

## **Adequacy of immersive Virtual Reality for the perception of daylight spaces: comparison of real and virtual environments**

K. Chamilothoni\*, J. Wienold and M. Andersen

*Laboratory of Integrated Performance in Design (LIPID), Ecole polytechnique fédérale de Lausanne (EPFL), Switzerland*

**Abstract** - This paper presents a novel experimental method which uses a Virtual Reality (VR) headset, aiming to provide an alternative environment for the conduction of subjective assessments of daylight spaces. This method can overcome the difficulty of controlling the variation of luminous conditions, one of the main challenges in experimental studies using daylight, while its novelty lies in the implementation of physically-based renderings into an immersive virtual environment. The present work investigates the adequacy of the proposed method to evaluate five aspects of subjective perception of daylight spaces: the perceived pleasantness, interest, excitement, complexity and satisfaction with the amount of view in the space. To this end, experiments with 29 participants were conducted, to comparing the user's perception of a real daylight environment and its equivalent representation in VR and testing the effect of the display method on the participants' perceptual evaluations, reported physical symptoms, and their perceived presence in the virtual space. The results indicate a high level of perceptual accuracy, showing no significant differences between the real and virtual environments on the studied evaluations. In addition, there was a high level of perceived presence in the virtual environment and no significant effects on the participants' physical symptoms after the use of the VR headset. Following these findings, the presented experimental method in VR seems very promising for use as a surrogate to real environments in investigating the aforementioned five dimensions of perception in daylight spaces.

**Keywords:** virtual reality, subjective user assessments, human perception, visual perception, lighting.

---

\* Corresponding author: Kynthia Chamilothoni, e-mail: [kynthia.chamilothoni@epfl.ch](mailto:kynthia.chamilothoni@epfl.ch)

# 1. INTRODUCTION

Daylight has an undeniable value in various aspects of architectural design, ranging from energy efficiency and user comfort to the perception of architectural space. Various studies have addressed the importance of perceptual qualities of daylight, such as the variation and composition of light for our experience of space [Parpaire and others 2002; Rockcastle and Andersen 2014; Rockcastle and others 2016], while emphasizing the need for a human-centric approach focusing on the occupant's point of view [Amundadottir and others 2017]. Although daylighting research has produced established metrics regarding human comfort and the energy performance of daylight, we are left with an inadequate understanding of the effect of luminous conditions on the user perception and experience.

A significant barrier in the acceleration of knowledge in this field is the difficulty of controlling the variation of luminous conditions in experimental studies. Although daylight is identified as one of the driving factors in architectural design [Zumthor 2006; Holl and others 2011], there are currently no methods that allow us to visualize and evaluate the dynamics and complexity of daylight in space, truly reproducing a user's experience, except when this space is finished and built. However, using real environments to investigate human perception is complex in parameters and resources, while daylighting research faces the particular problem of conditions that change over time, such as weather and sky [Bülow-Hübe 1995; Newsham and others 2010].

As a result, there has been a growing trend towards the use of virtual representations in research. Various studies have suggested that both photographs [Hendrick and others 1977; Newsham and others 2010; Cauwerts 2013] and renderings [Mahdavi and Eissa 2002; Newsham and others 2005; Cauwerts 2013] are a promising medium for investigating subjective impressions of space and light. Current rendering simulation tools can produce physically-based renderings, which provide accurate photometric data and allow the relation between these photometric measurements and subjective assessments, a necessary process to uncover the existence of links between stimulus and response.

The challenge lies of course in obtaining research findings that are valid, reproducible and generalizable from virtual to real environments. A key factor in this is the creation of virtual environments that are perceptually realistic and provide an experience that is "indistinguishable from normal reality" [Loomis and others 1999]. The user interaction and immersion have been identified as crucial parameters in creating virtual environments that can adequately substitute the human experience in the real space [Bishop and Rohrmann 2003; de Kort and others 2003; Newsham and others 2010; Cauwerts 2013]. Although there are few studies investigating the experience of an rendered immersive and interactive virtual environment [Franz and others 2005; Heydarian, Carneiro, and others 2015], none compares the evaluations of the virtual space with those of a corresponding real space, rendering the applicability of the experimental procedure and results questionable.

This paper introduces an experimental method using a novel projection technology, a Virtual Reality (VR) Headset, for the conduction of experiments investigating the perceptual effects of daylight. The proposed method produces an immersive virtual

environment from physically-based renderings, combining a wide field of view, interactive viewing mode, and stereoscopy for the main view direction. These advantages are specifically relevant for the identification of robustness and consistency in experimental results, as the mobility offered by VR Headsets allows the reproducibility and seamless conduction of experiments with identical stimuli. This technology has the potential to overcome important barriers in conducting experiments in real environments by allowing the control of luminous conditions and rapid alternation of visual stimuli, while offering a high degree of immersion, a factor which has been identified as crucial in reproducing the experience of a real space.

## **2. STATE OF THE ART**

In this section, we review relevant issues for the identification of a perceptually accurate immersive virtual environment: the physical discomfort from the use of the virtual reality device and the perceived presence in the virtual scene.

### **2.1. Factors impacting the perceptual accuracy of virtual environments**

Although simulated two-dimensional virtual scenes have been repeatedly identified as an adequate medium for investigating subjective impressions of space and light [Mahdavi and Eissa 2002; Newsham and others 2005; Cauwerts 2013], the immersion of the user in the virtual environment has been a recurring subject in various studies comparing virtual and real environments. This attribute regards not only the field of view provided by the device, an important factor when comparing real and virtual environments [Newsham and others 2010], but also the user interaction with the presented scene. It is suggested such an interaction would greatly improve the realism and thus the potential for user experience studies [Bishop and Rohrmann 2003; De Kort and others 2003]. Studies investigating the influence of different components in the visualization workflow, such as display type, field of view and user interaction, have identified an interactive panoramic view mode as the most perceptually accurate [Cauwerts 2013; Murdoch and others 2015], highlighting the importance of immersion and interactivity in the virtual scene.

Stereoscopy is another issue commonly raised when discussing the perceptual accuracy of virtual environments. Although binocular disparity in stereoscopic projections is primarily recognized due to its contribution in providing depth information [M. Loomis and others 1996], recent studies have highlighted its importance for multiple evaluation concepts in addition to depth perception [Lambooj and others 2011]. In the subject of subjective evaluations of daylit scenes, while stereoscopic projections have been found to be an adequate representation method of real environments [Charton 2002; Moscoso, Matusiak, Svensson, and others 2015], studies that compare the subjective assessment of scenes in monoscopic (2D) and stereoscopic (3D) projections suggest that there is little difference in how they are perceived. Although the 3D projection was deemed as more realistic than the equivalent 2D projection [Cauwerts and Bodart 2011], stereoscopy did not have a significant effect when directly comparing the subjective evaluations of

appraisal, emotion and perceived presence in the two projection modes [Baños and others 2008; Cauwerts and Bodart 2011].

To the authors' knowledge, no previous testing has been conducted on the use of a Virtual Reality headset regarding the perceptual effects of daylight. A recent study using this headset for the investigation of task performance in office environments (reading speed, comprehension, object identification) and comparing the performance indicators to benchmarks measured in a similar physical environment, concluded that there were no significant differences in performance between the virtual and real space [Heydarian, Pantazis, and others 2015]. A VR headset is used in studies investigating the effects of daylight composition on the subjective perception of the virtual space, showing that the technology allows the differentiation between daylight pattern distributions [Chamilothori and others 2016; Rockcastle and others 2017]. Another relevant study allows the users of a VR Headset to adjust the blinds and artificial light in a virtual office space, aiming to investigate the participant's lighting preferences for office related tasks [Heydarian, Pantazis, and others 2015]. The participants were invited to adjust the light condition in the room to their preference and their preferred settings were later compared with light maps that represented the chosen lighting setup of the user. Although this study is an important step in exploring the potential of immersive virtual environments, it presents two significant limitations: the rendering of the virtual space was not physically-based and the researchers had no control over the specific properties of the virtual space projection. The final projected scenes were rendered through a game engine with a photo-realistic real time renderer. As a result, there is no indication of the actual light measurement values of the scenes that the participants were shown and acted upon. The participant's preferences and behavior in the virtual environment were identified in light maps produced separately with physically-based calculations, corresponding to high dynamic range values of real world luminous conditions and not to those of the scene that was projected with the limited luminance range of the headset.

## **2.2. Perceived presence in the virtual space**

Along with the simulation accuracy of the virtual scene, the subject of presence is an emergent factor in creating a virtual environment that can adequately replicate our experience of a real space [De Kort and others 2003; Diemer and others 2015]. Presence is defined as the sense of 'being there' in the virtual environment [Slater and Wilbur 1997]. Schubert and others [2001] identified three dimensions of self-reported presence through factor analysis: spatial presence, involvement and realism. These factors correspond to the user's sense of being in the virtual space, their lack of awareness of the real world, and the perceived realism of the virtual scene in comparison with the real environment.

The subject of adverse physiological reactions has been identified as a possible factor of the perceived presence in virtual environments [Lessiter and others 2001; Schubert and others 2001]. A study by Van der Spek and Houtkamp [2008], inducing simulator sickness in a static virtual environment, found that the evaluation of the environmental dimensions of arousal and pleasantness was affected by the participants' physical discomfort. This finding is particularly relevant for the use of virtual reality as an empirical research tool,

emphasizing the need for a multi-criteria analysis for the adequacy of virtual reality as a substitute for real environments.

### **2.3. Problem statement**

In the pursuit of creating simulated environments that can be used as a substitute for real ones in empirical research, studies have highlighted several factors of importance. When investigating the effects of lighting on perceptual impressions, such as the perceived *pleasantness* or *interest* of an evaluated scene, the virtual environment needs to correspond to accurate photometric data. In addition, the user interactivity and immersion in the simulated space have been identified as essential in adequately reproducing the human experience in the real space. Although considerable research has been devoted to comparing and validating methods that aim to couple these features, so far existing methods are lacking in the depth of user immersion within the virtual environment.

This paper introduces a novel method for the creation of fully immersive scenes from photometrically accurate renderings, aiming to encourage the use of VR technologies in lighting research as an empirical tool. Following the approach of Bishop and Rohrmann [2003], who urged researchers to conduct validity assessments when introducing a new medium of simulation or projection, we present the results of an experimental study investigating the adequacy of the proposed method in three different dimensions: the perceptual accuracy of the method, the physiological effects of using the device, and the reported presence of the subjects in the virtual environment. To this end, we will investigate the difference between subjective evaluations in a real daylit environment and in its virtual representation, compare the reported physical symptoms of the participants before and after the use of the VR headset and assess the participants' perceived presence in the virtual scene through verbal questionnaires.

## **3. METHOD**

The method followed in this study consists of three main steps: the selection of visual stimuli in the real and virtual environments, the experimental design, and the approach regarding the statistical analysis of the experimental results, as illustrated in Fig. 1.

## Methodology

1. Selection of scenes in the real and virtual environment
  - a. Real Environment selection of experimental space
  - b. Virtual Environment generation of physically-based renderings  
tonemapping and projection in the VR Headset
2. Experimental design to investigate the adequacy of the method
  - a. Perceptual Accuracy evaluation of participant perception in the real and virtual environment (within-subject)
  - b. Physical Symptoms reported physical symptoms before and after using the VR headset (within-subject)
  - c. Reported Presence evaluation of perceived presence in the virtual environment after using the VR headset
3. Statistical analysis of experimental results
  - a. Perceptual Accuracy comparison of participant evaluation between the real and virtual environment
  - b. Physical Symptoms comparison of participant reported symptoms before and after using the VR headset
  - c. Reported Presence analysis of perceived presence in the virtual environment after using the VR headset
  - d. Effect of Presentation Order analysis of environment presentation order effect

**Fig. 1. Illustration of the methodological approach in the study.**

We will present the method in five subsections, corresponding to the software and equipment used (subsection 3.1), the workflow regarding the experimental stimuli in the real (subsection 3.2) and virtual (subsection 3.3) environments, the experimental design (subsection 3.4), and lastly the methodology behind the statistical analysis of the results (subsection 3.5).

### 3.1. Equipment

The VR headset used in this study is the Oculus Rift Development Kit 2 (DK2), which uses a 1920 x 1080 pixel low persistence OLED display with a refresh rate up to 75Hz, resulting in a resolution of 960 x 1080 pixels per eye. The display offers a 100° horizontal and 110° vertical field of view. Although the maximum luminance of the display is up to 300 cd/m<sup>2</sup>, in this experiment the maximum luminance measured on the display was 80 cd/m<sup>2</sup> and 40 cd/m<sup>2</sup> on the lens due to software limitations. In the development and execution of this study, the software used was Oculus Runtime 0.7.0.0, in combination with Unity Game Engine 4.9.6 and the corresponding Unity Package *OculusUtilities* (from <https://unity3d.com/> and <https://www.oculus.com/> respectively).

## 3.2. Visual stimuli: real environment

In order to test the perceptual accuracy of the proposed method, the authors set to compare the subjective evaluations of a real space and its representation in virtual reality. Aiming to keep in line with the characteristics of experimental spaces in relevant studies, the criteria that were established for the selection of the test room were a small-to-medium size [Moscoso, Matusiak, Svensson, and others 2015], implied office use [Heydarian, Pantazis, and others 2015; Murdoch and others 2015], and daylight access from the south façade to allow for experimental conditions with direct sunlight. The DEMONA (Module de démonstration en éclairage naturel) daylighting test module on the EPFL campus, shown in Fig. 2 (a), was selected as it fulfilled these criteria and was accessible by the participants. The test room was resembling an office with a desk and two chairs, basically achromatic surfaces –gray carpet and white walls and ceiling- and windows on the north and south façade. The north façade was covered, allowing daylight to enter only from the south (Fig. 2, b).

## 3.3. Visual stimuli: virtual environment

### 3.3.1 Generation of physically-based renderings

Using an existing 3D model (courtesy of LESO Laboratory, EPFL) as a starting point, the test room was modelled in Rhinoceros (<https://www.rhino3d.com/>) and then exported through the DIVA-for-Rhino toolbar (<http://diva4rhino.com/>) to Radiance (<http://www.radiance-online.org/>), an extensively validated physically-based lighting simulation tool [Ward Larson 1994]. The scene preparation and simulation protocol followed well-established workflows for producing high accuracy visualizations with Radiance, specifying the material properties from spectrophotometer measurements [Larson and Shakespeare 1998], generating sky descriptions based on radiation measures with *gendaylit* [Larson and Shakespeare 1998; Cauwerts 2013], described in Appendix 1.1, and using high accuracy rendering parameters [Reinhart 2005], shown in Table 1. The color and specularity of the surfaces and furniture in the experimental room were measured with a Konica Minolta CM-600d Spectrophotometer and translated to Radiance material properties, shown in Table 2. A series of perspective view high dynamic range (HDR) renderings were generated using *rpict* in Radiance the viewpoint shown in Fig. 2 (c), dividing the 360 degree field of view in 6 sections with 90 degree horizontal and vertical field of view.

**TABLE 1. Radiance *rpict* parameters for the perspective view renderings (-vtv)**

-vs	-vl	-ab	-s	-st	-lv	-ad	-as	-aa	-ar	-ps	-pj
0	0	5	1	0	0.00001	20000	10000	0.05	512	0	0

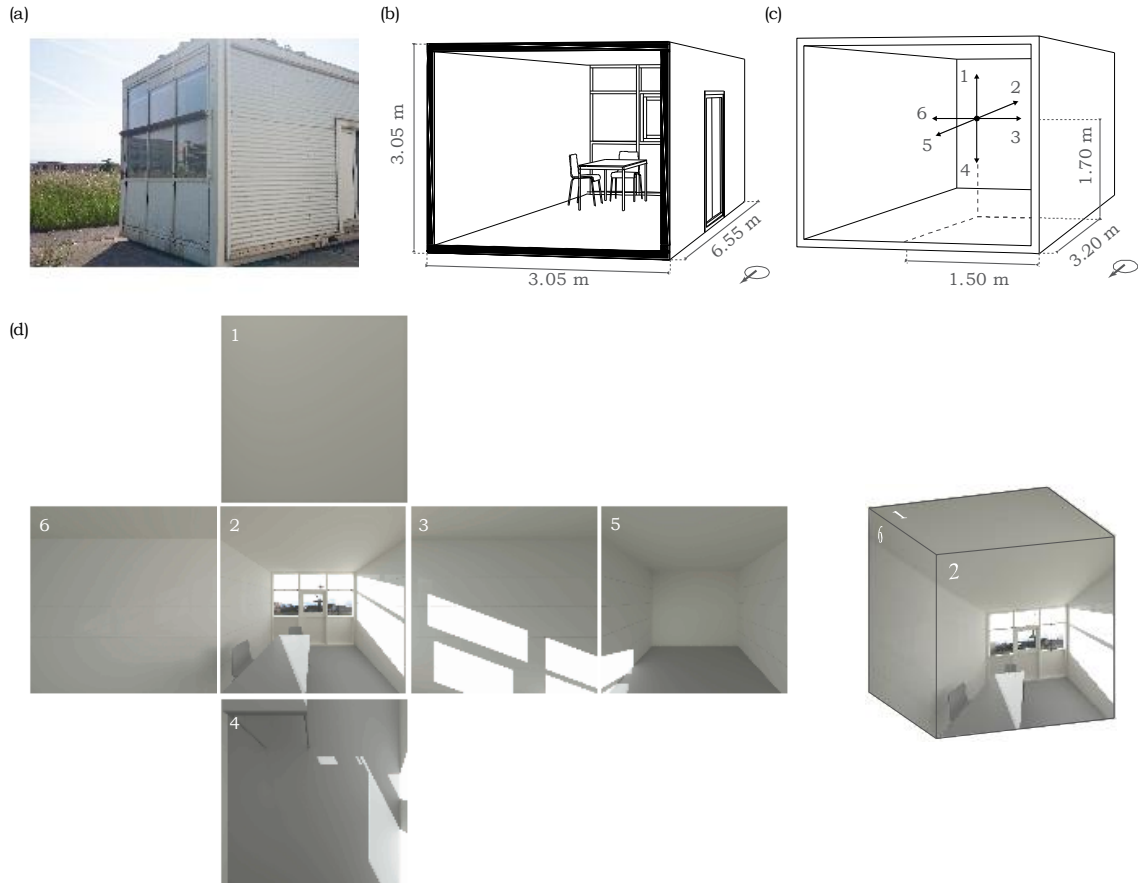
Resolution: 3600 x 3600 (scaled down to 1200x1200 using *pfilt*)

Rendering time: 16 hours for all six view directions

**TABLE 2. Radiance material properties for the main surfaces**

Surface	Type	R	G	B	Reflectance	Specularity	Visual Transmittance
Ceiling	plastic	0.93	0.92	0.86	83%	0	
Floor	plastic	0.46	0.47	0.48	19%	0	
Walls	plastic	0.92	0.92	0.89	82%	0.01	
Table	plastic	0.75	0.75	0.73	54%	0.01	
Window*	glass	0.96	0.96	0.96			88%

\*A non-coated double paned glazing was used for the window.



**Fig. 2. Photograph of the test room (a), illustration of the interior in the 3D model (b), placement of the viewpoint and six view directions (c) and example of resulting renderings of the 360° virtual space (d) divided in six sections of 90° vertical and 90° horizontal field of view.**

By keeping the viewpoint fixed and varying the view direction, the produced set of renderings forms an expanded cube, illustrated in Fig. 2 (d). The exposure of the HDR renderings was adjusted intuitively to match the appearance of the real space by using *pfilt* to apply a uniform exposure multiplier. The images were then converted to low dynamic range BMP files using *ra\_bmp* with a gamma correction factor of 2.2 and ensuring the application of identical settings for all six view directions.

To create the perception of depth with two dimensional images, we can generate a stereoscopic scene by projecting a different picture to each eye (Fig. 3, right) simultaneously. If the viewpoints of these images have a disparity in the horizontal axis in



a measure equal to the interpupillary distance ( $d$ ) of the subject's eyes (Fig. 3, left), the resulting image is perceived as three-dimensional.



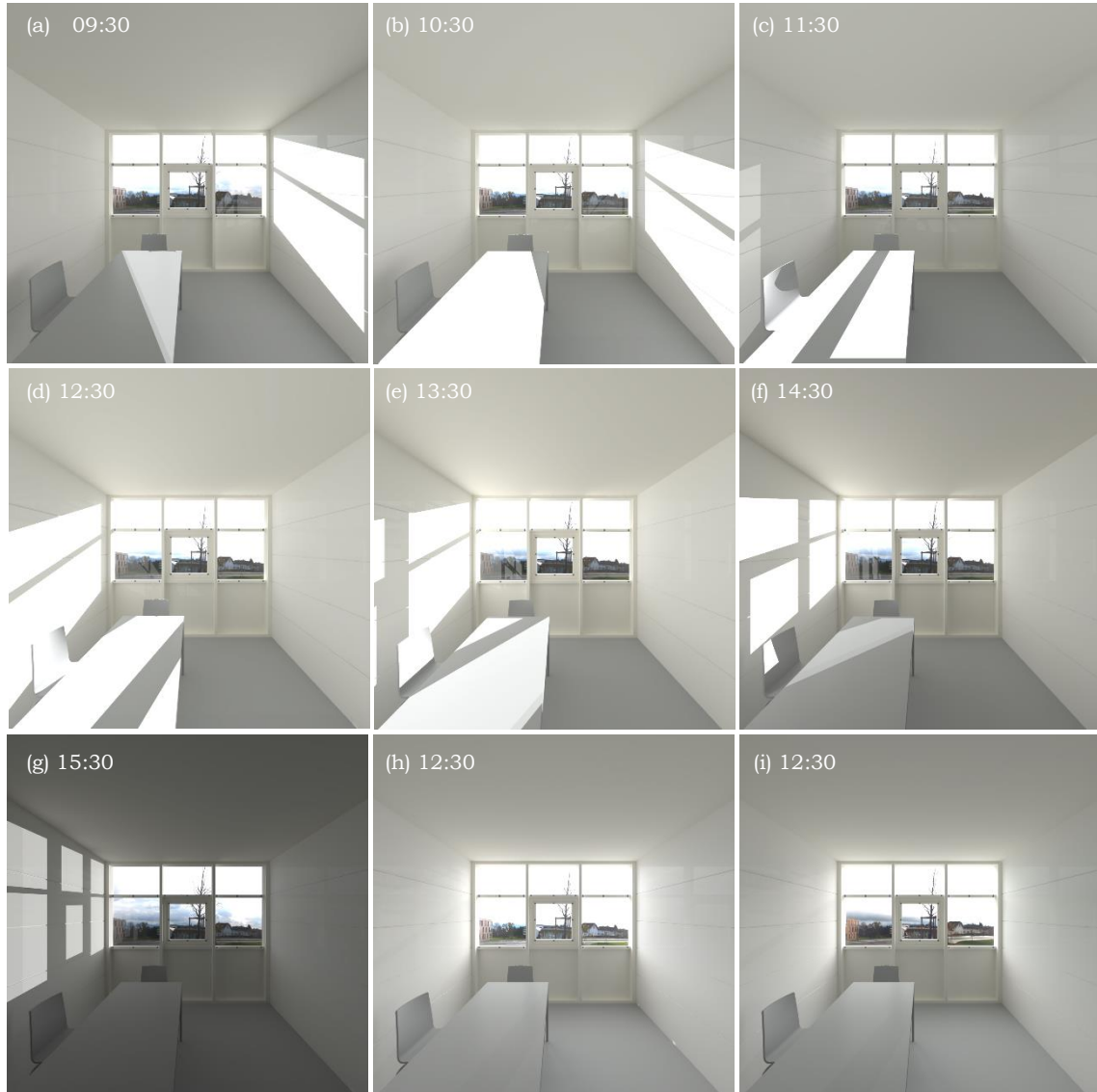
**Fig. 3. The principle of stereopsis used for the main view direction in the Virtual Reality Headset (left) and an example of projected stereoscopic image (right).**

Following this principle, the simulation procedure described above was repeated for two viewpoints in the digital model with a horizontal distance of 65 mm between them, suggested as the average interpupillary distance in the Oculus documentation [2015]. As the participants of the experiment would be mostly looking in front of them, shown as view direction 2 in Fig. 2 (c), the selected viewpoints correspond to the eyes of a participant looking in this view direction. Although the stereoscopic effect is correct for the view directions of front and back, it is not the case for the other view directions, where the two viewpoints will not correspond anymore to the eyes of a subject turning their head in the virtual space. This was a result of software limitations in the time of the study, although future work in Radiance aims to implement omnidirectional stereoscopic projection types [McNeil 2016]. When experiencing the virtual scene, the discrepancy was imperceptible as the non-stereoscopic view directions did not contain objects, leading to a minimal effect on the user's perception of depth.

In consideration of how prevalent is an order bias in relevant studies, where the subjects are first exposed to the real environment and these conditions are used to generate the virtual environment [Cauwerts and Bodart 2011; Murdoch and others 2015], we decided to compromise in virtual scene accuracy in favor of eliminating this bias and create the simulated scenes before the experiment. The experiment took place in November during the course of eight days, which were represented with a selection of virtual scenes corresponding to different times of day, sky type and view out. Due to time constraints, the variations of the scene were realized for one date only in November, although minor changes in sun angle were perceived in this period.

Aiming to address the daylight variation, seven scenes were rendered for clear sky type corresponding to every hour from 9:30 to 15:30, shown in Fig. 4 (a) to (g), and two scenes for overcast sky type, using 12:30 as the time of day, shown in Fig. 4 (h) and (i). The two scenes of overcast sky differed only in their view out of the window, as two different panoramas were mapped on the Radiance sky to account for the prevalent views in overcast conditions the weeks before the experiment.

For each scene, we rendered two sets of six images: each image corresponded to a perspective view of  $90^\circ$  horizontal and  $90^\circ$  vertical field of view from the same viewpoint, shown in Fig. 2, while each set of images corresponded to the viewpoint of one eye of a person looking towards the main frontal view direction of the scene, as shown in Fig. 3.

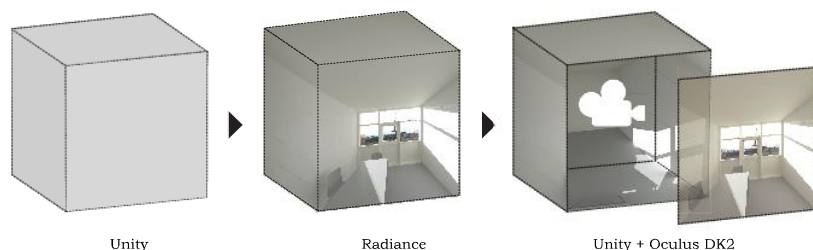


**Fig. 4. The front view direction for the right eye used in each of the nine scenes. The scenes from (a) to (g) were rendered in hourly time steps from 09:30 to 15:30 for clear sky type, and the scenes (h) and (i) were rendered for 12:30 and overcast sky type using two different panoramas mapped on the Radiance sky.**

### **3.3.2. Generation of 360 degree immersive scenes**

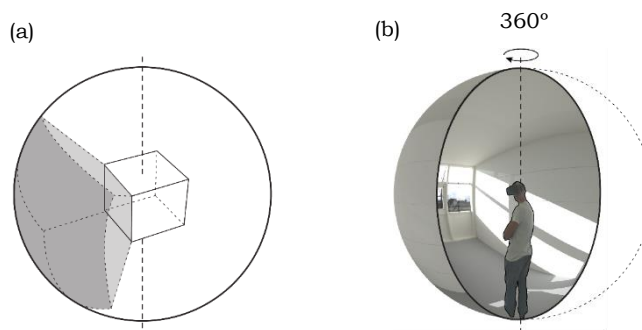
The renderings developed with Radiance were then imported in the game engine Unity. Drawing from the field of game design where a textured cubemap has been widely used to produce an immersive environment mapping [Greene 1986], we recreated this projection type in Unity. In a cubemap, or cube projection, the environment is projected onto a cube as seen from a particular viewpoint (illustrated in Fig. 2, (c) and (d)). A significant advantage of this projection mapping, as noted by Greene, is that it can be produced with any

rendering program that can generate perspective view visualizations. Figure 5 illustrates the process of creating a cubemap projection in Unity, applying the renderings from Radiance to the corresponding faces of a cube as textures (Appendix 1.2). Two cubes were created to simulate the principle of stereoscopy, each using as textures the renderings generated from the viewpoint of the equivalent eye, visible only to the corresponding half of the screen of the VR headset.



**Fig. 5. Illustration of workflow for the generation of immersive 360° scenes.**

A virtual camera was placed in the center of each cube (Fig. 5, right) using the *OVRCameraRig* script in Unity, which allowed the control of the camera through the head-tracking feature of the Oculus Rift headset. By placing the camera in the middle of this textured cube, the user's viewpoint corresponds to the one used to generate each set of renderings. Through this process, we can produce scenes for Oculus Rift where the user is able to look around and explore the space from a selected viewpoint. The scene is perceived as fully immersive and three-dimensional, as illustrated in Fig. 6.



**Fig. 6. Diagram of the word-to-cubemap projection for one cube face (a) and illustration of the freedom of movement in the projected virtual reality scene (b).**

### 3.4. Experimental design

In this study, we chose a within-subjects design where subjects were shown both the real and the virtual environment, counterbalancing the stimuli presentation order. The use of repeated measures has the advantage of eliminating the effect of variance between individual participants, as each subject acts as their own control. By randomizing the presentation order of the two environments within participants, the first environment they are exposed to functions as their individual reference. The experimental data was collected through questionnaire items, shown in Table 3 and grouped in three sections relating to the investigation of the perceptual accuracy of the method (Perceptual Impressions), the effect of using the VR headset on the user's physiological reactions (Physical Symptoms),

and the user's perceived presence in the virtual space (Reported Presence).

**TABLE 3. Overview: variables and questionnaire items**

---

**Independent Variables**

- S.1 Environment: real or virtual environment
- S.2 Presentation order: real space first or virtual environment first

**Dependent Variables**

*Perceptual Impressions*

- PI.1 How pleasant is this space?\*
- PI.2 How interesting is this space?\*
- PI.3 How complex is this space?\*
- PI.4 How exciting is this space?\*
- PI.5 How satisfied are you with the amount of view in this space?\*

*Physical Symptoms*

- PS.1 How sore do your eyes feel?\*
- PS.2 How fresh does your head feel?\*
- PS.3 How clear is your vision?\*
- PS.4 How fatigued do you feel?\*
- PS.5 Do you have any other symptoms? (*open question*)

*Reported Presence*

- RP.1 How much did your experience in the virtual space seem consistent with your experience in the real space?\*
  - RP.2 How much did you feel like 'being there' in the virtual space?\*
  - RP.3 How much did the virtual space become the reality for you?\*
- 

\* A scale from 1 to 5, 1 corresponding to 'Not at all' and 5 to 'Very much', was used for the marked questions.

The studied *perceptual impressions* relate to the component of *emotional response* rather than the *light appearance* in the perception of the luminous conditions [Van Erp 2008], placing the focus of the study in investigating further than the evident features of the luminous environment and exploring the impact of lighting condition on high order perceptions, such as complexity and interest [Boyce 2003]. Drawing from the pioneering work of Flynn [1973], Vogels identified two factors that can represent the *atmosphere* of a space, which is suggested as a more objective variable to measure the perception of luminous conditions: *cosiness* and *liveliness*, which correspond to the dimensions of *pleasantness* (affect) and *interest* (arousal) [Russell 1980] found in emotion theory [Vogels 2008]. The questionnaire items for *Perceptual Impressions* were based on the work of Vogels [2008] and Rockcastle and others [2016], adapted to unipolar scales and focusing on the dimensions of pleasantness, visual interest and complexity, with the addition of a question regarding the amount of view in the space, which has been shown to affect the perceived pleasantness of the space [Moscato, Matusiak, and Svensson 2015].

The questions for *Physical Symptoms* were based on a study investigating visual comfort with stereoscopic displays [Shibata and others 2011]. Lastly, the three items for

*Reported Presence* correspond to the dimensions of presence as identified by Schubert and others [2001]: realness, spatial presence and involvement (Table 3, *RP.1*, *RP.2*, and *RP.3*).

A 5-point scale with verbal anchors at the end points ( 1 corresponding to "Not at all" and 5 corresponding to "Very much") was used for all questions except for PS.5, "*Do you have any other symptoms?*", which had an open response. Both the order of the questions within each section and the polarity of the scale for each question were randomized. The 5-point range of the scale was selected as one of the most highly rated in terms of its ease of use and reliability [Preston and Colman 2000], and was chosen over the most commonly used 7-point range after a pilot study from the authors that showed that participants had difficulty answering a verbal questionnaire with 7-point scales as they had no visual reference of the scale range in virtual reality. The responses on the 5-point scale were assigned as ordinal in MATLAB, and were translated to interval –the corresponding numbers from 1 to 5- for the statistical analysis.

As the experimental data would be ordinal with paired responses, we chose the Wilcoxon Signed-Rank Matched-Pairs test [Siegel 1956] to perform the power analysis for our study. Having a final sample size of 29 subjects (28 for some attributes due to a technical problem), we conducted an a priori power analysis with the software GPower [Faul and others 2007] to estimate the power and effect size that can be obtained with this number of observations. Aiming for the conventional threshold of a statistical power equal to 0.80, our sample size was sufficient to detect medium to large effects as defined by the thresholds set by Cohen [1988], with an effect size  $d_z$  of 0.57 and 0.56 for 28 and 29 observations, respectively.

The counterbalancing of conditions between participants and the conducted power analysis ensure an adequate experimental design and population size for the purpose of this study. Our main focus in this study is the effect of the environment: whether there is a difference in the responses between the real and virtual space, and whether there is a difference in the reported physical symptoms after the use of the VR headset. However, we will also test the effect of the presentation order of the environment to investigate potential bias on the experimental results.

### **3.5 Statistical analysis**

Most studies using ordinal subjective assessments to investigate the agreement between a real environment and its virtual representation use tests of analysis of variance [Mahdavi and Eissa 2002; Bishop and Rohrmann 2003; Newsham and others 2010]. Such tests provide evidence of whether the null hypothesis of no difference between groups can be rejected [Siegel 1956], but cannot prove that the null hypothesis is true, a problem that has been highlighted in the literature [Salters and others 2012; Murdoch and others 2015]. A one-sample Kolmogorov-Smirnov test revealed the non-normality of our data for all the studied variables, leading us to apply non-parametric statistical tests. Knowing the above mentioned limitations of hypothesis testing and following the approach suggested by Murdoch and others [2015], we employed two indicators for the analysis of the data with paired responses: the result of a Wilcoxon Matched-Pairs Signed Ranks test at the significance level  $\alpha$  of 0.05 and Cohen's  $d$  effect size. In order to establish a threshold for

an accepted mean difference between the evaluations of the two environments, we are using Cohen's  $d$  to assess the relative magnitude of the observed differences. For two paired samples  $x$  and  $y$ , Cohen's effect size  $d_z$ , is based on their paired differences,  $z = x - y$  and is calculated by dividing the mean difference of two conditions by the standard deviation of the distribution of the differences, as shown in Equation 1.

$$d_z = \frac{\mu_z}{\sigma_z} \quad (1)$$

An absolute  $d$  of 0.2 has been suggested by Cohen [1992] as a small effect size. Using this cut-off point which is commonly found in the literature, we can establish a threshold of maximum accepted absolute effect size  $|d_{\max}|$  equal to 0.2, corresponding to a small effect size. Although Cohen's threshold have been criticized as too low, and thus too generous as a contrast measure [Ferguson 2009; Lipsey and Hurley 2009], for our purpose the lowest threshold is the most conservative. We can thus define the combination of a Wilcoxon Matched-Pairs Signed Ranks result of no statistical difference at a significance level  $\alpha$  of 0.05 and a  $|d_z|$  effect size equal or lower than 0.2 as a result of accepted similarity between the paired responses for an attribute.

#### **4. EXPERIMENTAL PROTOCOL**

The experimental study was conducted with 29 participants, 13 women and 16 men. As one participant didn't respond to all the questions due to a technical problem, the sample size is reported separately for each studied attribute in the following sections. Of the 29 participants, three were aged between 21 and 25 years old, fifteen were aged between 26-30, ten were aged between 31-35, and two were aged over 35 years old. The duration of the experiment was 30 minutes per participant, conducted in scheduled appointments from 9:00 to 15:30. At the start of each session, subjects were asked to read a document containing information about the experiment and sign a form of consent in order to proceed. After this step, they were asked to respond to a series of demographic questions concerning their age, gender and vision correction, followed by questions in random order regarding their physical symptoms before the experimental session (Table 3, *Physical Symptoms*).

The participants were randomly assigned to evaluate the real or the virtual space first, counterbalancing the order of stimuli between subjects. In the second case, they entered the room with their eyes closed, ensuring that they saw only the virtual environment first. In both cases, the participants were guided to stand on a mark in the center of the room (Fig. 7, middle) and were told they could explore the space by rotating on this spot. After having explored the environment, they were asked a series of questions in randomized order regarding their perception (Table 3, *Perceptual Impressions*).



**Fig. 7. Photographs showing subjects exploring the virtual (left) and the real (right) space, and the experimental space with the mark for the participant's position, corresponding to the viewpoint in the virtual space (middle).**

This procedure was repeated for the evaluation of both environments, the latter after confirming with the participants that the headset was adjusted correctly. After the subjects had experienced both environments and responded to the equivalent questions, they were asked to remove the headset (if applicable) and respond to a questionnaire regarding their perceived presence (Table 3, *Reported Presence*). Lastly, they were asked once again to evaluate their physical symptoms with the same questionnaire that was used in the beginning of the experiment. In each session, an HDR photograph of the real space that was evaluated for the particular session was captured with a camera placed at the back of the room (Fig. 7, middle). The scene shown in virtual reality in each experimental session was selected from the pre-rendered scenes based on its similarity with the daylight conditions in the real space. An example of the two environments is given in Fig. 8.



**Fig. 8. Photograph of the real space taken from the participant's viewpoint (left) and the corresponding virtual environment (right).**

## **5. RESULTS**

The results of the experimental study are presented in four sections, introducing the perceptual accuracy of the VR method (subsection 5.1), the effect of using the VR headset on the users' physical symptoms (subsection 5.2), the perceived presence of subjects in the virtual space (subsection 5.3) and the effect of the presentation order of the environment on subjective evaluations (subsection 5.4).

## 5.1. Perceptual accuracy

For each of the studied attributes in perception, each participant evaluated the space shown in both the real and the virtual environment. We can thus evaluate the perceptual accuracy of the VR method by calculating the difference between this pair of responses for each participant; a difference of zero would mean that the paired evaluations are identical.

**TABLE 4. Frequency distribution of absolute differences between the subjective evaluations in the real and virtual environment**

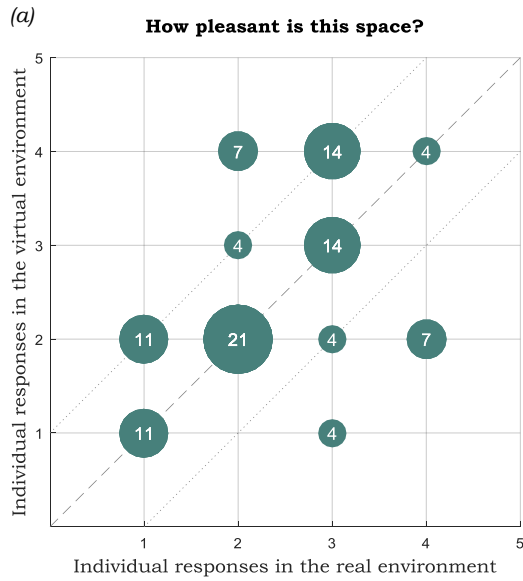
	Percentage of pairs with absolute difference					
	0	1	2	3	4	≤1
<i>How pleasant is this space?</i>	<b>50%</b>	32%	18%	0%	0%	<b>82%</b>
<i>How interesting is this space?</i>	<b>52%</b>	31%	14%	3%	0%	<b>83%</b>
<i>How complex is this space?</i>	<b>76%</b>	24%	0%	0%	0%	<b>100%</b>
<i>How exciting is this space?</i>	43%	46%	7%	4%	0%	<b>89%</b>
<i>How satisfied are you with the amount of view in this space?</i>	<b>52%</b>	45%	3%	0%	0%	<b>97%</b>

*Individual values are rounded to the nearest integer and are marked in bold if equal to or higher than 50%.*

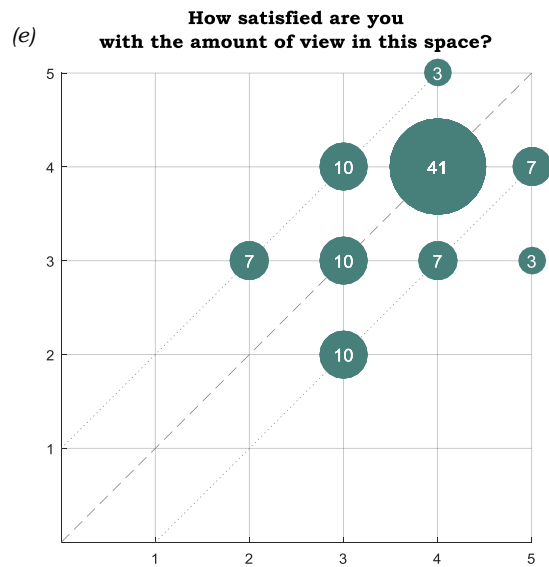
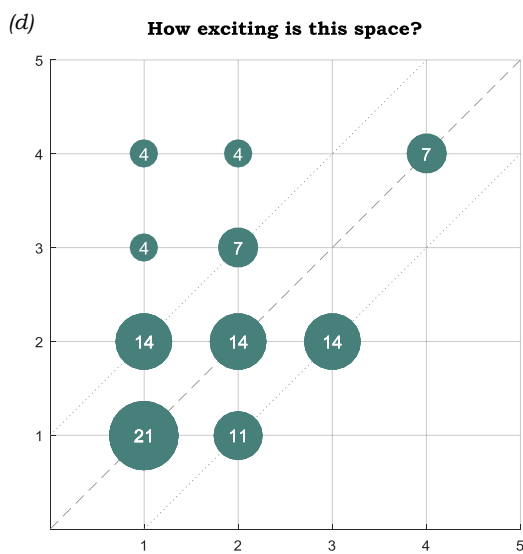
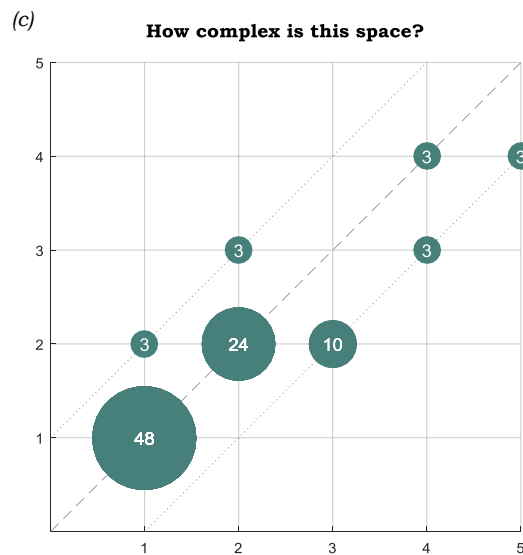
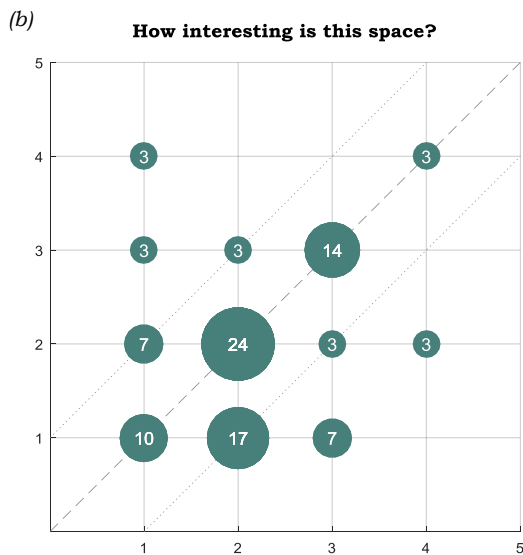
Table 4 shows the frequency distribution of the absolute differences between the paired responses in the two environments within our sample size. The pairs with a difference of zero equal to 50% or more of our sample for all questionnaire items, except for perceived excitement, while at least 82% of the responses for all questions had an absolute difference lower or equal to one between the pairs of evaluations in the real and virtual environment. These measures are particularly relevant for this study, as an absolute difference of one is the smallest possible non-zero difference when comparing responses in our verbal scale.

To better illustrate the agreement between responses, Fig. 9 shows a series of scatter plots with the paired responses of the participants in the real (x axis) and virtual (y axis) environment. In this graph, one can observe not only the agreement between the responses of the same participant -those that are closer to the diagonal line- but also the agreement of responses between participants in the different experimental conditions. Responses to questions of a more subjective nature (pleasant, interesting, and exciting) have a higher variation both within and between subjects.





**Fig. 9. Distribution of individual responses in the real (x axis) and virtual (y axis) environment for each of the studied attributes from each participant. The size and label of the circles correspond to the percentage of paired responses with identical values for the two evaluations, rounded to the nearest integer. The dashed and dotted diagonal lines mark the paired responses with a difference of zero and  $\pm 1$  between them, respectively.**



In contrast, the responses regarding the complexity and satisfaction with the amount of view show high agreement both between the responses in the real and the virtual space and also between participants, with a clustering of responses in the negative and positive range respectively. These observations show that a difference in the visual conditions – either between the two environments in one session, or between the experimental sessions- produced a stronger variation in impressions of pleasantness and visual interest.

**TABLE 5. Results of the Wilcoxon Matched-Pairs test for the effect of environment**

	N	$N_{diff}$	T	p-value
<i>How pleasant is this space?</i>	28	14	41	0.466
<i>How interesting is this space?</i>	29	14	64.5	0.460
<i>How complex is this space?</i>	29	7	20	0.453
<i>How exciting is this space?</i>	28	16	49	0.286
<i>How satisfied are you with the amount of view in this space?</i>	29	14	63	0.633

The statistical analysis of the responses with the Wilcoxon Matched-Pairs Signed-Rank test at a significance level  $\alpha$  of 0.05, shown in Table 5, indicates no significant differences between the responses in the real and the virtual environment for any of the studied variables. As discussed in the section 3.5, this test only uses the paired responses with a non-zero difference between them, marked as  $N_{diff}$  in Table 5, from a sample pool of N responses per question. This emphasizes the agreement between the responses in the two environments, as the result of non-statistically significant differences was observed in a sample much smaller than the original one; the remaining sample of paired responses had no difference between the evaluations in the two environments.

**TABLE 6. Overview of studied attributes and their perceptual accuracy based on the indexes of the Wilcoxon Signed-Ranks Matched-Pairs test and effect size  $d_z$**

	N	$N_{diff}$	$\mu_z$	$\mu  z $	$d_z$	
<i>How pleasant is this space?</i>	28	14	-0.179	0.679	-0.175	✓
<i>How interesting is this space?</i>	29	14	0.138	0.690	0.126	✓
<i>How complex is this space?</i>	29	7	0.103	0.241	0.212	(✓)
<i>How exciting is this space?</i>	28	16	-0.214	0.714	-0.208	(✓)
<i>How satisfied are you with the amount of view in this space?</i>	29	14	0.103	0.517	0.134	✓

An effect size  $d_z$  lower than or equal to 0.20, marked with ✓, indicates an acceptable result.

Although our data is ordinal, we are accepting the treatment of differences between the responses in the two presentation modes as interval. The inherent limitations of ordinal scales, such as the uncertainty of equidistance between the scale items due to subjectivity, are not present when considering the differences of the matched pairs where the participants act as a control for their own response. Assuming that each participant uses the same subjective distance in their evaluation of presentation modes, we can use the mean difference of the responses to gain insight into the similarity between the perception of virtual and real-world stimuli. Following the approach described in section 3.5, we are using two indexes to assess the perceptual accuracy of the studied method: the result of the Wilcoxon Signed-Ranks Matched Pairs test at a significance level  $\alpha$  of 0.05 and Cohen's  $|d_z|$  effect size. Since there were no observed statistically significant differences for any of

the studied attributes, a result of accepted perceptual accuracy corresponds to a  $|d_z|$  effect size equal or lower than our set threshold of 0.2, equivalent to a small effect size, and is symbolized with ✓ in Table 6.

From Table 6, we can see that three out of five attributes match our conditions for accepted perceptual accuracy. The attributes of perceived *complexity* and *excitement* fail to reach the threshold of a  $|d_z| < 0.20$  by a very small difference and can be considered marginally acceptable. We also briefly report the effect size Hedge's  $g$ , which is suggested as an unbiased form of Cohen's  $d$  and more suitable for small samples [Hedges and Olkin 1985]. The adjusted Hedge's  $g$  effect size is below the threshold of 0.20 for all the studied attributes, with the exception of perceived *excitement*. This difference with the results of using  $d_z$  is to be expected, as Hedge's  $g$  aims to correct for Cohen's  $d$  overestimation in small samples: for our analysis where an effect size below 0.2 is deemed as acceptable, Cohen's  $d$  is a more conservative measure.

In order to provide standardized measures for comparison with the literature, Table 6 contains both the mean difference  $\mu_z = \overline{x - y}$  and the mean absolute difference  $\mu_{|z|} = \overline{|x - y|}$  between the responses in the real and virtual environment for each studied attribute. It is worth noting that most of the mean differences  $\mu_z$  observed in our experiment are smaller than the maximum accepted mean differences in other relevant studies: 0.167 [Murdoch and others 2015], 0.667 [Moscoso, Matusiak, Svensson, and others 2015], 0.889 [Manyoky 2015], comparing values that are normalized based on a maximum possible difference of 4 units, corresponding to the 5-point scale used in our study. A discussion of possible inherent differences in such repeated measures due to the nature of the assessment can be found in the section 1.3 in the Appendix.

Another topic of interest is whether the agreement in the responses between the two environments could be a result of the repeated measures experimental design, as each participant was exposed to both environments. By comparing the evaluations of subjects that saw a similar condition (in this case, scenes with overcast sky) and using only the responses for the first environment they were exposed to, we can create a between-subjects dataset. A Wilcoxon Ranked-Sums test between 11 subjects that saw the real environment first and 13 that saw the virtual environment first, both groups in overcast sky conditions, indicated no significant differences between the two environments for a significance level  $\alpha$  of 0.05 for any of the studied attributes, in accordance with our findings from the paired-responses analysis.

In summary, the results indicate a high level of perceptual accuracy of the virtual reality method. A Wilcoxon Signed-Ranks Matched-Pairs test showed no statistically significant differences between the responses in the real and virtual environments for any of the studied attributes. In addition, the application of a second index for perceptual accuracy based on Cohen's  $d$  showed marginally acceptable results regarding how *complex* and how *exciting* the space was perceived, and acceptable for all other attributes, demonstrating that the perceptual evaluations in the virtual environment match closely those in the real environment.

## 5.2. Physical symptoms

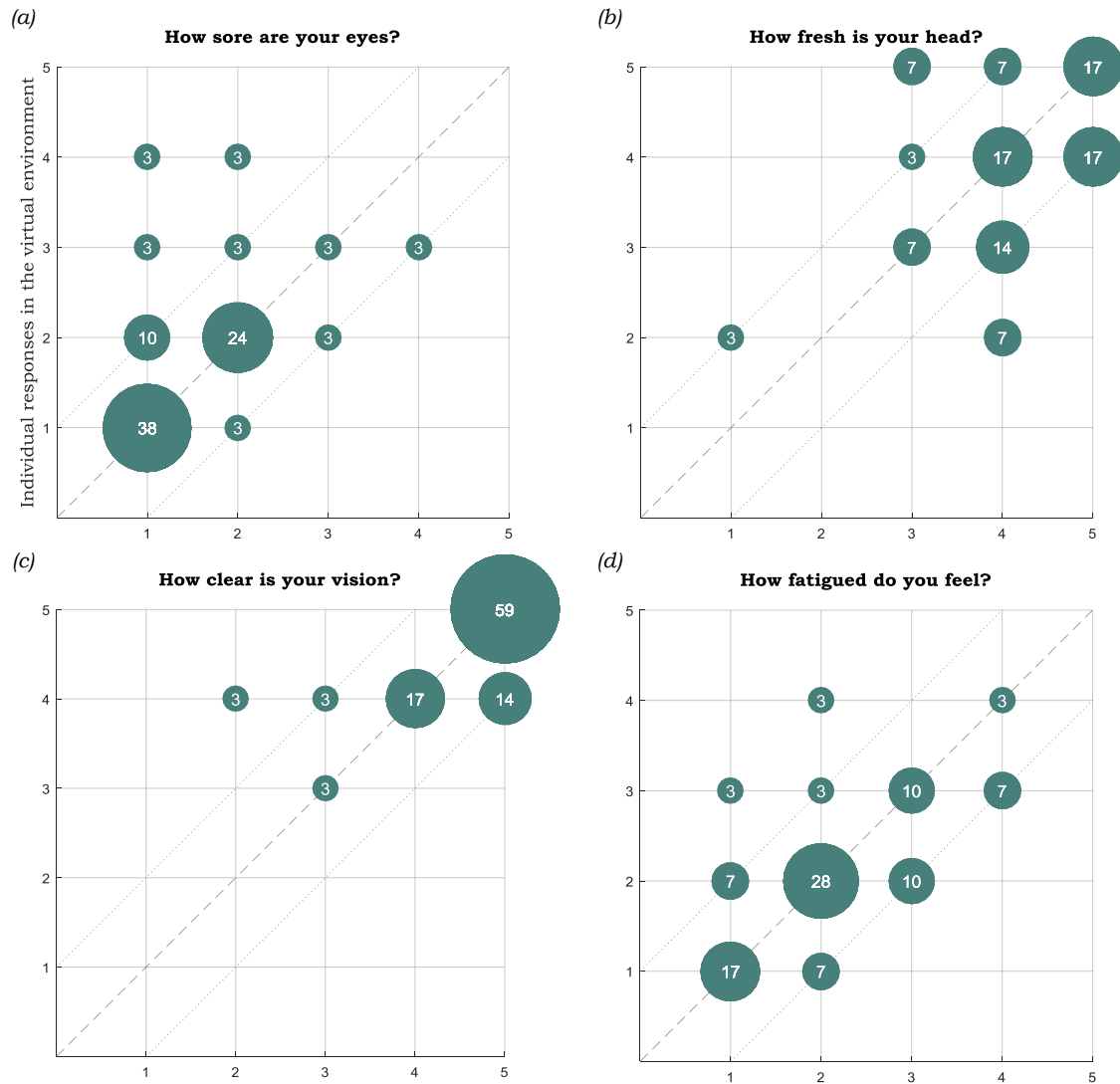
As the experimental data regarding the self-reported physical symptoms of the participants follows the same structure with *perceptual impressions*, consisting of paired ordinal responses before and after the experimental sessions, we will conduct the same analysis to investigate the possible adverse physiological effects from using the VR headset. The distribution of absolute differences within our sample size is shown in Table 7, binning the paired responses before and after the experimental session based on the absolute difference between the two evaluations.

**TABLE 7. Frequency distribution of absolute differences between the reported physical symptoms before and after the use of the virtual reality headset.**

	Percentage of pairs with absolute difference					
	0	1	2	3	4	$\leq 1$
<i>How sore do your eyes feel?</i>	<b>65%</b>	22%	6%	3%	0%	<b>87</b>
<i>How fresh does your head feel?</i>	41%	44%	14%	0%	0%	<b>85</b>
<i>How clear is your vision?</i>	<b>79%</b>	17%	3%	0%	0%	<b>96</b>
<i>How fatigued do you feel?</i>	<b>58%</b>	34%	6%	0%	0%	<b>92</b>

*Individual values are rounded to the nearest integer and are marked in bold if equal to or higher than 50%.*

With the exception of the question “*How fresh does your head feel?*”, more than 50% of the sample had identical responses before and after using the VR headset, while for all studied questions, more than 85% of the pairs had an absolute difference lower or equal to one. The distribution of the subjects’ responses are illustrated in Fig. 10, in a series of scatter plots where each data point corresponds to a pair of responses from the same participant, the x and y axes indicating the response before and after the experimental session, respectively.



**Fig. 10. Distribution of responses before (x axis) and after (y axis) the experimental session for each participant. The size and label of the circles correspond to the percentage of paired responses with identical values, rounded to the nearest integer. The dashed diagonal line marks the paired responses with a difference of zero while the dotted lines mark the responses with a difference of  $\pm 1$  in the evaluation before and after the use of the headset.**

We can observe a clustering of points in the top right quadrant of positive evaluations for both responses in the graphs (b) and (c), indicating a high agreement between the subjects. For all questions, there are few deviations from the diagonal lines marking a difference of zero and  $\pm 1$  between the paired responses, showing agreement within subjects in their responses before and after using the headset.

**TABLE 8. Results of the Wilcoxon Matched-Pairs Signed-Ranks test on the effect of using the VR headset on the participants' physical symptoms.**

	N	$N_{diff}$	T	p-value
<i>How sore do your eyes feel?</i>	29	10	12	0.131
<i>How fresh does your head feel?</i>	29	17	94	0.459
<i>How clear is your vision?</i>	29	6	12	1
<i>How fatigued do you feel?</i>	29	12	38.5	0.877

Similarly, the results of a Wilcoxon Matched-Pairs Signed-Ranks test, shown in Table 8, show no statistically significant differences for a significance level  $\alpha$  of 0.05 between the responses before and after the experimental session for the studied attributes.  $N_{diff}$  indicates the number of matched pairs with a non-zero difference used in the test, from the initial dataset of N pairs. Following the conservative approach described in section 3.5, we use two thresholds, an absolute effect size  $|d_z|$  equal or lower than 0.20, and the result of non-statistically significant differences shown in TABLE , to evaluate whether the use of the VR headset had an adverse effect on the subjects' self-reported physical symptoms.

**TABLE 9. Overview of the effect of using the VR headset on the participants' self-assessment of their physical symptoms for all questionnaire items.**

	N	$N_{diff}$	$\mu_z$	$\mu  z $	$d_z$	
<i>How sore do your eyes feel?</i>	29	10	-0.276	0.483	-0.313	
<i>How fresh does your head feel?</i>	29	17	0.172	0.724	0.172	✓
<i>How clear is your vision?</i>	29	6	0.035	0.241	0.061	✓
<i>How fatigued do you feel?</i>	29	12	0	0.483	0	✓

An effect size  $d_z$  lower than or equal to 0.20, marked with ✓, indicates an acceptable result.

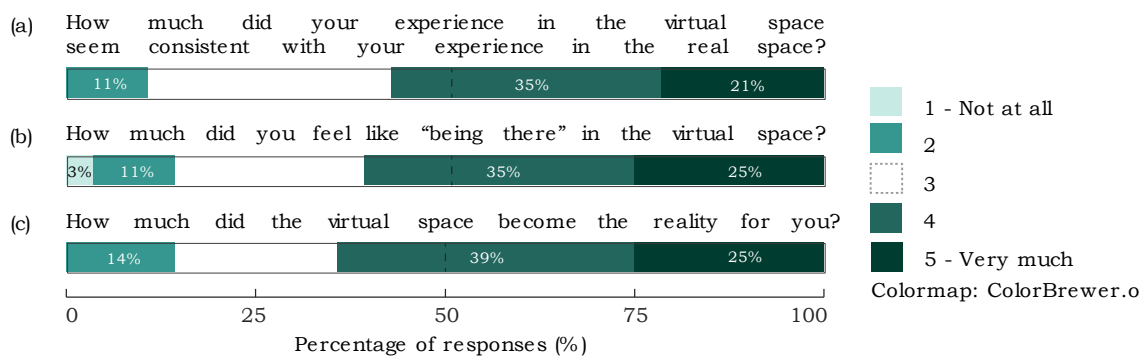
From the four questionnaire items, only the question “*How sore do your eyes feel?*” fails to meet these requirements. The results are identical when using Hedge’s  $g$  –suggested as an unbiased form of Cohen’s  $d$  [Hedges and Olkin 1985]- instead of  $d_z$ , as a complementary effect size measure. Lastly, the participants responded to the open question “*Do you have any other symptoms?*” (Table 3, PS.5) at the end of the experimental session. Out of 29 participants, 4 reported feeling slightly dizzy, while 25 reported no other symptoms, corresponding to 14% and 86% of our sample size. The responses to an open question at the end of the experimental session regarding any additional physical symptoms indicated slight dizziness as another minor effect, with 4 out of 29 participants reporting this as a symptom.

To summarize, the results regarding the participants' reported physical symptoms demonstrate no effect of using the VR headset, with the exception of the question “*How sore do your eyes feel?*”. A Wilcoxon Signed-Ranks Matched-Pairs test showed no statistically significant differences between the reported symptoms before and after using the headset for any of the questionnaire items. However, when using Cohen's  $d$  as a second index, the question “*How sore do your eyes feel?*” failed to meet our set threshold. Virtual-reality induced symptoms, often referred to as cybersickness, have been identified as a

recurrent issue in virtual reality [Sharples and others 2008]. Although a combination of subjective evaluations and physiological measurements is suggested as a more reliable assessment method [Kim and others 2005], the present experimental study relied on questionnaires to evaluate the perceptible symptoms that were experienced by the participants. Virtual reality applications that have low conflict between visual and proprioceptive senses, which is the case of our immersive virtual environment, are not expected to generate cybersickness (McCauley and Sharkey 1992), which is consistent with our own findings.

### 5.3. Perceived presence in the virtual reality environment

As discussed in earlier sections, the questionnaire items regarding the subjects' *perceived presence* in the virtual environment were designed to correspond to the three factors of presence defined by Schubert and others [2001], realness, involvement and spatial presence. The Cronbach's alpha reliability measure of 0.78 for the questionnaire indicates good internal consistency, showing that the items indeed measure the same structure [Cronbach 1951]. Figure 11 shows the distribution of the responses for each question.



**Fig. 11. Distribution of the participants' responses to the questions related to their perceived presence. Each stacked bar is labelled according to the percentage of responses in each point of the scale, out of a sample size of 29 participants. The midpoint of the scale (3) is not shown in order to highlight trends towards the extreme ends of the scale.**

Along with the distribution of responses, we are reporting the mean and standard deviation for each attribute –although our data is ordinal- to provide commonly used measures of comparison with the literature. From a sample size of 28 participants, 56% of the population responded in the positive range of the scale, while 11% responded in the negative range when evaluating the consistency of the experience between the real and the virtual environment ( $\mu=3.680$ ,  $\sigma=0.94$ ). Regarding their sense of being there in the virtual environment, 60% of the participants evaluated this attribute positively and 14% evaluated it negatively ( $\mu=3.685$ ,  $\sigma=1.09$ ), while 64% rated their sense of being there in the virtual environment in the positive range, contrary to 14% in the negative range ( $\mu=3.750$ ,  $\sigma=1.04$ ).

Using the mean of the responses as a point of reference, we can compare our results with those from the second experimental study described in Kuliga and others [2015], where participants engaged in a walk in a real environment and the corresponding 2D virtual environment projected on a large screen. Our reported mean values for the perceived

*involvement and consistency of experience* between the two environments are higher than those reported in the study by Kuliga and others, although by a very small margin. This is particularly interesting when considering the limitations mentioned in Kuliga and others regarding the limited realism and field of view of the virtual environment used in their study, possibly suggesting that the high interaction of the participants with the environment through their navigation in the virtual space compensated for these limitations.

#### 5.4. Effect of presentation order

The presentation order of the experimental stimuli has been identified as a possible influencing factor in relevant studies [Bishop and Rohrman 2003; Newsham and others 2010; Kuliga and others 2015]. In this section, we investigate the effect of presentation order on both the perceptual accuracy of the method and the reported presence of the participants.

We categorize the responses of the participants into two groups: responses of participants that saw the real environment first (RF) or virtual environment first (VF), with sample sizes  $N_{RF}$  and  $N_{VF}$  respectively, shown in Table 10. As in this analysis we are seeking a significant difference and using multiple questionnaire items, we use a more conservative threshold for statistical significance using a Bonferroni correction and dividing the conventional significance level  $\alpha$  of 0.05 with the number of items tested.

As we are interested in identifying an effect on the perceptual accuracy of the method, we test the differences between the paired responses in the two environments. By conducting a Wilcoxon Rank Sum test for independent ordinal data and using the conservative significance level  $\alpha'=0.05/5=0.01$ , we observe no statistically significant differences between conditions of presentation order for the perceptual accuracy of the studied attributes, with the exception of *perceived pleasantness* ( $p=0.000$ , ranksum statistic=134.5), as shown in Table 10.

**TABLE 10. Results of the Wilcoxon Rank Sum test for the effect of presentation order on the perceptual accuracy and perceived presence in the virtual space**

	$N_{RF}$	$N_{VF}$	T	p-value
Perceived Accuracy				
<i>How pleasant is this space?</i>	14	14	134.5	<b>0.000</b>
<i>How interesting is this space?</i>	15	14	220	0.816
<i>How complex is this space?</i>	15	14	231	0.903
<i>How exciting is this space?</i>	14	14	196.5	0.806
<i>How satisfied are you with the amount of view in this space?</i>	15	15	233	0.800
Perceived Presence				
<i>How much did your experience in the virtual space seemed consistent with your experience in the real space?</i>	14	14	179	0.251
<i>How much did you feel like 'being there' in the virtual space?</i>	14	14	191.5	0.649
<i>How much did the virtual space become the reality for you?</i>	14	14	189.5	0.517

A result of statistically significant difference,  $p < \alpha'$ , is marked with bold.

The means of the two datasets,  $\mu_{RF} = -0.714$  and  $\mu_{VF} = 0.357$ , are very enlightening in this case, as each dataset consists of the responses in the virtual space subtracted from their paired responses in the real space. A mean with a negative sign shows that the virtual environment was rated more positively than the real one and vice versa for a positive sign.



This finding indicates that the second environment that was experienced by the participant, independently of its type, was evaluated as more *pleasant*. Contrary to the findings of Newsham and others [2010], who identified an effect caused by a particular presentation order, reporting that there were fewer differences between real and virtual environments for participants who saw the real space first, we observed an effect of order which depends on the *sequence* rather of the *type* of presentation, impacting how *pleasant* the space was perceived.

We used the same procedure to test the effect of presentation order on the evaluations of the three questions regarding the reported presence in the virtual space, using a significance level of  $\alpha=0.05/3=0.016$ . No significant differences were observed between the two presentation orders for any of the three dimensions of presence. The absence of an effect of presentation order on the perceived *consistency in the experience of the two environments* possibly deviates from the results of Bishop and Rohrmann [2003], who found that participants that saw the real environment first evaluated the realism of the simulated environment more positively.

## **6. DISCUSSION**

The participants of the study were invited to discuss their thoughts on the experiment at the end of each experimental session. These comments, omitted for brevity, allow the identification of specific limitations of the current study. The participants pinpointed the difference in sky conditions in the two spaces and the lack of details in the virtual environment as a possible factor of discrepancies between the evaluations of the two environments. Another limitation was the representation of view from the window, as the scenery could change in the real space but not in the virtual one, as a result of the weather conditions or people passing. Although the statistical analysis of the subjective showed that the two environments were perceived very similarly, there were inevitable differences between the real conditions and those pictured in the virtual environment, such as the luminance of the scenes (Appendix 1.4). Even though it could be a limiting factor for the perceptual accuracy and reported presence in this experiment, this might not pose a problem for studies in the virtual environment if there is no need for immediate representation and comparison with the real environment. The positive results regarding the perceptual accuracy of VR method demonstrate that the immersive virtual environment could be used to adequately convey the visual experience of a real space for the studied perceptual attributes. These findings, along with the mobility offered by this technology, are encouraging for a wide range of possible applications, from education and practice in lighting design and architecture to lighting research, as a means to experience and evaluate luminous conditions in indoor spaces. However, the limited luminance range of the current head-mounted displays are a limiting factor for the investigation of visual discomfort, as it can be problematic to reproduce conditions inducing discomfort such as glare in the virtual reality environment.

Another important challenge in this study regards the tone-mapping of the scenes. High range dynamic images have to be compressed to the dynamic range of the selected display, making the subject of tone-mapping operators crucial in the creation of virtual

environments [Reinhard and others 2002; Salters and others 2012; Engelke and others 2013; Murdoch and others 2015].

The selection and application of a tone-mapping operator is particularly challenging in the context of an immersive virtual scene, as the content and contrast of the scene changes with the user's head movement. Current tone mapping operators are static and applied to the whole virtual scene, while an adaptive dynamic behaviour would correspond more accurately to human perception. Ongoing work from the authors investigates the perceptual accuracy of different tone-mapping algorithms for the representation of various lighting conditions in immersive virtual reality, using a newer device, the Oculus Rift CV1, which provides a higher resolution and refresh rate.

Due to the nature of this study, where we are seeking the absence of significant differences between the evaluations in the real and virtual environment, it is important to note that our findings cannot be generalized to other parameters without further investigation. As discussed in the section 3.5, our sample size was adequate for detecting medium to large effects, but limiting in identifying an effect of small magnitude. However, such an effect, as described by Cohen [1992], would be smaller than something noticeable by the naked eye of an observer, and thus would be unlikely to affect the overarching findings and usability of the proposed experimental method. Given the importance of validation for the establishment of any new method or technological advancement, further studies are encouraged to test the adequacy of immersive virtual reality scenes in different settings. Indicative settings could include omni-directional stereoscopic content, improvements on the luminance of the display, and different luminous conditions or levels of user interaction with the environment.

## 7. CONCLUSION

In this paper, the authors introduced a novel method for generating immersive virtual reality scenes based on photometrically accurate lighting simulations and tests the adequacy of this method as an alternative environment for subjective experiments on five aspects of the perception of daylight interior spaces (pleasantness, complexity, excitement, interest and satisfaction with the amount of view out). The adequacy of this method is assessed in three different areas: the perceptual accuracy of subjective evaluations in virtual scenes of daylight interiors, the effect of the virtual reality headset on the users' physiological reactions, and their perceived presence in the virtual environment.

While no significant differences were observed between the evaluations in the real and virtual environment or between the participants' reported physical symptoms before and after the experimental sessions in the respective non-parametric tests, the authors decided to adopt a more conservative approach, employing a small Cohen's *d* effect size as an additional threshold. Following the combination of the two indicators, the perceived impressions of *pleasant*, *interesting* and *satisfied with amount of view* were deemed sufficiently accurate, while *complex* and *exciting* had a marginally higher effect size. The use of the headset did not impact the self-reported physical symptoms of the participants based on the set threshold, with the exception of eye strain, corresponding to a small-to-

medium, rather than small, effect size. Lastly, the *reported presence* of the subjects in the virtual space showed very satisfactory results, comparable with other studies.

The experimental testing of the proposed method on the perceptual accuracy, adverse physiological effects and perceived presence demonstrated the adequacy of the produced virtual reality environment to be used as a surrogate to real daylight spaces in subjective experiments investigating the perception of the studied subjective attributes and highlighted its potential as a promising medium for empirical research, to be further studied against additional real conditions. Future work is encouraged to investigate the adequacy of the described method to evaluate other aspects of subjective perception and experience.

## **ACKNOWLEDGMENTS**

The authors would like to thank Prof. Steve Fotios for his valuable feedback, Prof. Jean-Marie Fürbringer and Peter Hansen for their technical support in the design of experiments, Lorenzo Cantelli for his help with the VR headset, and Dr. Parag Rastogi for his contribution in the analysis of solar radiation for the sky descriptions.

## **FUNDING**

This work was funded by the École Polytechnique Fédérale de Lausanne (EPFL) and through a grant awarded by the Velux Stiftung Foundation in the framework of Project 1022: "Identifying the impact of regional differences on the perceived quality of daylight architectural spaces: A comparison study across different latitudes".

## **DISCLOSURE STATEMENT**

The authors report no declarations of interest.

## **APPENDIX**

### **1.1. Radiance sky description for the virtual scenes**

The *gendaylit* script in Radiance can generate a sky description based on given sky radiation measures, date and local standard time. As the experiment would take place in November, we calculated the diffuse horizontal and direct normal irradiance for overcast and clear skies in this month. Using the Geneva 067000 (IWEC) EPW weather file, we selected the days in November with overcast and clear skies and their hourly values were used to interpolate the diffuse horizontal and direct normal irradiance for every half hour. The values corresponding to the hour and sky type selected for our simulations, shown in Table 11, were then used in the *gendaylit* script with the *-W* option to create descriptions of sky for each scene. The two scenes with overcast sky had an identical sky description and different images mapped on the Radiance sky.

**TABLE 11.** Diffuse horizontal irradiance (DHI) and direct normal irradiance (DNI) used in *gendaylit* to generate the description of sky for each of the rendered scenes.

Time of day	9:30	10:30	11:30	12:30	13:30	14:30	15:30	12:30
Sky type	clear	clear	clear	clear	clear	clear	clear	overcast
DHI	493.6	686.8	742.5	86.6	737.1	729.6	653.5	486.6
DNI	93.3	84.5	86.6	92.6	86.1	83.3	74.6	3

## 1.2. Material properties for the cubemap projection

In order to have full control over the environment projected in VR, we applied an unlit two-sided material to the cube, ensuring that the textures were unaffected by lighting sources in Unity and the scene appeared correctly from the user’s point of view. The material is created by manipulating the default *Unlit Shader* in Unity and adding the option *Cull off*, which enables the rendering of all faces of the objects in the Unity scene.

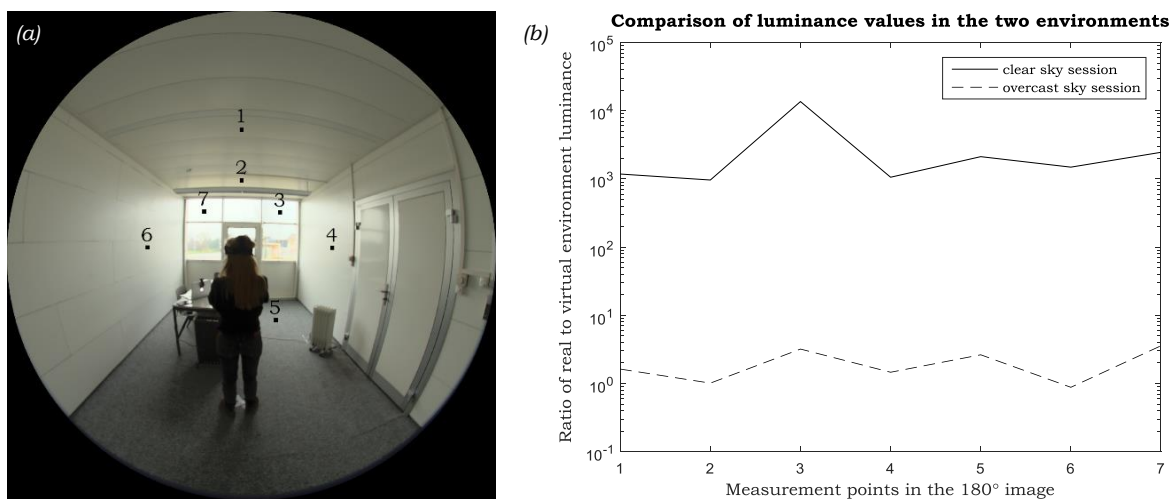
## 1.3. Inherent variation in subjective responses

A limitation of this experiment is the lack of repeated measures from the same participant for the same stimulus, which could provide information on the amount of variance that is expected in any assessment due to the subjective nature of questionnaire responses. To this end, we conducted a follow-up study including a randomized repeated verbal assessment for the same scene, either in the real or the virtual environment, and compared the responses of participants between their first and second evaluation. From a total of 56 paired responses, the mean absolute differences between the repeated measures for the combined attributes of how *pleasant*, *interesting* and *exciting* the space was perceived,  $\mu |z|_{repeated}$ , was 0.325, normalized for a 5-point scale. Interestingly, this value, while averaged for the three attributes of *pleasantness*, *interest* and *excitement*, corresponds roughly to half of our reported mean absolute differences between the real and the virtual environment shown in Table 6. Using this difference as a reference could help explain whether a part of the observed variance between environments is inherent in the nature of subjective evaluation, and to which extent it is an effect of the virtual reality medium.

## 1.4. Luminance comparison in the real and virtual environments

In order to provide a measure of the luminance discrepancy between the real environment and its virtual representation projected in the Oculus DK2, we compare the luminance in 7 reference points between the two environments. First, we group the 29 experimental sessions based on the sky type of the projected virtual scene in the session, either clear (N=5) or overcast sky (N=24). Using as reference the 180° HDR photograph of a randomly selected session from each group, transformed into an angular fisheye projection, we simulate the scene from the same viewpoint with the photograph as a 180° fisheye image with the equivalent rendering settings and ambient data as the virtual scene that was shown in this particular session. The luminance in the real space is directly measured from the HDR photograph, while for the projected images of the virtual space it is derived using the response curve of the Oculus DK2 display. As our reference for the luminance in the real space is an HDR photograph taken from a viewpoint behind the subject (Fig. 7), rather

than from their point of view, we choose to reproduce the virtual scene of the session from the same viewpoint. Although the resulting luminance measurements do not directly correspond to those from the subject's point of view, they allow for comparable assessment of the luminance deviation between the real and virtual scenes, shown in Fig. A1.



**Figure A1. Reference points on the fisheye images (a) and ratio of the luminance measurements in those points between the real and virtual environments for the equivalent sessions in both sky types, using a logarithmic scale for the y axis (b).**

## REFERENCES

- Amundadottir ML, Rockcastle S, Sarey Khanie M, Andersen M. 2017. A human-centric approach to assess daylight in buildings for non-visual health potential, visual interest and gaze behavior. *Build. Environ.* 113:5–21.
- Baños RM, Botella C, Rubió I, Quero S, García-Palacios A, Alcañiz M. 2008. Presence and Emotions in Virtual Environments: The Influence of Stereoscopy. *Cyberpsychol. Behav.* 11:1–8.
- Bishop I., Rohrmann B. 2003. Subjective responses to simulated and real environments: a comparison. *Landsc. Urban Plan.* 65:261–277.
- Boyce PR. 2003. *Human Factors in Lighting*. London and New York: Taylor & Francis.
- Bülow-Hübe H. 1995. Subjective reactions to daylight in rooms: Effect of using low-emittance coatings on windows. *Light. Res. Technol.* 27:37–44.
- Cauwerts C. 2013. *Influence of presentation modes on visual perceptions of daylit spaces (Doctoral Dissertation)*. Université catholique de Louvain.
- Cauwerts C, Bodart M. 2011. Investigation of 3D projection for qualitative evaluation of daylit spaces. In: *Proceedings of PLEA 2011*. Louvain-la-Neuve, Belgium.
- Chamilothori K, Wienold J, Andersen M. 2016. Daylight patterns as a means to influence the spatial ambience: a preliminary study. In: *Proceedings of the 3rd International Congress on Ambiances*. Volos, Greece.
- Charton V. 2002. *Etude comparative de la perception d'ambiances lumineuses en milieu réel et en milieu virtuel. [Doctoral Dissertation]*. [Lyon]: L'Institut National des Sciences Appliquées de Lyon.
- Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.
- Cohen J. 1992. Statistical power analysis. *Curr. Dir. Psychol. Sci.* 1.
- Cronbach LJ. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297–334.
- De Kort YA, Ijsselstein WA, Kooijman J, Schuurmans Y. 2003. Virtual laboratories: Comparability of real and virtual environments for environmental psychology. *Presence Teleoperators Virtual Environ.* 12:360–373.

- Diemer J, Alpers GW, Peperkorn HM, Shiban Y, Mühlberger A. 2015. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Front. Psychol.* 6.
- Engelke U, Stokkermans MGM, Murdoch MJ. 2013. Visualizing lighting with images: converging between the predictive value of renderings and photographs. In: Rogowitz BE, Pappas TN, de Ridder H, editors. p. 86510L.
- Faul F, Erdfelder E, Lang A-G, Buchner A. 2007. G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39:175–191.
- Ferguson CJ. 2009. An effect size primer: A guide for clinicians and researchers. *Prof. Psychol. Res. Pract.* 40:532–538.
- Flynn JE, Spencer TJ, Martyniuk O, Hendrick C. 1973. Interim Study of Procedures for Investigating the Effect of Light on Impression and Behavior. *J. Illum. Eng. Soc.* 3:87–94.
- Franz G, von der Heyde M, Bühlhoff HH. 2005. An empirical approach to the experience of architectural space in virtual reality—exploring relations between features and affective appraisals of rectangular indoor spaces. *Autom. Constr.* 14:165–172.
- Greene N. 1986. Environment Mapping and Other Applications of World Projections. *IEEE Comput. Graph. Appl.* 6:21–29.
- Hedges L, Olkin I. 1985. Statistical Methods in Meta-Analysis. In: *Stat Med.* Vol. 20.
- Hendrick C, Martyniuk O, Spencer TJ, Flynn JE. 1977. Procedures for Investigating the Effect of Light on Impression Simulation of a Real Space by Slides. *Environ. Behav.* 9:491–510.
- Heydarian A, Carneiro JP, Gerber D, Becerik-Gerber B, Hayes T, Wood W. 2015. Immersive virtual environments versus physical built environments: A benchmarking study for building design and user-built environment explorations. *Autom. Constr.* 54:116–126.
- Heydarian A, Pantazis E, Carneiro JP, Gerber D, Becerik-Gerber B. 2015. Towards Understanding End-user Lighting Preferences in Office Spaces by Using Immersive Virtual Environments. In: *Proceedings of the 2015 International Workshop on Computing in Civil Engineering.* Austin, Texas: American Society of Civil Engineers.
- Holl S, Kwinter S, Safont-Tria J. 2011. *Color, Light, Time.* 1 edition. Müller, Lars.
- Kim YY, Kim HJ, Kim EN, Ko HD, Kim HT. 2005. Characteristic changes in the physiological components of cybersickness. *Psychophysiology* 42:616–625.
- Kuliga SF, Thrash T, Dalton RC, Hölscher C. 2015. Virtual reality as an empirical research tool — Exploring user experience in a real building and a corresponding virtual model. *Comput. Environ. Urban Syst.*
- Lambooij M, IJsselsteijn W, Bouwhuis DG, Heynderickx I. 2011. Evaluation of Stereoscopic Images: Beyond 2D Quality. *IEEE Trans. Broadcast.* 57:432–444.
- Larson GW, Shakespeare RA. 1998. *Rendering with Radiance - The Art and Science of Lighting Visualization.* Morgan Kaufmann Publishers.
- Lessiter J, Freeman J, Keogh E, Davidoff J. 2001. A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory. *Presence Teleoperators Virtual Environ.* 10:282–297.
- Lipsey M, Hurley S. 2009. Design Sensitivity: Statistical Power for Applied Experimental Research. In: *The SAGE Handbook of Applied Social Research Methods.* 2nd ed. Thousand Oaks: SAGE Publications, Inc. p. 44–76.
- Loomis JM, Blascovich JJ, Beall AC. 1999. Immersive virtual environment technology as a basic research tool in psychology. *Behav. Res. Methods Instrum. Comput.* 31:557–564.
- M. Loomis J, Da Silva JA, Philbeck J, Fukusima S. 1996. Visual Perception of Location and Distance. *Curr. Dir. Psychol. Sci.* 5:72–77.
- Mahdavi A, Eissa H. 2002. Subjective Evaluation of Architectural Lighting via Computationally Rendered Images. *J. Illum. Eng. Soc.* 31:11–20.
- Manyoky M. 2015. Visual-acoustic simulation for landscape impact assessment of wind parks. Diss., Eidgenössische Technische Hochschule ETH Zürich, Nr. 22725.
- McNeil A. 2016. *Rendering omni-directional stereo images with Radiance.*
- Moscoso C, Matusiak B, Svensson UP. 2015. Impact of window size and room reflectance on the perceived quality of a room. *J Arch Planning Res.*

- Moscoso C, Matusiak B, Svensson UP, Orleanski K. 2015. Analysis of Stereoscopic Images as a New Method for Daylighting Studies. *ACM Trans. Appl. Percept.* TAP 11:21.
- Murdoch MJ, Stokkermans MGM, Lambooi M. 2015. Towards perceptual accuracy in 3D visualizations of illuminated indoor environments. *J. Solid State Light.* 2:1.
- Newsham GR, Cetegen D, Veitch JA, Whitehead L. 2010. Comparing lighting quality evaluations of real scenes with those from high dynamic range and conventional images. *ACM Trans. Appl. Percept.* 7:1–26.
- Newsham GR, Richardson C, Blanchet C, Veitch JA. 2005. Lighting quality research using rendered images of offices. *Light. Res. Technol.* 37:93–115.
- Oculus Rift Developer Guide, Version 0.6.0.0. 2015.
- Parpaire K, Baker NV, Steemers KA, Compagnon R. 2002. The Luminance Differences index: a new indicator of user preferences in daylight spaces. *Light. Res. Technol.* 34:53–66.
- Preston CC, Colman AM. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychol. (Amst.)* 104:1–15.
- Reinhard E, Stark M, Shirley P, Ferwerda J. 2002. Photographic Tone Reproduction for Digital Images. In: *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA: ACM. [SIGGRAPH '02]. p. 267–276.
- Reinhart CF. 2005. *Tutorial on the Use of Daysim/Radiance Simulations for Sustainable Design*. Ottawa, Canada: NRC.
- Rockcastle S, Amundadottir ML, Andersen M. 2016. Contrast measures for predicting perceptual effects of daylight in architectural renderings. *Light. Res. Technol.*
- Rockcastle S, Andersen M. 2014. Measuring the dynamics of contrast & daylight variability in architecture: A proof-of-concept methodology. *Build. Environ.* 81:320–333.
- Rockcastle S, Chamilothoni K, Andersen M. 2017. Using Virtual Reality to Measure Daylight-Driven Interest in Rendered Architectural Scenes. In: *Proceedings of Building Simulation 2017*. San Francisco, CA, USA.
- Russell JA. 1980. A circumplex model of affect. *J. Pers. Soc. Psychol.* 39:1161–1178.
- Salters B, Murdoch M, Sekulovski D, Chen S-H, Seuntjens P. 2012. An evaluation of different setups for simulating lighting characteristics. Vol. 8291. p. 82911M–82911M–13.
- Schubert T, Friedmann F, Regenbrecht H. 2001. The Experience of Presence: Factor Analytic Insights. *Presence Teleoperator Virtual Environ.* 10:266–281.
- Sharples S, Cobb S, Moody A, Wilson JR. 2008. Virtual reality induced symptoms and effects (VRISE): Comparison of head mounted display (HMD), desktop and projection display systems. *Displays* 29:58–69.
- Shibata T, Kim J, Hoffman DM, Banks MS. 2011. The zone of comfort: Predicting visual discomfort with stereo displays. *J. Vis.* 11:11–11.
- Siegel S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. 1st edition. New-York: McGraw-Hill.
- Slater M, Wilbur S. 1997. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence Teleoperators Virtual Environ.* 6:603–616.
- Spek ED v d, Houtkamp JM. 2008. Mixing Emotions, How Physical Discomfort Influences the Affective Appraisal of Virtual Places. In: *2008 International Conference Visualisation*. p. 18–22.
- Van Erp T. 2008. *The effects of lighting characteristics on atmosphere perception [Master's thesis]*. Eindhoven University of Technology.
- Vogels I. 2008. Atmosphere Metrics - Development of a Tool to Quantify Experienced Atmosphere. In: Westerink JHDM, editor. *Probing Experience: From Assessment of User Emotions and Behaviour to Development of Products*. Dordrecht, the Netherlands: Springer. p. 25–41.
- Ward Larson G. 1994. The RADIANCE lighting simulation and rendering system. In: *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM. [SIGGRAPH '94]. p. 459–472.
- Zumthor P. 2006. *Atmospheres : architectural environments - surrounding objects*. Basel: Birkhäuser - Publishers for Architecture.