCAM	Ph42	Static	**1.14**	−0.19	0.10	**0.38**	**0.29**	**0.89**	**0.69**	**0.43**	−0.19	**0.73**	0.50
B	Orig	Ind	iCAM	Ph42	Static	−0.22	**0.39**	**0.41**	**0.85**	**0.57**	**−0.76**	−0.02	**−0.38**	**−0.36**	**−0.48**	0.44
G	Orig	Ind	Rein	HB	Static	**0.30**	**−1.18**	−0.03	**0.37**	0.12	0.09	0.07	**−0.36**	**−0.32**	**−0.71**	0.35
J	Orig	Ind	Rein	NEC46	Static	**0.57**	**−0.36**	0.07	−0.18	0.28	−0.11	**0.56**	0.22	**−0.43**	**−0.56**	0.33
L	Update	Ind	Rein	NEC46	Static	0.16	**−0.59**	−0.26	−0.29	0.15	**−0.41**	0.07	−0.13	−0.42	**−0.66**	0.31
H	Orig	Ind	Lin	HB	Static	0.21	0.11	**0.47**	−0.24	**0.85**	−0.08	0.06	0.11	**−0.27**	0.09	0.25
A,E,I	Orig	Ind	Rein	Ph42	Static	−0.16	**−0.32**	**0.20**	**0.31**	**0.38**	−0.12	−0.07	**−0.21**	−0.12	**−0.34**	0.22
Q	Update	Ind	Rein2	Proj	Static	0.23	**−0.52**	−0.10	0.13	0.21	−0.14	**0.39**	0.08	0.12	**−0.37**	0.23
R	Update	Ind	Rein2	Proj	Pano	0.16	**−0.76**	−0.23	**0.30**	0.22	0.02	0.21	−0.07	−0.13	−0.09	0.22
O,S	Update	Ind	Rein2	NEC46	Static	0.11	**−0.49**	−0.03	−0.14	**0.29**	**0.36**	0.25	−0.17	0.09	−0.23	0.22
N	Update	Ind	Rein2	Laptop	Pano	0.20	**−0.48**	0.15	−0.10	**0.36**	−0.01	0.19	−0.18	−0.28	−0.12	0.21
T	V-Ray	V-Ray	Rein2	NEC46	Static	0.07	**−0.43**	−0.09	−0.24	0.21	−0.29	**0.30**	−0.09	0.00	**−0.36**	0.21
M	Update	Ind	Rein2	Laptop	Static	0.01	**−0.39**	0.22	−0.07	**0.36**	−0.16	0.19	−0.13	0.10	−0.14	0.18
F	Orig	Ind	Rein	3D	Static	−0.14	−0.17	0.19	−0.16	**0.34**	−0.10	−0.04	−0.05	−0.11	**−0.35**	0.16
P	Update	Ind	Rein2	NEC46	Pano	−0.03	**−0.64**	0.00	0.05	0.22	0.02	0.09	**−0.37**	−0.04	−0.16	0.16
K	Photo	-	-	NEC46	Static	0.02	**−0.33**	−0.19	−0.07	0.19	0.21	0.03	−0.06	−0.04	−0.24	0.14

Columns in the table correspond to the ten perceptual attributes studied, while rows refer to different virtual presentations (abbreviations are noted in the text). Numbers in the cells are the marginal mean difference between the baseline experiment and the virtual presentation, colored to indicate significant differences according to the LMM (red), and non-significant differences (green). The final column is the average of all attributes' absolute mean difference per virtual presentation (average of absolute values in each row), and provides the sort order for the table (largest to smallest difference)

The results given in Table 3 outline marginal mean differences (on a 1–7 scale) between the baseline experiment and each virtual presentation for each of the ten perceptual attributes, after accounting for the modeled LMM factors. Shaded cells indicate situations in which the LMM analysis found a significant difference ($p < 0.05$) between real and virtual, unshaded indicates no significant difference. Table rows are sorted in order of decreasing average absolute mean difference, so worst-to-best by that measure.

It is immediately clear from the uneven distribution of shaded cells in the table that some variations (rows) perform much better than others, showing the impact of the pipeline variables under study. The worst variations are those rendered with LightTools (C, D), which were poorer in visual quality, as well as those using the iCAM TMO (B, D), which performs complete chromatic adaptation and thereby masks the color temperature changes in the scene. While the table is not in chronological order, the alphabetic order of our presentations over time exposes the trend that we have progressed from more to fewer significant differences over time. The table also shows that some attributes (columns) are apparently more difficult than others: for example, the Brightness column indicates significant differences in more variations than all other attributes, while the Diffuseness column shows differences in the least number of variations. Based on observed significant differences, high-level attributes such as Pleasantness and Coziness seem to be slightly better conveyed than physical attributes such as Brightness and Uniformity.

Arguably the best, the photograph presentation (K) has the smallest number of significant differences, only for Brightness, and also has the smallest average mean difference. The best-case virtual presentation was created with a pipeline of an accurate and updated scene model, Reinhard TMO, calibrated TV-size display and panoramic view (P). Hence, for these two presentations it is interesting to look more closely at the individual light conditions. Figure 5 therefore graphically depicts the difference in mean opinion scores (DMOS) with the real space for the presentations K and P per attribute

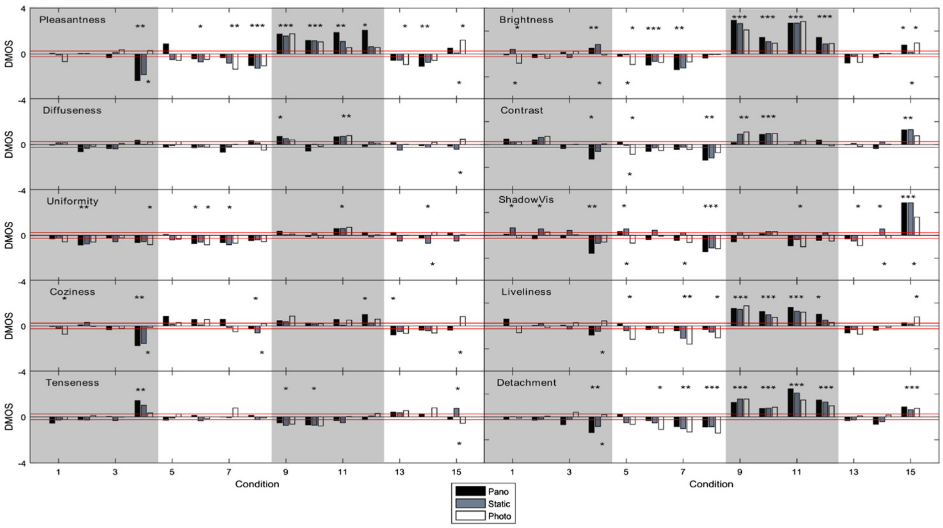

Fig. 5 Cohen's d comparison. Difference in mean opinion score (DMOS) for presentations P (Pano), O (Static), and K (Photo) as compared to the baseline. Red lines indicate the DMOS threshold corresponding to a small Cohen's d effect size

and condition. Additionally, in order to systematically verify the added value of the panoramic view, we also included the static version of this best-case visualization (O). We do not include some other good performers such as the 3D display (F) and the laptop-based static presentation (M), simply to keep the figures and discussion manageable. On the figure, the red lines indicate a mean difference of +/−0.25 on a 7-point scale, the threshold computed for what can be considered a small Cohen's d effect size. The asterisks above the bars indicate a significant difference between each presentation and the real space, and the asterisks below the bars indicate a significant difference between the photos and the static visualizations.

Looking first at which cases have the fewest LMM-based significant differences, photo appears best, with only 36 out of 150 comparisons different. Pano follows with 40, and static with 52. Since, as stated in section Statistical Analysis, lack of significance does not indicate that there is a match for the remaining conditions, we look at which cases have the most DMOS bars within the boundaries of small Cohen's d effect size. From this perspective, static is the best with 51 matches, followed by photo with 49 and pano with 34. Overall, these numbers showed that although two-thirds or more of the comparisons were not significantly different from the real space, roughly only one-third of the differences were small enough to be considered a match according to our Cohen's d small effect size threshold. Looking closely at Fig. 5, it is apparent that some lighting conditions match better than others. For many attributes the mean differences seem to be especially large for the dimmed light conditions (9–12) compared to the rest of the conditions.

Comparing these three presentations, photo can still be considered the best, with the fewest significant differences and almost as many matches as static visualization. Static conflictingly has both the most matches and the most significant differences, and pano simultaneously has the fewest matches and the fewest significant differences. Comparing static and pano, despite pano's good performance in Table 3, the Cohen's d conclusion is that static matches more conditions, especially for some attributes, notably Coziness. It is hard to conclude that there is value added by the look-around capability of the panoramic view.

Additionally, it is clear from Fig. 5 that the perceptual evaluations of the photographs behave similarly to both virtual variations. Looking specifically at the static case, while the LMM results show significant differences from the real space, they differ from the photographs in only 19 out of 150 comparisons. Hence, despite having differences with the real space, the static visualizations are nearly equivalent to the photographs.

In addition to the best panoramic (P) and the best static (O) virtual presentations in Table 3 there lie some other interesting variants worth looking at. The 3D presentation (F) performed well even without the model and TMO updates that helped in other cases, so it should be pursued further. The laptop (M,N in static and pano, respectively) presentations' performance shows that a small screen is not an obstacle to good perceptual presentation, a useful fact for portability, and the good performance of the V-Ray (T) presentation is a boon for rendering efficiency, with our admonition that getting good results with V-Ray requires some parameter fiddling. Going forward, it appears that visualization pipelines resulting in static or panoramic presentations using either Indigo (M,N,O,P) or V-Ray (T) renderers will provide robust results on TV-sized (O,P) or laptop-sized (M,N) displays. The additional expense (in terms of rendering time) of panoramic presentation may not be justified by the minimal gain in accuracy it seems to provide. 3D (F) is worth studying further with the caveat that it requires a specialized display.

Major effects of pipeline components

A basic question to address in this work is which pipeline elements have the strongest effect on the desired outcome of perceptual accuracy. This cannot be explicitly determined with the variable level choices and incomplete design, but indeed we have gained great insight into what is important. One way to look at it is that the weakest link limits the whole pipeline, somewhat analogously to how the worst component of an imaging system limits overall image quality. The physical simulation, including the models of the scene and light distribution, has an obvious effect on realism and affects not only scene accuracy (for example luminaires of the wrong shape or with incorrect light distribution), but also the intensity and contrast of the final image. We attempted to separate these effects – those of modeling versus those of presentation – by using photographs in one of our experimental presentations (straight photos, but nonetheless photos which included manual choices in overall intensity), and in so doing we verified that a photograph may indeed be considered nearly as good as ground truth.

If we focus on the presentation, the most critical factor is tonemapping, which is there to account for the dynamic range limitations of displays and affects everything from the intensity and contrast of the image to the sharpness of details. It also determines image intensity and color bias in terms of exposure and white balance, which are analogous to visual adaptation and thus are often intentionally partially corrected.

Looking at the perceptual attributes, brightness remains the most difficult to convey properly, despite its seemingly basic nature. This is presumably due in large part to the luminance adaptation capabilities of the human visual system, but it may also be cognitive. The brightness of some dim scenes is consistently overestimated. In discussion with experiment participants and experts, it seems that a low-intensity image may be interpreted as an underexposed image of a normally-lit room rather than a properly-exposed image of a dim room (see Fig. 6). However, a bedroom might be more likely to be interpreted as dimly-lit than an office, based on experience in the real world. In

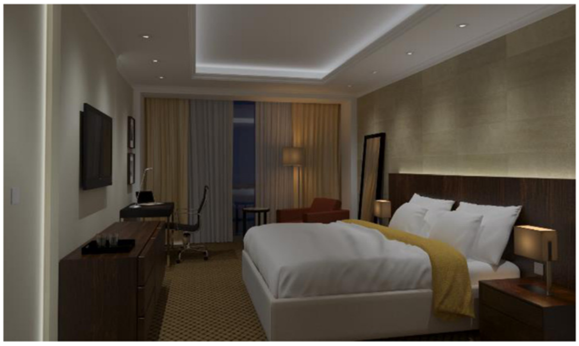

Fig. 6 Low-intensity image. Is this an underexposed image of a normally-lit room or a properly exposed image of a dim room?

ongoing experiments we are further probing the perception of brightness as a function of image intensity, scene context, and presentation viewing conditions.

Display also has an important but not dominating effect on perceptual accuracy, at least with the range of displays we have available today. Imagining an ideal display for our purposes, a first wish would be for high brightness and high dynamic range (specifically, HDR with no spatial limitation, compared to what is typical today), which would likely obviate the TMO. We found with a high brightness display that linear tonemapping appears to perform better than the Reinhard TMO for relatively low-contrast scenes, but the display's dynamic range limitations prevented us from using it for general application. An ideal display might have luminance levels sufficient to create retinal afterimages and intra-ocular glare just as a real luminaire naturally would. Looking at a bare LED of a million cd/m^2 leaves a visual impression that cannot be matched with today's displays, thus using simulations to judge visual comfort or glare remains out of reach. As an aside, we have experimented separately with modeled glare added synthetically to renderings and found that it does to some extent correlate with increased apparent brightness, but it never affects comfort or creates afterimages like increased physical intensity would. This remains an avenue for future research.

Secondary to HDR, an ideal display would also incorporate 3D stereoscopic full field of view for immersive interaction, hopefully less obtrusively than a head-mounted display, and a capable low latency real-time graphics engine behind it. Our 3D display presentation performed quite well even with a simpler scene model, and we would in the future explore the apparent advantage of stereoscopic presentation. Recent gaming engines are rapidly getting good enough to do accurate physical simulation of lighting systems, and mobile stereoscopic head-mounted displays are developing quickly. We have made first tests with such systems and believe they hold great potential. Once the human interface becomes transparent and intuitive, this kind of display system will flourish, and they will be very valuable for architectural lighting previews.

Critique of meta-analysis

As we mentioned in the Statistical Analysis section, our null hypothesis is that there is no difference between virtual presentations and the baseline real-world experiment. This could be considered worrisome because type-II error (false negative) is a favorable outcome for us, as it would seem to support our goal of a perceptual match between virtual and real. The real risk is that we might miss significant differences because we have too few observations or a sloppy experimental practice. It was with this in mind that we adopted the Cohen's d analysis.

Another way of looking at this problem might be a methods-comparison approach, as applied in daylight simulation research by Moscoso et al. [23]. They describe the application of the Bland-Altman method to find limits of agreement (LoA) [24]. At first this sounds promising, but in fact this method requires manually choosing a priori a mean difference threshold, essentially defining the size of a mean difference deemed important. In their work, Moscoso et al. chose a threshold of 1 unit on a 7-point scale, a threshold that would make all but 5 of our observed significant differences disappear, providing a very optimistic result. Picking arbitrarily we might have chosen a threshold of 0.5 units on the 7-point scale.

Any threshold is of course related to the size of the effects found, and we can conclude that even if we have missed something in the real vs. virtual comparison, it would be much smaller than the observed effect of lighting condition. Stepping back to the LMM results in Table 3, we observe that the largest non-significant mean differences in all the LMMs is 0.42 units out of the 7-point scale, indicating that any missed effect is likely smaller than 0.42 out of 7 points. This is quite small compared to the range of responses affected by the lighting conditions, as seen in Fig. 1, with observed ranges from 1 or 2 units for some perceptual characteristics up to about 6 for the most extreme (and note that for all perceptual attributes, lighting condition was found to be significant in our LMM analyses).

Future development

Future work could address a few points where our approach is not fully comprehensive. The presentation of brightness differences remains tricky and relies on manually chosen parameters in the Rein TMO. We have extensive experience in choosing these parameters, including a published study of how they may be influenced by the viewing environment [25], but we would certainly prefer a reliable objective way to set them. Our ongoing research continues to address brightness perception, for example in one experiment that directly compares perceived brightness matches between a physical luminaire and a displayed virtual luminaire, considering the limited display luminance, TMO parameters, and synthetic glare. In another recent experiment we considered an indoor environment additionally lit by both a wider range of artificial light and daylight, which of course increases scene luminance levels, where we found small but significant differences between real and virtual for Brightness as well as a few other attributes [13].

As advanced display technology and real-time graphics engines continue to develop, our 3D visualization pipeline may need to be tweaked to take full advantage of new capabilities. Positive improvements could be expected from virtual reality (VR), which could potentially improve adaptation effects through immersion and extreme FOV beyond the good performance we saw with a static 3D and panoramic presentations, and HDR displays, which will drastically increase the available display dynamic range and, as we expect based on our experience with a high-brightness (but not high dynamic range) display, simplify the TMO greatly.

Further research on additional perceptual attributes relative to lighting would also be welcome progress. Important aspects such as color rendering, spectral engineering, and glare have not been addressed, but they could be, especially with advanced displays, by following our methodology of proving perceptual accuracy.

Conclusions

Through this body of work, our goal was to build, test, and improve a perceptually accurate visualization pipeline for simulated images of lighting systems in context. Through a series of experiments, we have shown that we can create visualizations of lit environments that are practically as good as photographs in terms of accuracy for primary perceptual attributes, with very small differences in perceptual attributes as compared to a real-world environment. The most important factors influencing accuracy are the models behind the simulation, the tonemapping operator, and the display presentation. Display technology limitations mean some attributes such as glare and immersive field of view cannot be properly conveyed, but with our visualization pipeline physical lighting characteristics of a scene are generally conveyed well and higher-level attributes such as atmosphere are conveyed very accurately with some exceptions for dimmed lighting conditions. In general, Brightness remains a difficult attribute to convey accurately, due in part to the ambiguity of dimmed scenes and in part to the ability of the human visual system to adapt to different levels of scene intensity.

We have observed that any weak point in the visualization pipeline limits its perceptual accuracy, as in system image quality. We have iteratively improved our approach to ensure that visualizations are presented in as perceptually accurate a way as possible. Our success with this goal and insight into the factors affecting it continue to improve with ongoing experiments. We look forward to improvements in virtual reality and display technology that will improve luminance levels, dynamic range, and field of view. All of these will likely enhance adaptation and presence, as well as potentially perceptual accuracy.

Abbreviations

3D: three-dimensional; 3D: Philips 3D LCD; ANOVA: analysis of variance; BRDF: bi-directional reflectance distribution function; DMOS: difference in mean opinion scores; FOV: field of view; HB: high-brightness LCD; HDR: high dynamic range; iCAM: Kuang's iCAM07 TMO; ICBE: Internal Committee for Biomedical Experiments; Ind: indigo renderer; Laptop: HP laptop display; LCD: liquid-crystal display; Lin: linear TMO; LMM: linear mixed model; LT: lighttools simulation software; MOS: mean opinion score; NEC46: NEC 46-inch LCD; Pano: panoramic presentation; Ph42: Philips 42-inch LCD; Proj: projector; Rein: Reinhard's 2002 photographic TMO; SSL: solid-state lighting; Static: static presentation; STD: standard deviation; TMO: tonemapping operator.

Competing interests

The authors declare that they have no competing interest.

Authors' contributions

MM led the project that included this research, [co-] designed all experimental variations, and drafted much of this paper. MS co-designed and executed several experiments, performed the statistical analyses, and drafted the results section. ML co-designed and executed one experiment and assisted with analyses and paper draft. All authors edited and approved the final manuscript.

Acknowledgements

The authors would like to thank Ulrich Engelke, Bart Salters, Chris Damkat, Dragan Sekulovski, Pieter Seuntiens, and Shih-Han Chen for their contributions over the course of this work.
Submitted to: Journal of Solid State Lighting *Thematic Series on SID Display Week SID/IES Lighting Track Lu & Larimer, Eds.*

Received: 9 September 2015 Accepted: 14 December 2015
Published online: 29 December 2015

References

1. Appel A (1968) Some techniques for shading machine renderings of solids. In Proceedings of the April 30–May 2, 1968, spring joint computer conference (AFIPS'68 (Spring)). ACM, New York, NY, USA, 37–45. doi:10.1145/1468075.1468082
2. Whitted T (1980) An improved illumination model for shaded display. Commun. ACM 23, 6 (June 1980), 343–349. doi:10.1145/358876.358882
3. Kajiya JT (1986) The rendering equation. SIGGRAPH Comput. Graph. 20, 4 (August 1986), 143–150. doi:10.1145/15886.15902
4. Ward GJ (1994) The RADIANCE lighting simulation and rendering system. In Proceedings of the 21st annual conference on Computer graphics and interactive techniques (SIGGRAPH'94). ACM, New York, NY, USA, 459–472. doi:10.1145/192161.192286
5. Drago F, Myszkowski K (2001) Validation proposal for global illumination and rendering techniques. Computers & Graphics 25(3):511–518
6. Villa C, Parent E, and Labayrade R (2010) Calibrating a display device for subjective visual comfort tests: selection of light simulation programs and post-production operations, in Proc. CIE Light Efficiency, (Vienna)
7. Villa C, and Labayrade R (2010) Psychovisual assessment of tone-mapping operators for global appearance and colour reproduction, in Proc. CGIV, 189–196 (Joensuu)
8. Schielke T. Validity of simulations for lighting and brand image evaluation. Lighting Research and Technology 1477153515589116, 2015. doi:10.1177/1477153515589116
9. Newsham G, Richardson C, Blanchet C, Veitch J (2005) Lighting quality research using rendered images of offices. Lighting Research & Technology 37(2):93–112. doi:10.1191/1365782805li132oa

10. Salters B, Murdoch MJ, Sekulovski D, Chen S, Seuntiens P (2012) An evaluation of different setups for simulating lighting characteristics, in Proceedings of SPIE 8291, 1–13 (San Francisco)
11. Engelke U, Stokkermans MGM, Murdoch MJ (2013) Visualizing lighting with images: converging between the predictive value of renderings and photographs, in proceedings of SPIE 8651, 1–10 (San Francisco)
12. Murdoch MJ, Stokkermans MGM (2014) Effects of image size and interactivity in lighting visualization, in Proceedings of SPIE 9014, (San Francisco)
13. Stokkermans MGM, Chen Y, Murdoch MJ, Vogels IMLC, Heynderickx IEJ (2015) "Effect of daylight on atmosphere perception: comparison of a real space and visualizations". in Human Vision and Electronic Imaging XIX. Proc. SPIE 9394. SPIE, San Francisco, CA
14. Murdoch MJ, Stokkermans MGM, Lambooij M (2015) "50.3 invited paper: perceptual accuracy in the visualization of lighting scenes". in SID symposium digest of technical papers. San Jose, CA, SID
15. Vogels I (2008) "Atmosphere metrics," in probing experience. Springer, Netherlands, pp 25–41
16. Vogels IMLC, de Vries M, van Erp TAM (2008) Effect of coloured light on atmosphere perception, in proceedings AIC, (Stockholm)
17. Seuntiens PJH, Vogels IMLC (2008) Atmosphere creation: the relation between atmosphere and light characteristics, in Proceedings 6th Conference on Design & Emotion, (Hong Kong)
18. Reinhard E, Stark M, Shirley P, Ferwerda J (2002) Photographic tone reproduction for digital images. ACM Transactions on Graphics 21:267–276
19. Reinhard E. Parameter estimation for photographic tone reproduction. J. Graph. Tools 7, 1 (November 2002), 45–52. doi:10.1080/10867651.2002.10487554
20. Kuang J, Johnson GM, and Fairchild MD (2007) iCAM06: A refined image appearance model for HDR image rendering. J. Vis. Commun. Image R.18, 406–414
21. McCulloch CE, Searle SR (2001) Generalized, linear, and mixed models. Wiley, New York
22. Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum, New Jersey
23. Moscoso C, Matusiak B, Svensson UP, and Orleanski K (2015) Analysis of stereoscopic images as a new method for daylighting studies. ACM Trans. Appl. Percept. 11, 4, Article 21 (January 2015), 13 pages. doi:10.1145/2665078
24. Altman DG, Bland JM (1983) Measurement in medicine: the analysis of method comparison studies. The Statistician 32:307–317
25. Stokkermans M, Murdoch MJ, and Engelke U (2012) "Preference for key parameter of tone mapping operator in different viewing conditions." in Experiencing Light. Eindhoven