Building and Environment 146 (2018) 258–267

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv

A Bayesian method of evaluating discomfort due to glare: The effect of order bias from a large glare source

M.G. Kent[a,*], T. Cheung[b], S. Altomonte[c], S. Schiavon[d], A. Lipczyńska[b]

[a] Department of Architecture and Built Environment, University of Nottingham, Nottingham, UK
[b] Berkeley Education Alliance for Research in Singapore, Singapore
[c] Architecture et Climat, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
[d] Centre for the Built Environment, University of California, Berkeley, CA, USA

ARTICLE INFO

Keywords:
Bayesian method
Discomfort glare
Experimental bias
Order bias
Statistical analysis

ABSTRACT

Replicating scientific findings is a fundamental aspect of research. However, in studies of discomfort due to glare, it is difficult to make comparisons between the results of different experiments since the statistical tests usually reported do not allow independent findings to be directly compared to each other. Here we present an alternative Bayesian approach that can address this problem. To show how this approach works, we performed a laboratory test with 55 participants to validate the effect of order bias previously detected in a similar study evaluating discomfort due to glare but, this time, under a large luminous source. Using the luminance adjustment procedure, the glare source was varied to meet four sensations of discomfort due to glare. Adjustments were performed under three different order sequences: ascending, descending, and randomised. Test participants provided glare settings using a newly proposed evaluation scale. The effect of order bias detected in the original study was compared to the data obtained with the same methodological procedure in the new experiment using Bayesian inferential tests. The results showed a close replication, highlighting that the order bias effect found in the original study was also present in the new experiment. The wide application of Bayesian methods in the design and analysis of experimental studies may improve the accuracy and validity of glare models.

1. Introduction

Discomfort due to glare is one of the challenges of building façade design. While studies have found that visual discomfort is a significant problem in many conventional buildings, occupants have reported glare five times more often in green-rated buildings [1]. A study based on 2540 occupant responses, collected from 11 countries and 36 different "sustainable" buildings, has also shown that glare – particularly from daylit windows – remains a pertinent issue [2]. To minimise the risk of glare, various models have been developed to provide precise measures of discomfort from a visual scene, with the objective of quantifying the perceived levels of glare based on physical measurements [3]. However, these models often give a low prediction accuracy [4]. Among many models recommended in the literature and in international standards, Table 1 presents a selection of key experimental studies used to derive prediction models of discomfort glare, also illustrating the subjective criteria that observers used to evaluate the glare sources.

From a methodological perspective, the studies presented in Table 1 [5–7] – together with many others – relied invariably on frequentist approaches (e.g., null hypothesis significance testing (NHST)) to analyse the predictive performance of the models proposed. However, NHST testing has several limitations, such as:

- Statistical significance is dependent on both the size of the sample and the magnitude of the effect, which cannot be measured using NHST alone [8,9]. This implies that, when large samples are used, statistically significant findings can be detected even though the magnitude of the effect is not practically relevant. For example, Altomonte and Schiavon [10] showed that even the smallest variations in occupant satisfaction scores between LEED and non-LEED rated buildings produced highly significant differences ($p \leq 0.001$) due to a large amount of sample data available ($n = 21\,250$).
- NHST tests do not provide any evidence that two or more studies will produce similar findings (i.e., no reliable information about the replication of experimental findings). In fact, when replicating an effect across studies with fixed sample sizes, but with different observers, statistical significance levels (p-values) can vary considerably [11]. Even small changes to the means, correlation

* Corresponding author.
E-mail address: michael.kent2@nottingham.ac.uk (M.G. Kent).

https://doi.org/10.1016/j.buildenv.2018.10.005
Received 13 June 2018; Received in revised form 1 October 2018; Accepted 3 October 2018
Available online 05 October 2018

Table 1
Key studies of discomfort due to glare.

Study	Prediction Model	Evaluation Criteria
Petherbridge and Hopkinson (1950)	Glare Constant/IES-GI	Multiple-Criterion Scale
Hopkinson and Bradley (1960)	Daylight Glare Index (DGI)	Multiple-Criterion Scale
Wienold and Christoffersen (2006)	Daylight Glare Probability (DGP)	4-Point Glare Scale

coefficients, or regression coefficients can lead to large variations in the calculated *p*-values, and therefore on the conclusions that are drawn from the data [9]. This can be problematic when comparisons are made between significant and non-significant results [12].

- Since differences in statistical significance (*p*-values) are not always statistically significant themselves, the comparisons made when using NHST analyses can often be misleading [9,13].

These conflicts arise also in discomfort glare research, whereas the strength of the significant relationships detected between evaluations of visual discomfort and calculated glare index values can vary considerably across different studies, even when the same prediction model has been used (e.g. [14–17]). For example, Tuaycharoen and Tregenza [18] showed that the correlation coefficients (r = 0.72–0.86) measuring the relationship between calculated Daylight Glare Index (DGI) values and the evaluations of discomfort due to glare reported by observers on Hopkinson's multiple-criterion scale were statistically significant ($p \leq 0.01$). Conversely, similar studies [19,20], which also used the DGI and the multiple-criterion scale, reported smaller correlation coefficients (r = 0.28–0.56) and did not show a statistically significant ($p > 0.05$) relationship between the same variables. Since results from separate studies often lead to inconsistent conclusions, we believe that NHST should not be used as the sole analysis method to support the statistical inferences derived from discomfort glare experiments.

The use of different statistical tools that can build on the work of previous studies may lead to a more robust and reliable characterisation of discomfort due to glare. An alternative approach to NHST is to use a Bayesian framework, whereby information from previous studies can be used to inform the analysis of data obtained in a new experiment [21]. Bayes' rule describes the probability of the occurrence of an outcome based on the conditions that might be related to it [22], positing how a degree of belief from previous knowledge should change once accounting for new evidence [23]. Bayesian inference treats data as fixed and model parameters as random variables [24]. A Bayesian approach is, thus, distinctly different from frequentist statistics, since it assumes that each unknown parameter has a *posterior* probability distribution that describes the uncertainty about population parameter values. The aim of the analysis is to estimate the posterior distribution given the data. The posterior density is the normalised product of a *prior* distribution, reflecting initial beliefs, and the *likelihood* from the data [25]. Once new data are collected, the prior is combined with the likelihood to produce a posterior distribution. In so doing, Bayesian models of analysis can deal with the complexity of real-world settings and overcome some of the limitations of controlled laboratory studies [26]. Clearly, since the Bayesian approach relies on knowledge from previous research, for it to be applied to inform new experimental studies there is a need to make data publicly available along with the original study findings. For example, in a recent article, Bayesian inferences were applied to analyse the effect of personalised control systems on the levels of visual satisfaction in daylit offices [27]. Using this analytical approach, previous knowledge of human visual preferences was combined with newly collected information to develop personalised visual satisfaction profiles within private workstations. Another important application of a Bayesian approach is to examine whether a new experiment can successfully replicate the results found in earlier research [13].

One further critical aspect of any scientific investigation is to verify whether the conclusions drawn from an original study were not the result of an experimental or analytical bias (i.e., a random error). In the context of discomfort glare research, we previously identified a substantive effect of order bias (i.e., the sequence in which the magnitude of discomfort glare was evaluated using a multiple-criterion scale and a luminance adjustment task) in the procedure used by Petherbridge and Hopkinson [6] to obtain the Glare Constant, which is at the basis of many successive glare models [28]. To ensure that our previous conclusions – based on an experimental setup using a small glare source – were not the result of a random error, a new experiment was designed. We applied a Bayesian approach to validate the previously detected effect of order bias, using a Hopkinson-like luminance adjustment task but under slightly modified experimental conditions. Informing the alternative hypothesis with the data from the earlier experiment [28], we used the same procedure in a new experimental setting with a large artificial window and a different sample of test participants. On this basis, in this paper we aim to: (1) demonstrate how a Bayesian approach can be used as a suitable alternative to NHST when analysing experimental findings derived from independent glare studies; (2) replicate the detection of the order bias effect when using a luminance adjustment procedure to evaluate the subjective degrees of discomfort due to glare from a large glare source. Therefore, while the effect of order bias is of relevant interest to this study, it was used primarily to illustrate how a Bayesian approach can reinforce the experimental conclusions derived from independent discomfort glare studies.

2. Method

2.1. Experimental setting

The new experiment was carried out in a test room located at the Berkeley Education Alliance for Research in Singapore (BEARS) centre, within the SinBerBEST testbed (Fig. 1). The room contained an artificial window (known as Daylight Emulator (DLE)), backlit by an array of cool and warm LEDs, capable of emitting light with a spectral composition approximate to daylight. The DLE provided a variable luminance in the range between 952 and 10 005 cd/m^2.

The artificial window featured three separate panes of glass, each of 0.98 × 2.15 m^2, surrounded by metal sill frames of 0.08 m in width and depth. Behind each glass pane, a fabric sheet membrane was mounted in front of the DLE. The membrane had uniform transmission properties allowing direct light from the DLE to be evenly distributed across the area of the window. Each window pane was equipped with a fabric roller blind, which remained fully retracted during the experiment. The room surfaces had reflectance properties of: ρ_{wall} = 0.56, ρ_{floor} = 0.72, $\rho_{ceiling}$ = 0.72 (these were estimated using the Munsell system). A workstation (chair, desk, and desktop computer) was placed inside the room at a position parallel to the window. This arrangement was informed by the study from Osterhaus and Bailey [29] and was preferred over an internal spatial organisation where the workstation was orthogonal to the artificial window. Since a parallel arrangement produced higher glare index values, we believed this would have increased the likelihood of detecting an order bias effect. The surface of the desk had a reflectance of ρ_{desk} = 0.56, dimensions of 1.80 × 0.75 m^2, and a height of 0.74 m from the floor. We used a flat 24" liquid crystal display computer screen (hp zdisplay z24i, mean self-luminance = 150 cd/m^2) to present visual tasks to test participants. The screen was mounted on the desk top.

2.2. Photometric measurements

We used a Charged Coupled Device (CCD) Canon EOS 7OD camera with a 4.5 mm f/2.5 EX DX GSM 180° sigma fish-eye-lens, a luminance

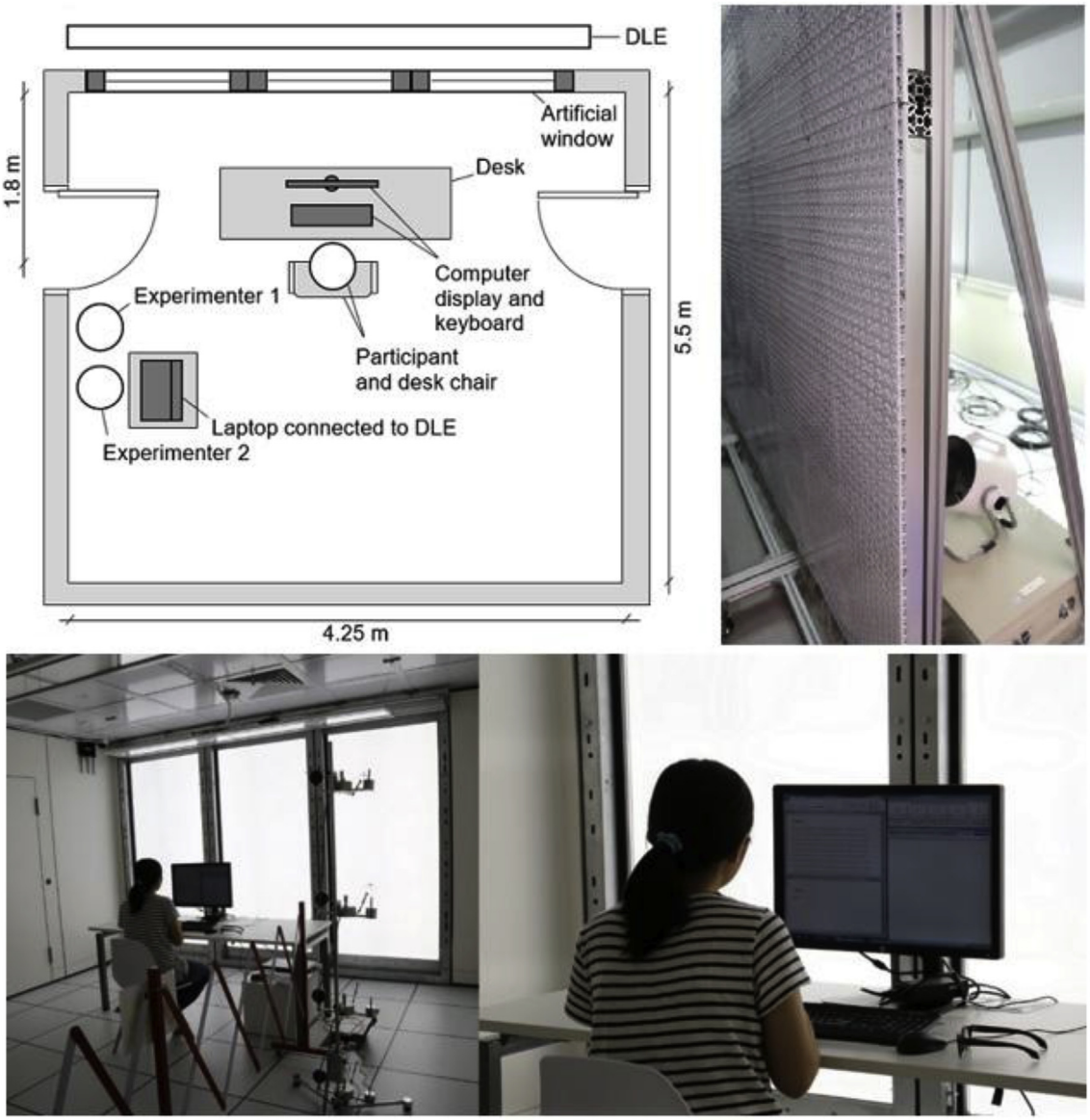

Fig. 1. Plan view of the experimental setup showing the position of the experimenters and of the participant (*top left*). Note: Both entrance doors to the test room were kept closed during the experiment; Image of the DLE showing an array of warm and cool LEDs with a cooling system used to prevent the DLE from overheating (*top right*); A test participant sat at the viewing position performing a visual task (*bottom left and right*).

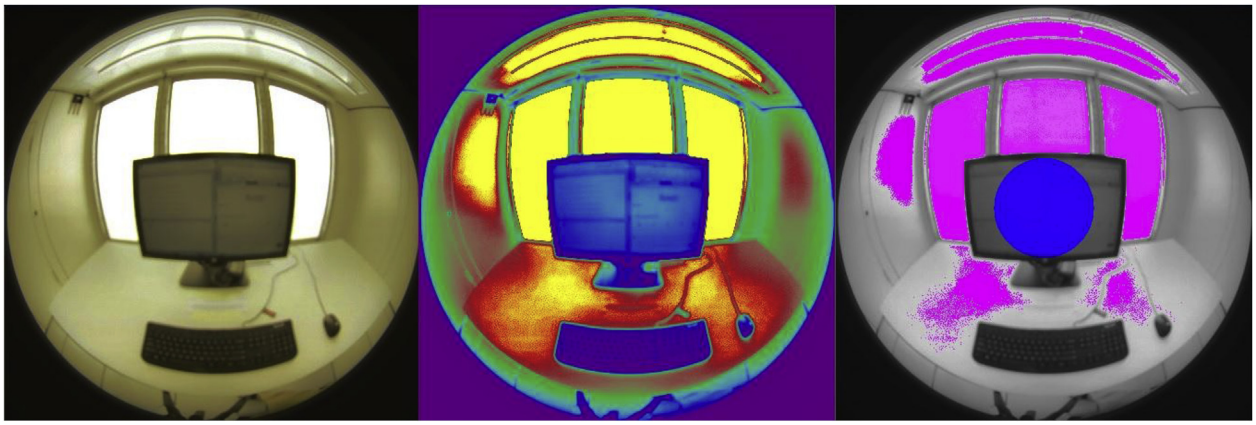

Fig. 2. Example of HDRI constructed from the seven LDRIs captured using the CCD camera (*left*); False colour Photosphere luminance map with Radiance image formatting (*centre*); *Evalglare* image with a blue circle at the centre of the screen representing the point of visual fixation. Pixels highlighted in pink surrounding the screen represent identified glare sources (*right*). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

meter (LS-100, Minolta, Japan – with an accuracy ± 2% cd/m^2) and an illuminance chromameter (CL-200a, Minolta, Japan – with an accuracy ± 2% lux) to obtain photometric measurements. The luminous conditions were captured using a series of Low Dynamic Range Images (LDRI) with varying exposure values and vertical illuminance measurements taken from the viewing position.

The artificial window luminance could be gradually increased or decreased using a Digital Multiplex (DMX) controller, which was operated by the experimenters via a customised software on a laptop. To achieve precise photometric measurements in repeated procedures, the luminance output of the window needed to be calibrated with the software. To do this, the DMX controller was increased at evenly adjusted intervals at which seven LDRIs were captured using the CCD camera, while several spot-point luminance measurements and a single vertical illuminance and correlated colour temperature (CCT) readings were also collected [30]. The seven LDRIs taken at each DMX interval were combined into a Radiance-formatted High Dynamic Range Image (HDRI) using the software *Photosphere*, which merged several LDRIs into a single HDRI [31]. The HDRI images could then be evaluated using the *Evalglare* software. The glare search algorithm adopted by *Evalglare* used a task definition criterion whereby a visual fixation area covering the screen was outlined within the image (Fig. 2). This corresponded to the test participants' point of visual focus during the experiment [7].

Based on a chosen sensitivity parameter, any pixel with a luminance value that was five times greater than the average luminance of the fixation area was treated as a glare source [7]. This implied that the luminance of the window did not necessarily correspond to the only glare source in the evaluations made by *Evalglare* under different DMX settings (i.e., other glare sources could also appear in different areas of the visual scene, for example the ceiling, walls, table surfaces, etc., as shown in Fig. 2).

Table 2 presents, at each DMX interval, the illuminance entering the lens of the CCD camera, the illuminance and CCT reaching the sensor of the chromameter, the average luminance of the glare sources, and the DGI values calculated from the HDRIs evaluated by *Evalglare*. For each DMX interval, the illuminances received at the CCD camera lens and at the sensor of the chromameter showed only minor differences ($\leq$2%). Therefore, we concluded that the measurements collected from the CCD camera were reliable for further analysis.

As shown in the measurements presented in Table 2, if no background illumination was provided by the suspended ceiling luminaries, the CCT of the visual scene remained relatively constant when varying the luminance of the artificial window. Therefore, to avoid any potential confounding influence of colour temperature on glare evaluations, these lights were turned off during the experiment.

Since the DGI index was originally derived from experiments that had also varied the luminance of a large artificial window [5], we used this glare model to verify the presence of an order bias effect on the subjective glare evaluations. In addition, considering that adjustments were performed on a large area glare source, the background luminance of the visual scene (i.e., the walls, ceiling, table, floor, etc.) was no longer independent from the luminance of the artificial window. Since this had an impact on the adaptation level, the DGI – designed specifically to evaluate discomfort due to glare from large area sources [5] – was considered to be the most suitable index to analyse the order bias effect. Although the DGI was calculated from the CCD camera images evaluated using *Evalglare*, its values were based on Equation (1) [5]:

$$\mathrm{DGI} = 10 \log \sum_{i=1}^{n} 0.478 \frac{L_s^{1.6} \cdot \Omega^{0.8}}{L_b + (0.07 \cdot \omega^{0.5} \cdot L_s)} \quad (1)$$

whereby, L_s = source luminance (cd/m^2), Ω = solid angle of the source modified by its position index (sr), L_b = background luminance (cd/m^2), and ω = solid angle of the source (sr).

Table 2
Photometric conditions measured for each DMX setting.

DMX Setting	Camera Illuminance (lux)	Chromameter Illuminance (lux)	Chromameter CCT (K)	Average Luminance of Glare Sources (cd/m^2)	DGI
10	412	415	5001	535	12
40	1628	1594	4995	2581	16
70	2562	2597	4996	3402	21
100	3516	3558	4996	4159	23
130	4526	4517	5006	5842	24
160	5497	5429	5003	6871	25
190	6371	6368	5008	7524	27
220	7158	7147	5030	8645	29
250	8102	8153	5018	10 082	29

2.3. BEARS glare scale

Although the DGI features four predefined thresholds of glare sensation – namely: "just perceptible", "just acceptable", "just uncomfortable", and "just intolerable" [5] – in our experiment, participants were asked to make judgments of visual discomfort using a newly developed scale: the "BEARS scale of subjective glare evaluation" (Fig. 3).

In fact, when reviewing existing glare scales in the literature (i.e. [6,7,32]), we found that response labels used to define the perceived magnitude of discomfort glare contained terminology that could be misinterpreted by participants or could even bias the aim of our experiment. As an example, on the original multiple-criterion scale on which the DGI is based [5], the second and third response labels refer to the borderline criteria of "just acceptable" and "just uncomfortable", respectively. When reporting these criteria in a luminance adjustment procedure using an ascending order of discomfort, this implies that a glare source judged as "acceptable" would be increased until reaching the threshold of "just uncomfortable", before becoming "uncomfortable". However, when reversing the order of presentation of the glare stimulus (i.e., when using a descending luminance adjustment sequence), it would not be logical to transition from an "uncomfortable" to an "acceptable" glare sensation by a threshold that is labelled "just uncomfortable". In this case, in fact, a decreasing "uncomfortable" glare sensation would start to become "acceptable" at the threshold, hence it might be expected that some participants would rather use the "just acceptable" criterion at the borderline between these two levels of glare sensation. Another problem is related to overlapping semantic terminology. For example, it is reasonable to assume that something that is uncomfortable may be, at the same time, acceptable. Unclear response labels found on glare scales may be another reason why large inconsistencies are often found due to uncertainty over the meaning of magnitude descriptors [33]. We believe that a detailed analysis on the use of glare scales is urgently needed. However, for this study, we developed a new scale featuring alphabetical thresholds (borderline criteria) and numerical classifications ("regions" of glare sensations with time-based descriptors) instead of response labels for participants to report their perceived levels of discomfort due to glare.

The new scale was based mainly on Petherbridge and Hopkinson's multiple criterion scale [6]. Since Hopkinson [34] proposed that observers should only be requested to describe their degree of visual discomfort with a limited number of glare criteria, we followed the same 4-point level of measurement structure used in most glare scales, e.g. Refs. [6,7,35]. The 4-points on the BEARS scale correspond to borderline alphabetical criteria (A, B, C, and D), which participants were instructed to refer to when making glare settings in a luminance adjustment procedure – similar to the study by Hopkinson and Bradley [5]. Test participants were asked to imagine these criteria as the change-over points between glare "regions" (e.g., criterion A refers to

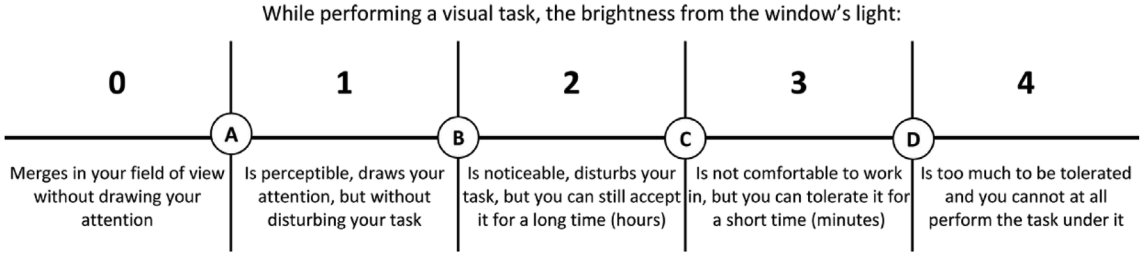

Fig. 3. BEARS scale of subjective glare evaluation.

the point when the discomfort transitions from 0 to 1 on the glare scale). Time-based descriptors were used to define five numerical "regions" of discomfort (0, 1, 2, 3 and 4) on the scale. These descriptors were largely based on previous studies of glare [29,36]. Throughout the experiment, a paper-based version of the BEARS scale was placed on the desk and an electronic version was always displayed at the top right corner of the computer screen. This allowed participants to refer to the descriptors at any point needed.

2.4. Experimental procedure

Since the study primarily sought to replicate the effect of order bias detected in our previous study, the same experimental procedure found in Kent et al. [28] was utilised. The influence of order bias on subjective glare evaluations was analysed using a luminance adjustment procedure with three different sequences of borderline discomfort criteria:

- Ascending: A, B, C, and D
- Descending: D, C, B, and A
- Randomised: e.g., A, D, C, and B

At the beginning of the experiment, participants were asked to sit on the chair so that their head was at the correct viewing position. The experimenters provided a set of instructions, including a definition of discomfort glare, the meaning of the borderline criteria and of the numerical classifications on the BEARS glare scale, and a description of how the test would run. To help reinforce the participants' understanding of the glare scale, they were asked to perform a trial of the experimental procedure without data being recorded.

At the start of the experiment, the luminance of the window was adjusted to an initial setting (anchor) corresponding to a DGI of 16, this describing the glare source on the Hopkinson scale as "just perceptible" [37]. Since previous research established that the initial setting influences the final glare setting given in a luminance adjustment task [15], only one anchor was used at the start of each block of trials. The anchor was used only when providing the first alphabetical glare criterion. Once the source was adjusted to the next alphabetical glare criterion, the luminance set by the participant became the new anchor.

During each block of trials, the experimenters adjusted the luminance of the window at a controlled pace according to the instructions given by the participant. Participants were asked whether they would like the luminance of the window to be increased, decreased, or kept at its current brightness to reach each of the four alphabetical criteria of discomfort. When participants vocally indicated that the glare source had reached the requested discomfort sensation, the corresponding DMX value was recorded. DGI values for each DMX setting were obtained from a polynomial fit calibration line based on measurements shown in Table 2. The test procedure was repeated until the participant had provided all four criteria of glare sensation under each of the three different sequences (ascending, descending, and randomised). To mask unwanted procedural effects, the sequences were presented to participants under a randomised order [38].

While making glare evaluations, participants were instructed to perform a visual task as adjustments were made to each borderline criterion of visual discomfort [30]. For this experiment, an alpha-numerical pseudo-text task was presented on the computer screen and had to be manually typed by the participant into a space provided on the display. The use of randomised pseudo-text characters was preferred to other visual tasks in order to minimise the risk of learning or experience, which could have occurred if normal text (i.e., newspaper articles) had been used [30]. Coherent with our previous work [39], the text was set to an Arial font, 12 points, double line spacing, and each character was separated using triple spacing.

A total of 55 participants (24 female and 31 male) took part to the experiment. Subjects varied in nationality and cultural background but were all fluent in English, 53 were right-handed and two left-handed, the mean age was 31.6 (SD = 9.7), 29 wore glasses, all were self-certified as having no other eye problems. Test subjects were paid for their participation to the study. The UC Berkeley Committee for Protection of Human Subjects approved the research protocol (CPHS #2017-03-9758), and all subjects signed an informed consent form before the experiment.

2.5. Statistical analyses

Graphical (Quantile plots) and statistical (Shapiro-Wilk) tests revealed that the data distributions were normal about the mean parameter [38]. A Bayesian repeated-measures Analysis of Variance (BRM-ANOVA) was ran to compare against each other the mean DGI values for each borderline criterion of glare sensation across the three different order sequences. These tests used the data from both the original study [28] and from the new experiment to determine whether the same order bias effect could be detected when both datasets were considered in the same analysis.

To determine if the same order bias effect was present in the data from the new experiment, we compared how much more likely the data fell under the *alternative* hypothesis (close replication of an order bias) than the *null* hypothesis (no replication of an order bias). Since these hypotheses are typically denoted by H_1 (*alternative*) and H_0 (*null*), the outcome of the Bayesian analysis was interpreted by the Bayes Factor (BF_{10}), whereby "10" indicates the ratio of probabilities between the *alternative* and *null* hypotheses calculated according to Equation (2) [40]:

$$BF_{10} = \frac{p(data|alternative)}{p(data|null)} \tag{2}$$

whereby, $p\ (data|alternative)$ is the probability that the data is likely to fall under the *alternative* hypothesis containing effect sizes measuring the order bias from the original study, and $p(data|null)$ is the probability that the data is likely to fall under the *null* hypothesis with an effect size equal to zero [13]. Therefore, any data supporting the *null* hypothesis would show no evidence of an order bias effect.

When the data in a new experiment shows that a similar effect size to an original study is detected, the outcome of the analysis will favour the alternative hypothesis, i.e., a close replication [41]. Therefore, a $BF_{10} > 1$ shows that the data support a close replication, while values of $BF_{10} < 1$ show that the data do not support a replication of the effect previously detected [42]. The replication Bayes Factor method

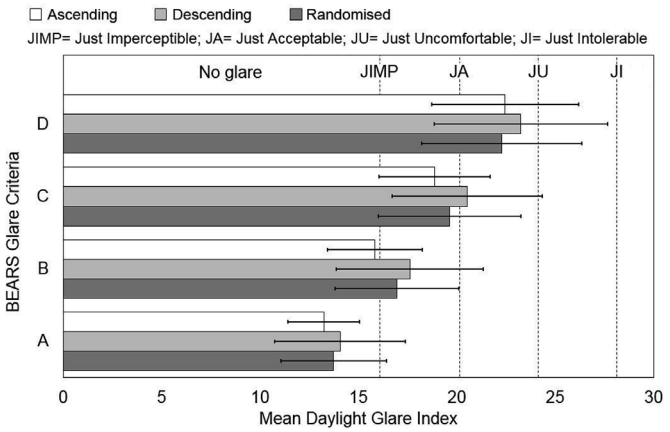

Fig. 4. Mean DGI values for the three sequences and the four glare criteria on the BEARS scale. The error bars show the standard deviations.

was applied to the BRM-ANOVA model by utilising Equation (3), which – when rearranged – gives Equation (4) [41]:

$$BF_{10}(data_{\text{original}},\ data_{\text{new}}) = BF_{10}(data_{\text{original}}) \cdot BF_{10}(data_{\text{new}}|data_{\text{original}}) \tag{3}$$

$$BF_{10}(data_{\text{new}}|data_{\text{original}}) = \frac{BF_{10}(data_{\text{original}},\ data_{\text{new}})}{BF_{10}(data_{\text{original}})} \tag{4}$$

whereby, $data_{\text{original}}$ denotes the original data found in Kent et al. [28] and $data_{\text{new}}$ is the new data generated from the experiment herein described.

Since the BRM-ANOVA only provided information that a close replication of the order bias effect was detected in at least one of the comparisons made between the three order sequences, *post-hoc* testing was performed using Bayesian paired-samples *t*-tests to determine if a close replication was present in all, or only in some, of the comparisons. To derive the prior probability distributions informing the *alternative* hypothesis, an initial Bayesian *t*-test with default priors (i.e., distributions containing no previous knowledge of the order bias) was ran using the data obtained from Kent et al. [28]. To provide a measure of the order bias, the distributions were based on the effect size (i.e., Cohen's δ = population mean difference/population standard deviation) [43]. While the prior is a distribution that contains knowledge from an original study, when this information is combined with the data from a new experiment this gives a posterior distribution, i.e., an updated estimate of the effect of interest [23,44]. To specify the prior distributions in the new analysis, the mean average effect size ($\tilde{\delta}$) and the standard deviation of the posterior distributions derived from our original study data [28] were used. This created prior distributions that closely resembled the shape of the posterior distributions found in the previous analysis [45].

Since the prior distribution under the *alternative* hypothesis in a Bayesian paired-samples *t*-test provides a relative range of effect size values, an integral is used to estimate a weighted average value for the effect size across its distribution [13]. Therefore, the Bayes Factor shows the extent to which the observed data are more likely to have occurred under the *alternative* hypothesis rather than under the *null* (5):

$$BF_{10} = \frac{\int p(data|\delta,\ \ alternative)p(\delta|alternative)\mathrm{d}\delta}{p(data|null)} \tag{5}$$

whereby, δ = Cohen's effect size.

To account for the uncertainty about the effect size under the *alternative* hypothesis, typically a prior distribution is created instead of relying on a single absolute estimate. This contains the relative range of plausible effect sizes and usually reflects the anticipated magnitude of the effect of interest [13]. To remind again, the prior distribution under the *null* hypothesis simply states that the effect size is equal to zero. It is important to note that the Bayes Factor only provides a probabilistic evaluation of the relative support between two competing hypotheses and no real indication of the actual effect size [46].

Therefore, we also used the effect size found in the posterior distribution combining both the original and the new datasets to interpret the order bias effect. The effect sizes appeared as a posterior distribution under the *alternative* hypothesis with values seen as regions along a probability density curve calculated according to Equation (6) (Gronau et al., 2017):

$$p(\delta|data) = \frac{\int_0^\infty p(data|\delta,\ \sigma^2)p(\sigma^2)\delta\sigma^2 p(\delta)}{p(data)} \tag{6}$$

whereby, $p(\delta|data)$ = marginal probability of the posterior distribution (i.e., the probability of obtaining the parameter (δ) based on the observed data [47]) and σ = the variance.

The mean values for each posterior distribution were used to interpret the differences across the two groups. Since the prior distributions from the original data were derived from a probability distribution curve with an effect size centred on a Cohen's $\delta = 0$ (in fact, no available data could be used to inform the prior before the original study), this could have resulted in posterior distributions containing effect sizes underestimating the true magnitude of the differences. Therefore, the less conservative benchmarks proposed by Cohen for effect sizes denoted as small, moderate, and large($\delta \geq 0.20$, 0.50 and 0.80, respectively) were used to interpret the outcome [48].

3. Results

In Fig. 4, we compare the mean DGI values for each criterion of glare sensation given on the BEARS scale (A, B, C and D) under the three different order sequences (ascending, descending, and randomised) for the data collected in the new experiment. The x-axis presents the mean DGI values calculated from the captured HDRI images. At specific intervals along the x-axis, the interpretation of the DGI is given based on the criteria used in the original Hopkinson's multiple-criterion scale [37], whereby benchmarks are provided for Just Imperceptible, Just Acceptable, Just Uncomfortable, and Just Intolerable (DGI ≥ 16, 20, 24, 28, respectively). The y-axis shows the alphabetical discomfort glare sensations reported by test participants distributed according to the order sequences. The interpretation of the glare settings using the window luminance as the main outcome variable can be found in Appendix B.

Fig. 4 suggests that the experimental order sequence influenced the mean DGI values. For the same criterion of glare sensation, a tendency for higher mean DGI values can be observed when adjustments were performed in a descending sequence. In fact, upon completing the block trials, one participant wrote: "*When the brightness of the window is increasing* [ascending sequence]*, the perception of brightness is different compared to when the brightness is being decreased* [descending sequence]. *The tolerance is higher when it is coming down from a brighter light than increasing it from a darker one*". This would imply that, when using the descending order sequence, if the first glare setting had been overestimated (i.e., when adjusting the window luminance to criterion D), the following settings would be biased towards this point (anchored), thereby influencing subsequent evaluations made to the other criteria on the scale. Interestingly, for glare settings made under the randomised sequence, the mean DGI values generally fell between the settings given under the ascending and descending orders. These trends support previous findings from our original study [28]. When comparing the mean DGI values across the three order sequences, the largest variations produced a difference of approximately two DGI units. Although these differences, at first glance, may wrongly give the impression that the order bias had an irrelevant impact on the resultant glare evaluations made by test participants, these differences were large

Table 3
Results of the BRM-ANOVA with informed priors.

BEARS scale criteria	BF_{10} ($Data_{original}$)	BF_{10} ($Data_{original}$, $Data_{new}$)	BF_{10} ($Data_{new}$ | $Data_{original}$)
A	31	63	2
B	2	1083	547
C	0.67	998	1490
D	0.24	5	21

Bayes Factor (BF_{10}): $0.10 \leq BF_{10} < 0.33$ is Anecdotal evidence for H_0; $0.33 \leq BF_{10} < 1$ is No evidence; $1 \leq BF_{10} < 3$ is Anecdotal evidence for H_1; $3 \leq BF_{10} < 10$ is Moderate evidence for H_1; $10 \leq BF_{10} < 30$ is Strong evidence for H_1; $30 \leq BF_{10} < 100$ is Very strong evidence for H_1; $BF_{10} \geq 100$ is Extreme evidence for H_1.

enough to change the interpretation of the outcome using Hopkinson's glare criteria for two out of the four glare criteria (i.e., B and C).

Table 3 presents the results of the BRM-ANOVA. This shows, for each borderline criterion on the BEARS scale, the Bayes Factors corresponding to the original data ($Data_{original}$), the Bayes Factors for the combined data ($Data_{original}$, $Data_{new}$), and the Bayes Factors with informed prior distributions ($Data_{new} \mid Data_{original}$).

The results of the analysis reported in Table 3 show that the evidence supporting a close replication of the order bias (i.e., the *alternative* hypothesis) is extreme ($BF_{10} \geq 100$) for the glare criteria B and C, strong ($10 \leq BF_{10} < 30$) for the glare criterion D, and anecdotal ($1 \leq BF_{10} < 3$) for the glare criterion A. Since a high Bayes Factor gives supportive evidence that the new data detected the same order bias found in the original study, we can conclude that we obtained a close replication and that the order bias was present both when discomfort due to glare was evaluated using a small (original study) and a large (new experiment) luminous source.

A *post-hoc* analysis was then performed using Bayesian paired-samples *t*-tests, whereby all combinations between the sequences for each level of glare sensation were compared against each other to detect the variations found in the BRM-ANOVA. Directionality of the hypotheses was informed by inspecting the central tendencies from graphical displays of the data [49]. Since there was no convincing evidence of a prevailing relationship between the sequence of glare criteria and mean DGI values, two-tailed hypotheses were applied [50].

Table 4 presents the results of the Bayesian paired-samples *t*-tests. This shows, for each criterion on the BEARS scale, the comparison (sequences) under examination, the mean and standard deviation of the effect size ($\bar{\delta}_{prior}$) values used to inform the prior distribution, the Bayes Factor (BF_{10}), the average mean effect size ($\bar{\delta}_{posterior}$) extracted from the posterior distribution, and the 95% upper (CI_U) and lower (CI_L) confidence intervals about the effect size.

The results of the analysis reported in Table 4 show that the evidence supporting a close replication of the order bias (i.e., the *alternative* hypothesis) is extreme ($BF_{10} \geq 100$) in two cases, strong ($10 \leq BF_{10} < 30$) in two cases, moderate ($3 \leq BF_{10} < 10$) in two cases, anecdotal ($1 \leq BF_{10} < 3$) in two cases, while no evidence ($0.33 \leq BF_{10} < 1$) was detected in three cases. Evidence supporting no replication of the order bias (i.e., the *null* hypothesis) is anecdotal in one case ($0.10 \leq BF_{10} < 0.33$). In this context, a higher Bayes Factor shows that the posterior distribution effect size closely resembles the prior distribution effect size. The largest Bayes Factors appeared when comparisons were made between the ascending and descending order sequences (e.g., for criteria B and C). This shows a high degree of evidence supporting a close replication of the order bias effect in the new experiment. The lowest discomfort glare criterion (A) showed smaller Bayes Factors for each of the three comparisons with respect to the other borderline glare criteria. While this was anticipated considering that the initial analysis (Table 3) concluded there was only anecdotal evidence to support a close replication of the order bias effect for this criterion, we suspect that this might be due to differences in the response labels and semantic descriptors found on the glare scales across the two studies. In fact, since the original study [28] used the lowest response label of “just imperceptible” found on Hopkinson's multiple-criterion scale [6], the interpretations made by observers may have differed.

The average effect sizes of the posterior distributions generally have a substantive influence ranging from moderate (Cohen's $\bar{\delta}$ absolute value: $0.50 \leq \bar{\delta}_{posterior} < 0.80$) in two cases, to small ($0.20 \leq \bar{\delta}_{posterior} < 0.50$) in eight cases, and negligible ($\bar{\delta}_{posterior} < 0.20$) in two cases. When comparing the ascending and descending order sequences, the effect size was always above the negligible threshold, and its sign was consistently negative. Supporting our previous observations, the effect sizes suggest that, under the descending order sequence, glare settings were made by participants to higher luminance values and, therefore, lower levels of discomfort due to glare were perceived when adjustments were made under this sequence.

4. Discussion

In this experiment, we obtained statistically significant and practically relevant evidence of an effect of order bias in the discomfort glare evaluations made by test participants under a large artificial window. These findings verify the conclusions made from the results of our previous work having detected the order bias effect when a small artificial glare source was used to evaluate subjective degrees of

Table 4
Results of the Bayesian *t*-tests with informed priors and effect sizes ($\bar{\delta}$).

BEARS scale criteria	Comparison	$\bar{\delta}_{prior}$	Standard Deviation	BF_{10}	$\bar{\delta}_{posterior}$	CI_U, CI_L
A	Asc. vs. Des.	−1.15	0.36	0.47	**−0.33**	−0.61, −0.06
	Asc. vs. Ran.	−0.50	0.32	1.66	**−0.30**	−0.54. −0.04
	Des. vs. Ran.	0.56	0.30	0.30	**0.20**	−0.08, 0.45
B	Asc. vs. Des.	−0.72	0.33	842	**−0.60**	−0.86, −0.33
	Asc. vs. Ran.	−0.34	0.28	28	**−0.39**	−0.64, −0.15
	Des. vs. Ran.	−0.09	0.13	0.65	0.05	−0.14, 0.23
C	Asc. vs. Des.	−0.27	0.29	1131	**−0.53**	−0.79, −0.27
	Asc. vs. Ran.	0.24	0.34	2.11	**−0.26**	−0.51, 0.00
	Des. vs. Ran.	0.59	0.29	5	**0.38**	0.13, 0.62
D	Asc. vs. Des.	−0.02	0.28	3	**−0.25**	−0.49, 0.00
	Asc. vs. Ran.	0.31	0.27	0.43	0.13	−0.10, 0.37
	Des. vs. Ran.	0.27	0.27	16	**0.35**	0.11, 0.60

Note: Asc. = Ascending, Des. = Descending, Ran. = Randomised.
Bayes Factor (BF_{10}): $0.10 \leq BF_{10} < 0.33$ is Anecdotal evidence for H_0; $0.33 \leq BF_{10} < 1$ is No evidence; $1 \leq BF_{10} < 3$ is Anecdotal evidence for H_1; $3 \leq BF_{10} < 10$ is Moderate evidence for H_1; $10 \leq BF_{10} < 30$ is Strong evidence for H_1; $30 \leq BF_{10} < 100$ is Very strong evidence for H_1; $BF_{10} \geq 100$ is Extreme evidence for H_1.
Effect Size: $\bar{\delta} < 0.20$ is negligible; $0.20 \leq \bar{\delta} < 0.50$ is small; $0.50 \leq \bar{\delta} < 0.80$ is moderate; $\bar{\delta} \geq 0.80$ is large. Values in bold denote a substantive and practically relevant effect size ($\bar{\delta} \geq 0.20$).

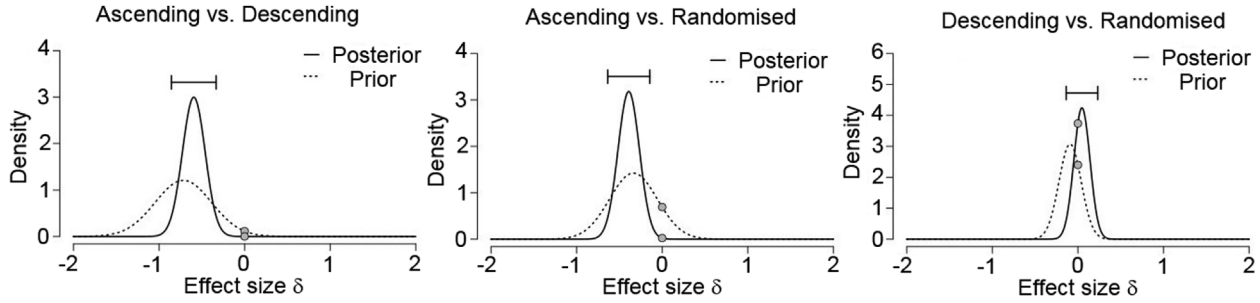

Fig. 5. Prior and posterior distributions for the effect size for glare criterion B.

discomfort due to glare.

While the main purpose of our analysis was to evaluate how much evidence favoured a close replication between the new experimental data and our original study, the posterior distributions derived from the analysis can also be used to give a more robust evaluation of the order bias. Since this outcome can be expressed as a distribution containing effect sizes that are composed from the data derived from both the original study and the replication experiment, the final parameter estimates are considered as combined values from the two sets of data.

Fig. 5 shows the effect sizes under the prior and posterior distribution curves for the glare criterion B on the BEARS scale (see Appendix C for the distribution curves relative to the other criteria on the scale). Since the three comparisons between the order sequences provided statistical evidence to support a close replication corresponding to, respectively, “extreme”, “moderate” and “no evidence” outcomes, this glare criterion was selected to illustrate the prior and posterior distributions.

For each comparison, the plots show how the prior distributions containing information from our original study [28] are updated in consideration of the new data from the replication experiment in order to create posterior distributions. Therefore, the effect sizes under the posterior distributions are based on the data from both studies.

For the “ascending vs. descending” and “ascending vs. randomised” comparisons, the mean effect sizes under the prior and posterior distributions appear to be relatively similar. This indicates equivalent differences in the glare settings made to the same criterion using different order sequences from both the original and the replication study. In fact, for both cases, the Bayes Factors (Table 4) show supportive evidence favouring the alternative hypothesis, which suggests that the same order bias was detected in both studies.

For the “descending vs. randomised” comparison, the mean average effect sizes for the prior and posterior distributions do not appear to be as closely related as seen in the previous comparisons made. In fact, the Bayes Factor did not provide as much supportive evidence towards the alternative hypothesis (Table 4). This suggests that, while replication was successful for most comparisons made across both studies, in this instance the same influence of order bias could not be detected.

It must be noted that, since the original study used a smaller sample size ($n = 20$) than the new experiment ($n = 55$), some results may have occurred simply due to chance. Therefore, when reproducing the same order bias using a larger number of observations, it should be expected that a more reliable approximation of the true effect is found [47]. In fact, comparing the effect sizes under the distributions could provide an opportunity to determine whether similar findings can be detected, hence offering some validation to the conclusions drawn in the original study [21].

The two experiments from where the data for this study were drawn, conducted respectively in Nottingham (UK) and in Singapore, were both following the same method of luminance adjustment, which is a fundamental procedure first used by Hopkinson [35] to evaluate the degrees of discomfort due to glare. Changes in experimental settings or materials are common in repeated studies, this hypothetically introducing variables that could confound the results. In our study, however, the adoption of the same luminance adjustment procedure in both experiments enabled the use of the Bayesian analysis. The main differences between the two experiments were the type of glare source (small vs large), the glare scale used (multiple-criterion scale vs BEARS), and the sample of test subjects (one group recruited at the University of Nottingham and one at BEARS in Singapore). The results of our study confirmed the same effect of order bias, signalling that the differences between the two experimental settings did not mask the ultimate finding; an order effect can bias the evaluation of discomfort due to glare in a luminance adjustment task, and is also resilient to changes in the experimental procedure.

Since little attention has been placed on the semantic descriptors found in response labels used in past studies, we proposed a new scale for evaluating subjective degrees of discomfort due to glare. While further investigation is needed to ensure that the participants' understanding of response label descriptors matches the experimenter's intended interpretation, we believe that our new (BEARS) scale avoided the use of the misleading semantics found in other glare scales. However, the development of a more robust scale that can be used to evaluate the sensation of discomfort due to glare still requires urgent consideration.

4.1. Limitations

Before any conclusions on the practical applications of these results can be made, some methodological limitations need to be acknowledged. Similar to the original study's experimental procedure, it should be considered that, within each block trial, the tests began the adjustments with the glare source set to a low luminance anchor. Since glare evaluations have been found to be influenced by the initial anchor [15], we believe that participants could have given different settings if a higher luminance anchor had been used (i.e., the artificial window would have been set to higher luminance values for the same degree of glare sensation).

The literature shows that many different glare scales have been used and, while these are closely related to the multiple-criterion scale [6], their response labels used to describe glare sensation have varied considerably [28]. There are some studies that have applied response labels to prediction models that had been originally derived using a different set of glare criteria [51]. Based on such inconsistencies, we proposed a new glare scale. While we have avoided the use of semantics to directly describe a given threshold of discomfort, variations in the response labels from our original study may have influenced the participants’ understanding of the descriptors used on the scale [52]. Since the glare scales were not identical across both studies, this might have influenced the magnitude of evidence supporting a close replication of the order bias effect.

Finally, it should be noted that, under the luminance adjustment procedure, the calculated DGI values were consistently lower than the thresholds provided by Hopkinson (Fig. 4). It could also be postulated that, if the glare source had contained visual content – for example, a

view to an exterior environment – test participants might have given different glare settings. This is coherent with studies that have shown that observers became more tolerant to discomfort due to glare when a daylit (window) glare source contained pleasant information [17,18]. Furthermore, one participant reported that they would have provided higher glare settings if the maximum luminance of the glare source had been known prior to giving their adjustments. This might be associated to a range bias effect [53] and will be the focus of future experiments.

5. Conclusions

In a test room equipped with a large artificial window, we performed a laboratory test with 55 participants to validate the effect of order bias on subjective degrees of discomfort due to glare. This effect had been previously detected in a similar experiment using a small luminous source. The results confirmed the presence of an order bias effect, hence supporting the conclusions drawn from our previous study. Moreover, we used a Bayesian approach to statistically verify the level of replication and quantify the magnitude of its effect. The Bayesian approach gave us the following advantages: (1) we were able to identify the same order bias when evaluating discomfort due to glare with a small and a large source; (2) estimates of the order bias from our original study and from the new experiment were combined to provide a more precise evaluation of the effect of interest. The posterior estimates derived in this study can also be used to inform prior distributions in further analyses; (3) conclusions drawn from the new data supported those derived in our previous study. This ruled out the possibility that the order bias effect detected in the original study had occurred simply due to a random error.

We estimated the magnitude of the order bias to cause at least a two-unit change in the DGI, which is large enough to change the interpretation of the outcome as described by Hopkinson's glare criteria. Even if these results may not have an immediate application to building practice, we believe it is essential to consider more robust analytical approaches to support the conclusions drawn when comparing the results from different glare studies. Despite the inconsistencies that have been identified between glare models, these are still recommended by international lighting design standards [54–56]. We believe that the new research approach proposed in this paper will benefit the lighting research community and, consequently, the building standardization and design industries.

Since many replication studies report poor levels of predictive power when evaluating glare models using NHST, as a research community, we need to move beyond the conventional null hypothesis significance testing approach and encourage the wider use of Bayesian methods (or at least to publish the original data with any paper publication). Although in our study we used a Bayesian analysis to replicate the detection of an order bias effect, this approach can also serve as a method for evaluating multiple sets of data to provide robust evaluations of certain statistical parameters (i.e., correlation or regression coefficients). The use of such approaches when analysing experimental data may lead to more reliable comparisons between different glare studies. Furthermore, when multiple studies use the same subjective glare criteria, the Bayesian approach could also be applied to update parameter estimates that are used to describe the thresholds of discomfort. These thresholds can be adopted by building designers to minimise the risk of discomfort due to glare from large light sources, such as in the case of a window, hence leading to more reliable predictions when using glare models.

Acknowledgements

This work was supported by: the Engineering and Physical Sciences Research Council [grant number EP/N50970X/1]; an International Collaboration Fund awarded by the University of Nottingham; and, the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a centre for intellectual excellence in research and education in Singapore.

Acknowledgement is given to Professor Steve Fotios (University of Sheffield) for reviewing the experimental procedure and the proposed BEARS glare scale, and to Professor E.M. Wagenmakers (University of Amsterdam) for support on the statistical analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.buildenv.2018.10.005.

References

[1] M.B. Hirning, G.L. Isoardi, V.R. Garcia-Hansen, Prediction of discomfort glare from windows under tropical skies, Build. Environ. 113 (2017) 107–120, https://doi.org/10.1016/j.buildenv.2016.08.005.
[2] G. Baird, J. Thompson, Lighting conditions in sustainable buildings: results of a survey of users' perceptions, Architect. Sci. Rev. 55 (2012) 102–109, https://doi.org/10.1080/00038628.2012.667941.
[3] I. Konstantzos, A. Tzempelikos, Daylight glare evaluation with the sun in the field of view through window shades, Build. Environ. 113 (2017) 65–77, https://doi.org/10.1016/j.buildenv.2016.09.009.
[4] P. Tregenza, M. Wilson, Daylighting: Architecture and Lighting Design, Routledge, Abingdon, 2011.
[5] R.G. Hopkinson, R.C. Bradley, A study of glare from very large sources, Illum. Eng. 55 (1960) 288–294.
[6] P. Petherbridge, R.G. Hopkinson, Discomfort glare and the lighting of buildings, Trans. Illum. Eng. Soc. 15 (1950) 39–79, https://doi.org/10.1177/147715355001500201.
[7] J. Wienold, J. Christoffersen, Evaluation methods and development of a new glare prediction model for daylight environments with the use of CCD cameras, Energy Build. 38 (2006) 743–757, https://doi.org/10.1016/j.enbuild.2006.03.017.
[8] R.P. Carver, The case against statistical significance testing, revisited, J. Exp. Educ. 61 (1993) 287–292.
[9] A. Gelman, H. Stern, The difference between “Significant” and “Not Significant” is not itself statistically significant, Am. Statistician 60 (2006) 328–331, https://doi.org/10.1198/000313006X152649.
[10] S. Altomonte, S. Schiavon, Occupant satisfaction in LEED and non-LEED certified buildings, Build. Environ. 68 (2013) 66–76, https://doi.org/10.1016/j.buildenv.2013.06.008.
[11] G. Cumming, Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better, Perspect. Psychol. Sci. 3 (2008) 286–300, https://doi.org/10.1111/j.1745-6924.2008.00079.x.
[12] S. Nieuwenhuis, B.U. Forstmann, E.-J. Wagenmakers, Erroneous analyses of interactions in neuroscience: a problem of significance, Nat. Neurosci. 14 (2011) 1105–1107, https://doi.org/10.1038/nn.2886.
[13] J. Verhagen, E.-J. Wagenmakers, Bayesian tests to quantify the result of a replication attempt, J. Exp. Psychol. Gen. 143 (2014) 1457–1475, https://doi.org/10.1037/a0036731.
[14] K. Fisekis, M. Davies, M. Kolokotroni, P. Langford, Prediction of discomfort glare from windows, Light. Res. Technol. 35 (2003) 360–369, https://doi.org/10.1191/1365782803li095oa.
[15] M.G. Kent, S. Fotios, S. Altomonte, Discomfort glare evaluation: the influence of anchor bias in luminance adjustments, Lighting Research & Technology, 2017, https://doi.org/10.1177/1477153517734280 1477153517734280.
[16] R. Kittler, M. Kocifak, S. Darula, Daylight Science and Daylight Technology, Springer, New York, 2012.
[17] N. Tuaycharoen, P.R. Tregenza, Discomfort glare from interesting images, Light. Res. Technol. 37 (2005) 329–341.
[18] N. Tuaycharoen, P.R. Tregenza, View and discomfort glare from windows, Light. Res. Technol. 39 (2007) 185–200, https://doi.org/10.1177/1365782807077193.
[19] T. Iwata, M. Shukuya, N. Somekawa, K. Kimura, Experimental study on discomfort glare caused by windows: subjective response to glare from a simulated window, J. Archit. Plann. (Trans. AIJ) 432 (1992) 21–33.
[20] T. Iwata, M. Tokura, M. Shukuya, K. Kimura, Experimental study on discomfort glare caused by window part 2: subjective response to glare from actual windows, J. Archit., Plann. Environ. Eng. (Trans. AIJ) 439 (1992) 19–31, https://doi.org/10.3130/aijax.439.0_19.
[21] M.J. Bayarri, A.M. Mayoral, Bayesian analysis and design for comparison of effect-sizes, J. Stat. Plann. Inference 103 (2002) 225–243, https://doi.org/10.1016/S0378-3758(01)00223-3.
[22] T. Bayes, An essay towards solving a problem in the doctrine of chances, Phil. Trans. 53 (1763) 370–418, https://doi.org/10.1098/rstl.1763.0053.
[23] W.M. Bolstad, Introduction to Bayesian Statistics, John Wiley, 2016.
[24] J. Kruschke, T. Liddell, The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-analysis, and Power Analysis from a Bayesian Perspective, Social

Science Research Network, Rochester, NY, 2016https://papers.ssrn.com/abstract=2606016 , Accessed date: 21 May 2018.
[25] P.D. O'Neill, A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods, Math. Biosci. 180 (2002) 103–114, https://doi.org/10.1016/S0025-5564(02)00109-8.
[26] A. Yuille, D. Kersten, Vision as Bayesian inference: analysis by synthesis? Trends Cognit. Sci. 10 (2006) 301–308, https://doi.org/10.1016/j.tics.2006.05.002.
[27] J. Xiong, A. Tzempelikos, I. Bilionis, N.M. Awalgaonkar, S. Lee, I. Konstantzos, S.A. Sadeghi, P. Karava, Inferring Personalized Visual Satisfaction Profiles in Daylit Offices from Comparative Preferences Using a Bayesian Approach, Building and Environment, 2018, https://doi.org/10.1016/j.buildenv.2018.04.022.
[28] M.G. Kent, S. Fotios, S. Altomonte, Order Effects when Using Hopkinson's Multiple Criterion Scale of Discomfort Due to Glare, Building and Environment, 2018.
[29] W.K.E. Osterhaus, I.L. Bailey, Large area glare sources and their effect on visual discomfort and visual performance at computer workstations, Conference Record of the 1992 IEEE Industry Applications Society Annual Meeting, vol. 2, 1992, pp. 1825–1829, , https://doi.org/10.1109/IAS.1992.244537.
[30] M.G. Kent, S. Altomonte, R. Wilson, P.R. Tregenza, Temporal effects on glare response from daylight, Build. Environ. 113 (2017) 49–64, https://doi.org/10.1016/j.buildenv.2016.09.002.
[31] H. Cai, T.M. Chung, Improving the quality of high dynamic range images, Light. Res. Technol. 43 (2011) 87–102, https://doi.org/10.1177/1477153510371356.
[32] J.B. de Boer, D.A. Schreuder, Glare as a criterion for quality in street lighting, Trans. Illum. Eng. Soc. 32 (1967) 117–135, https://doi.org/10.1177/147715356703200205.
[33] S. Fotios, Research Note: uncertainty in subjective evaluation of discomfort glare, Light. Res. Technol. 47 (2015) 379–383, https://doi.org/10.1177/1477153515574985.
[34] R.G. Hopkinson, The multiple criterion technique of subjective appraisal, Q. J. Exp. Psychol. 2 (1950) 124–131, https://doi.org/10.1080/17470215008416585.
[35] R.G. Hopkinson, Discomfort glare in lighted streets, Trans. Illum. Eng. Soc. 5 (1940) 1–32, https://doi.org/10.1177/147715354000500101.
[36] M. Velds, User acceptance studies to evaluate discomfort glare in daylit rooms, Sol. Energy 73 (2002) 95–103, https://doi.org/10.1016/S0038-092X(02)00037-3.
[37] R.G. Hopkinson, Glare from windows, Construc. Res. Dev. J. 2 (1970) 98–106.
[38] A. Field, Discovering Statistics Using IBM SPSS Statistics, fourth ed., Sage, London, 2013.
[39] S. Altomonte, M.G. Kent, P.R. Tregenza, R. Wilson, Visual task difficulty and temporal influences in glare response, Build. Environ. 95 (2016) 209–226, https://doi.org/10.1016/j.buildenv.2015.09.021.
[40] R.D. Morey, J.N. Rouder, Bayes Factor approaches for testing interval null hypotheses, Psychol. Methods 16 (2011) 406–419, https://doi.org/10.1037/a0024377.
[41] A. Ly, A. Raj, A. Etz, M. Marsman, Q.F. Gronau, E.-J. Wagenmakers, Bayesian Reanalyses from Summary Statistics: a Guide for Academic Consumers, Open Science Framework, (2017), https://doi.org/10.17605/OSF.IO/7DZMK.
[42] C. Harms, A Bayes Factor for Replications of ANOVA Results, ArXiv:1611.09341 [Stat] (2016) http://arxiv.org/abs/1611.09341 , Accessed date: 4 February 2018.
[43] H. Jeffreys, Theory of Probability, third ed., Oxford University Press, Oxford, 1961.
[44] R.E. Kass, A.E. Raftery, Bayes Factors, J. Am. Stat. Assoc. 90 (1995) 773–795, https://doi.org/10.1080/01621459.1995.10476572.
[45] Q.F. Gronau, A. Ly, E.-J. Wagenmakers, Informed Bayesian T-tests, ArXiv:1704.02479 [Stat] (2017) http://arxiv.org/abs/1704.02479 , Accessed date: 4 February 2018.
[46] H.S. Stern, A test by any other name: P values, Bayes Factors, and statistical inference, Multivariate Behav. Res. 51 (2016) 23–29, https://doi.org/10.1080/00273171.2015.1099032.
[47] A. Field, An Adventure in Statistics: the Reality Enigma, SAGE Publications, 2016.
[48] J. Cohen, A power primer, Psychol. Bull. 112 (1992) 155–159.
[49] D. Hauschke, V.W. Steinijans, Directional decision for a two-tailed alternative, J. Biopharm. Stat. 6 (1996) 211–218, https://doi.org/10.1080/10543409608835134.
[50] G.D. Ruxton, M. Neuhäuser, When should we use one-tailed hypothesis testing? Method/ Ecol/ Evol/ 1 (2010) 114–117, https://doi.org/10.1111/j.2041-210X.2010.00014.x.
[51] J.Y. Suk, M. Schiler, K. Kensek, Investigation of existing discomfort glare indices using human subject study data, Build. Environ. 113 (2017) 121–130, https://doi.org/10.1016/j.buildenv.2016.09.018.
[52] A.W. Gellatly, D.J. Weintraub, User Reconfiguration of the De Boer Rating Scale for Discomfort Glare, The University of Michigan, Industry Affiliation Program for Human Factors in Transportation Safety, 1990.
[53] E.C. Poulton, Quantitative subjective assessments are almost always biased, sometimes completely misleading, Br. J. Psychol. 68 (1977) 409–425, https://doi.org/10.1111/j.2044-8295.1977.tb01607.x.
[54] CIE, Discomfort Glare in Interior Lighting, Commission Internationale de l'Éclairage, 1995.
[55] SLL, The SLL Code for Lighting, Chartered Institution of Building Services Engineers (CIBSE), 2012.
[56] IESNA, I.E.S. Of N. America, The Lighting Handbook: Reference and Application, Illuminating Engineering Society of North America, 2011.