Measuring Discomfort from Glare: Recommendations for Good Practice

S. Fotios[a], M. Kent[b]

[a] School of Architecture, University of Sheffield, Sheffield, UK
Steve.fotios@sheffield.ac.uk

[b] Berkeley Education Alliance for Research in Singapore, Singapore
michaelkent@berkeley.edu

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

S. Fotios & M. Kent (2020) Measuring Discomfort from Glare: Recommendations for Good Practice, *LEUKOS*, DOI: 10.1080/15502724.2020.1803082

Abstract

This article presents a review of the methods used for subjective evaluation of discomfort from glare, focusing on the two procedures most frequently used in past research – adjustment and category rating. Evidence is presented to demonstrate that some aspects of these procedures influence the evaluation, such as the range of glare source luminances available in an adjustment procedure, leading to biased evaluations and which hence reduce the reliability and validity of the data. The article offers recommendations for good practice when using these procedures and also suggests alternative methods that might be explored in further work.

Key Words: discomfort glare, lighting, experimental methods

1. Introduction

The human visual system can adapt to, and work well in, a wide range of ambient light levels, from strong sunlight to moonlight. The visual system does this through a range of mechanisms including, for example, an increase in the diameter of the pupil at lower light levels. At any given moment, however, the visual system can adapt to only a limited range of luminances. If that range is too great, the eye cannot adapt, and regions of the scene that are excessively bright can lead to discomfort. Common examples of such situations leading to discomfort include the headlights of oncoming vehicles when driving after dark and direct sunlight through windows in daytime.

This article concerns discomfort from glare. Glare is defined by the International Commission on Illumination (CIE) as a "*condition of vision in which there is discomfort or a reduction in the ability to see details or objects, caused by an unsuitable distribution or range of luminance, or by extreme contrasts*" [CIE 2019a]. Discomfort glare is further defined by the CIE as "*glare that causes discomfort without necessarily impairing the vision of objects*" [CIE 2019b]. *Comfort* is a state of physical ease and freedom from pain or constraint [Oxford Dictionaries 2019], a pleasant feeling of being relaxed and free from pain [Cambridge Dictionary 2019a]: *Discomfort* is a feeling of being uncomfortable, physically or mentally [Cambridge Dictionary 2019b]. Hence the term 'discomfort' is used to distinguish between a subjective sensation (discomfort from glare) and an impairment to visual performance (disability from glare) – in other words, between the psychological glare (discomfort) and physiological glare (disability) [Osterhaus 2005]. A given visual scene may induce one, both, or neither of these outcomes.

Discomfort from glare is not well understood. Despite the existence of over 80 experimental studies of discomfort from glare in various contexts [Pierson et al 2018] there is still no agreed model for predicting the likely presence and/or severity of discomfort. One reason why there is no agreed model is that there is a large variance in findings, both between subjects and between studies. As demonstrated in previous reviews of preferred light levels [Fotios and Cheal 2010], spatial brightness [Fotios et al 2015] and correlated color temperature (CCT) preference [Fotios 2017], considering differences in experimental design can explain some of the variance.

This article presents a review of the methods used for subjective evaluation of discomfort from glare, showing those aspects of experimental design which can lead to biased evaluations and hence reduce the reliability and validity of the data. Reliability is the extent to which results are consistent over time, accurately represent the intended population, and can be reproduced under a similar methodology [Golafshani 2003]. Validity describes how well an experiment is actually measuring what it was intended to measure, and hence how truthful the results are [Golafshani 2003]. The discussion focusses upon subjective, quantitative evaluations of discomfort magnitude because that is what the majority of past studies have used. While evaluation using physiological measurement and behavioral observation is also possible, very few studies have done so, and hence there are limited data available.

In studies of discomfort from glare, it is typical for the level of discomfort to be varied by changing the luminance of either the glare source or its background. While this article is phrased in terms of glare source luminance, the findings are applicable also to variation of the background luminance. This article does not include discussion of photometric measurements or disability from glare, and it does not provide a review of previous studies or models of discomfort from glare.

This article first describes standard psychophysical test procedures, followed by a discussion of experimental biases associated with the adjustment and category rating procedures, which are the most commonly used procedures in past studies of discomfort from glare. This leads to a series of recommendations for good practice and suggestions for alternative procedures. Following the recommendations might reduce the variance associated with discomfort evaluations, whereas

alternative procedures might lead to different evaluations. Both steps are essential if understanding of discomfort from glare is to be advanced.

2. Commonly used test procedures

There are four basic psychophysical procedures for explicit quantitative measurement: adjustment, matching, discrimination and category rating [CIE 212:2014]. They can be categorized according to the ability to modify the stimulus and the nature of the reference stimulus as shown in Figure 1. Of the four procedures, two are commonly used in studies evaluating the discomfort from glare – adjustment and category rating. While equally valid as test procedures, matching and discrimination have rarely been used to evaluate discomfort from glare. This may be because they are two-interval tasks in which the test scene is compared with a visual reference scene (e.g. as two side-by-side scenes observed simultaneously, or, as two scenes observed one after the other at the same spatial location), requiring an additional visual scene to be set up.

Interaction with the visual scene	**Absolute measurement** No external reference present	**Relative measurement** Presence of an external reference
Passive (No interaction)	*Rating*	*Discrimination*
Active (Interaction required)	*Adjustment*	*Matching*

Figure 1. Basic procedures for explicit quantitative measurement.

Note: *External reference*: a second relevant visual scene is presented whilst assessment of the test scene is made, although not necessarily simultaneously.
Interaction: within the trial, the visual scene itself can be changed by the actions of the participant. In brightness studies this interaction is limited to one dimension – variation in quantity, such as luminance or illuminance, at a calibration point.

Adjustment is a single-interval task. A single-interval task is one in which only a single visual scene is observed and judgements are made against an internal (memory) reference. In contrast, a two-interval task is one in which two visual scenes are observed: the scene being judged and a visual comparison. The luminance of the glare source is adjusted (increased and/or decreased) until the scene resembles a particular level of discomfort. Recent studies using adjustment include Tuaycharoen and Tregenza [2005], Fekete et al [2010] and Kim and Kim [2011]. In some experiments, adjustment is used to define the so-called Border between Comfort and Discomfort (BCD), the criterion introduced by Luckiesh and Guth [1949] following the earlier *boundary* between comfort and discomfort of Luckiesh and Holladay [1925]. In other work, adjustment is made to each of several degrees of discomfort, which is known as the multiple criterion method (MCM) initiated by Hopkinson [1940]. Adjustment by the participant may be made through direct control of luminance (e.g. a rotary control dial) or indirectly, with the test participant giving commands (e.g. higher or lower) to an experimenter who carries out the action. The output of a trial is the glare source luminance at the setting made. Different visual scenes (e.g. light sources of different spectral power distribution, size or location) are presented individually, in

succession, and the task is carried out in isolation of an external reference. Experimental biases in adjustment-based studies of discomfort from glare are discussed in Section 3.

Category Rating is usually a single interval task in which the participant is required to describe the degree of discomfort experienced when observing a visual scene by allocating it to one of a series of categories. For example, this may be a 7-point response scale with category labels varying from 'imperceptible' to 'intolerable' [Ngai and Boyce 2000]. There is no consensus as to the number of response points nor the labels of each category and hence these vary between studies. In discomfort from glare studies, category rating is typically used as a single-interval task in which different visual scenes are presented and evaluated individually, in succession. The output of a trial is the discomfort category, usually quantified by the integer associated with that category. Recent studies using rating include Tuaycharoen and Tregenza [2007], Rodriquez and Pattini [2014] and Tyukhova and Waters [2018]. Experimental biases in studies of discomfort from glare using category rating are discussed in Section 4.

Matching presents participants with two scenes in spatial or temporal juxtaposition. One scene is the reference and remains unchanged. Participants are instructed to vary the glare source luminance of the second (test) scene until it matches as near as possible the degree of discomfort portrayed by the reference scene. This action is usually carried out directly by the participant but may also be carried out indirectly by the experimenter following a command from the participant. The output of a trial is the ratio of the glare source luminances at equal discomfort. Matching is, however, rarely used in discomfort studies. One example is the "comparative method" of Luckiesh and Holladay [1925].

Discrimination requires the participant to report which of two scenes presents the greater degree of discomfort (also known as paired comparison). The two scenes are presented in spatial or temporal juxtaposition and the conditions of both are fixed for a given trial. Discrimination is usually (but not necessarily) a forced choice task, where the response of equal discomfort is not permitted. The output is the frequency of responses by which a particular scene is considered to be the greater discomfort. To enable subsequent estimation of the luminance ratio for equal discomfort the discrimination task is repeated with the luminance (glare source or its background) of one or both of the visual scenes varied through several steps. Discrimination, however, has rarely been used [Collins 1962; Tuaycharoen and Tregenza 2005 (experiment 2); Waters et al 1995].

Bargary et al [2014] used a staircase procedure. This is essentially a series of discrimination evaluations with a separate, rather than simultaneous (or sequential), mode of evaluation. For a given stimulus, the observer reported if discomfort was absent or present. The glare source luminance was varied in fixed steps, a sequence of increments or decrements, and the evaluation repeated at each step. For trials starting with no discomfort, the luminance gradually increased until discomfort was found, at which point the sequence reversed until comfort was achieved: the average of several such reversals was used to estimate the mean luminance for discomfort threshold.

3. Experimental bias: Adjustment

This section describes biases that can occur with adjustment tasks. For most of the described biases, it is known that they occur but not *why* they occur, which is likely a combination of psychological and physiological factors that are difficult to discern, and beyond the scope of this article.

3.1 Stimulus range bias

Stimulus range bias describes the influence on subjective evaluations of the range of stimuli available to the test participant [Poulton 1989]. Range effects have been found to affect many sensory responses when using the adjustment procedure, including preferred color [Logadóttir et al 2013], preferred

brightness [Fotios and Cheal 2010; Logadóttir et al 2011, Uttley et al 2013], and loudness [Parker and Schneider 1994; Poulton 1977].

In the context of the adjustment procedure for discomfort from glare, the range refers to the minimum and maximum luminances available via the control device. This range is chosen by the experimenter. Regardless of any alleged validation to their choice, these ranges strongly influence the response gained from test participants and therefore constitute an experimental bias.

Table 1 shows the results of an MCM adjustment procedure in which test participants were required to set glare source luminances representing four degrees of discomfort (in a random order) for three ranges of available stimulus magnitudes (as modified by variation in the upper luminance available with the control device) [Kent et al 2019a]. The results show that mean luminances for a particular degree of discomfort increased as the upper limit of the stimulus range increased. These differences were statistically significant (p<0.01).

Consider the luminance associated with discomfort degree 4, the highest degree of discomfort (Table 1). The mean luminance set with the low stimulus range (4,169 cd/m^2) is smaller than that for the middle (5,544 cd/m^2) and high (6,539 cd/m^2) luminance ranges. Furthermore, it is also smaller than the mean luminances set for a lower degree of discomfort (discomfort degree 3) in the middle and high ranges. In other words, a change in stimulus range caused a change in luminance settings similar to one whole criterion step on the discomfort scale. The interpretation made by the experimenter from such results (typically discomfort threshold X is associated with glare source luminance Y) depends on the range available for luminance settings. The choice of luminance range is rarely, if ever, discussed in previous studies other than those specifically investigating range bias.

Table 1. Mean luminances for four degrees of discomfort set using MCM adjustment with three stimulus ranges (minimum and maximum luminances that could be set using the adjustment control) [Kent et al 2019a]. These are results for all trials, with participants having direct control over the adjustment.

Luminance range	Luminance range (cd/m^2)		Mean luminance (cd/m^2)			
	Min	Max	Discomfort degree 1 (low)	Discomfort degree 2	Discomfort degree 3	Discomfort degree 4 (high)
Low range	441	5106	1417	2160	3209	4169
Middle range	441	7288	1931	2976	4408	5544
High range	441	9469	2314	3490	5036	6539

Further demonstration of range bias when using the adjustment procedure was reported by Lulla and Bennett [1981] – see Appendix 1. A comparison of the influence of different aspects of experimental design suggests the largest effect sizes are those associated with range bias and anchoring [Kent et al 2019a].

Range bias offers an alternative explanation to proposed influences on discomfort evaluations. Kim and Kim [2011] imply that the discomfort tolerance of Koreans is different to that of the test participants of Luckiesh and Guth [1949] – whom we assume to be North Americans. Both studies used an adjustment task to set the BCD with a glare source in central vision. Kim and Kim used a luminance range of 0-160,000 cd/m2 resulting in a mean luminance of 5,253 cd/m2 (see their Table 2), higher than the average luminance of 2,844 cd/m2 (830 foot lamberts – see their Table 1) found by Luckiesh and Guth when using a luminance range extending from zero to 103,000 cd/m2 (30,000 fL). (Note that Kim and Kim cite the geometric mean reported by Luckiesh and Guth and not the arithmetic mean which was

891 fL). In both studies, the average BCD luminance is approximately 3% of the available range. Rather than being an effect of ethnicity, as implied by Kim and Kim, the difference between the two studies can also be explained as an effect of range bias – i.e., the larger luminance range led to the greater estimate of BCD.

3.2 Anchor effects

When using an adjustment procedure, the action of making the adjustment must have a starting point. The luminance of the glare source at the start is known as an anchor because the setting subsequently made is weighted towards the anchor: a low anchor leads to a lower setting than when commencing from a high anchor. Anchors can affect a large range of judgements, including responses to general knowledge questions, economic evaluations, and social values [Chapman and Johnson 1999; Mussweiler and Strack 2001]. Within the field of lighting, an anchor effect has been demonstrated in studies using an adjustment procedure to investigate brightness [Fotios and Cheal 2010, Logadóttir et al 2011, Uttley et al 2013] and color [Logadóttir et al 2013] as well as discomfort from glare [Osterhaus and Bailey 1992; Kent et al 2019b].

Kent et al [2019b] repeated a luminance adjustment task with three anchors (labeled low, medium and high) used in a randomized order. The results (Table 2) revealed significant differences (p<0.001) in mean luminance settings with change in the anchor.

Table 2. Results of luminance adjustment procedure where four discomfort sensations were evaluated with three anchors [Kent et al 2019b]. For all discomfort sensations, the higher anchor resulted in a higher mean luminance.

Anchor	Source luminance (cd/m²)	Mean luminance cd/m² (and standard deviation)			
		Just Imperceptible	Just Acceptable	Just Uncomfortable	Just Intolerable
Low	1,627	1,784 (1,031)	3,043 (1,534)	4,517 (2,027)	8,238 (4,135)
Medium	5,414	3,192 (1,341)	4,350 (1,982)	5,858 (1,982)	10,130 (3,388)
High	8,999	5,663 (2,923)	7,224 (3,037)	9,031 (3,232)	13,548 (4,858)

One approach that intends to counter the influence of anchors is to set the variable stimulus to values far above and far below the expected threshold value prior to successive trials and use the mean of the two subsequent settings as the best estimate [Gescheider 1997].

3.3 Order effects

In an early luminance adjustment study conducted by Petherbridge and Hopkinson [1950], the stimulus was adjusted to four levels of discomfort in ascending order: just imperceptible, just acceptable, just uncomfortable and just intolerable. The observers were instructed to vary the luminance of the glare source to meet the lowest of the four discomfort levels, and then, incrementally, the other three. The previously set luminance would then be an anchor for the next trial. Thus, this study did not follow current standard practice, which would be for the adjustments to each level to be made in a randomized or counterbalanced order. Later work has revealed that the order likely influenced the results.

Pulpitlova and Detkova [1993] used a secondary sequence in addition to Petherbridge and Hopkinson's ascending-only order, with this secondary order being a near reversal of the original. The results (Table 3) indicated that the mean luminance settings in the secondary sequence were consistently higher than those in the ascending order for all four discomfort criteria. This is supported by Kent *et al.* [2018] who conducted an adjustment experiment similar to that of Petherbridge and Hopkinson but with three approaches to the order in which the four discomfort sensations were employed; ascending, descending and randomized. Differences between the three orders were significant for three degrees of discomfort

(just imperceptible, just acceptable and just uncomfortable) but were not suggested to be significant for just intolerable settings.

Table 3. Mean luminance settings when adjustments were made to four levels of discomfort glare in different sequential orders (data from Pulpitlova and Detkova 1993).

Level of discomfort	Luminance (cd/m^2)	
	Ascending Order (JP, JA, JU, JI)	**Secondary Sequence (JU, JI, JA, JP)**
Just Perceptible	418	1042
Just Acceptable	1330	2189
Just Uncomfortable	2836	3110
Just Intolerable	4501	5501

Note: JP= Just Perceptible, JA= Just Acceptable, JU= Just Uncomfortable, JI= Just Intolerable

3.4 Direct versus indirect control

There are two routes by which an adjustment action may be made. In some past studies, the observer was required to directly vary the luminance, such as by using a control dial e.g. [Hopkinson 1950; Luckiesh and Guth 1949; Petherbridge and Hopkinson 1950]. With this approach, the observer has direct control over the variable stimulus and is free to adjust the variable stimulus in any manner they choose until they reach the final setting. In other studies this control is indirect, with the experimenter making the adjustments according to the vocal instructions provided by the test observer e.g. [Kent et al 2015; Tuaycharoen and Tregenza 2005]. There are two reasons why this may make a difference. First is the perception of personal control. The perceived level of personal control over an environment plays a large role on occupant performance, satisfaction and behavior [Lee and Brand 2005]. Second, the observer may employ a different level of precision when giving instructions to a second person rather than having direct control. The observer may accept an otherwise imperfect setting to reduce the need to yet again say “increase” or “decrease” to change the glare source brightness, which would be an undesirable outcome [Kent et al 2019a]. Conversely, an observer may have limited motor control or unfamiliarity with the dimming controls, in which case having an experimenter that is trained to adjust the glare source brightness based on observer instruction may lead to more reliable adjustment. This is especially true if the dimming control system is highly sensitive or non-linear.

An experiment was conducted to compare settings made using direct and indirect adjustment [Kent et al 2019a]. The glare source was a large artificial window facing the participant. Settings were made to four degrees of discomfort, in a randomized order, with three stimulus ranges, in a randomized order. The anchor in these trials was the mid-point of the available range. In a repeated measures design, the 42 test participants completed this task using both direct and indirect control. Direct control was achieved using a mouse click on a screen command.

Table 4 shows the luminances set for four sensations of discomfort when using the adjustment procedure for indirect and direct control [Kent et al 2019a]. For each of the four discomfort sensations, the luminance set under the direct control was higher than those under the indirect method of control. The differences across the two conditions were significant for Just Uncomfortable (p<0.001), Just Perceptible (p<0.01) and Just Intolerable (p<0.01) but were not suggested to be significant for Just Noticeable. The differences across all four sensations of discomfort show small effect sizes: a small effect means that something is happening (i.e. it is practically meaningful) but may only be revealed with careful study. A small effect may be relevant if it is sufficient to change the conclusion drawn from an investigation. While the results show that the method of controlling the variable stimulus matters when

evaluating discomfort from glare, it is unclear which method provides the most valid results in practice, only that there is a difference.

Table 4. Luminances for four sensations of discomfort when compared across direct (participant) and indirect (experiment) control during an adjustment procedure [Kent et al 2019a]. Note: The luminances set are the average across the three stimulus ranges used.

Degree of discomfort	Control method		$\Delta Mean^{NHST}$	Effect size (*r*)
	Direct (cd/m²)	*Indirect (cd/m²)*		
Just Perceptible	1,888	1,723	165 **	0.34
Just Noticeable	2,875	2,735	140 n.s.	0.24
Just Uncomfortable	4,218	3,793	425 ***	0.49
Just Intolerable	5,417	5,102	315 **	0.36

*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$; n.s. = not significant

3.5 Visual task

To study peripheral glare in laboratory experiments requires that a visual task or target is given upon which to focus participant's visual fixation. In the earliest glare studies, observers were asked to evaluate the discomfort when directly viewing the glare source. In more recent studies visual tasks have ranged from a simple symbol [Berman et al 1994] to tasks intending to represent normal working [Wienold and Christoffersen 2006]. The choice of task is expected to influence the discomfort evaluation because different types of task require different degrees of cognitive attention and thus affect the cognitive resource available for the discomfort evaluation. Different types of tasks provide different abilities to maintain fixation and reduce glances toward the supposedly peripheral glare source.

Kent et al [2019c] evaluated discomfort due to peripheral glare using luminance adjustment and category rating procedures with two visual tasks. One was a simple fixation marker, and the second was a series of pseudo-text, a task expected to demand a greater degree of cognitive attention. The results demonstrate that the visual task significantly influenced the evaluation of discomfort, albeit a small effect size. When engaged in the pseudo-text task, participants were more tolerant to glare, seen as settings of higher luminance in the adjustment task and lower ratings of discomfort in the category rating task.

Kent et al [2019c] used their tasks to promote foveal fixation and hence maintain the glare source in peripheral vision: other studies have examined the effect of task difficulty itself. Osterhaus and Bailey [1992] found participants were less sensitive to glare when the evaluation was made immediately following a letter-counting task, which agrees with Kent et al [2019c]. While the studies of Sivak et al [1989] and Flannagan et al [1990] suggest at first the opposite, with higher discomfort being reported when the task was more difficult, this may not be the case. In their studies, this was a gap detection task, with variation in gap size used to vary the degree of difficulty. Figure 2 shows mean glare ratings plotted against the percentage of correctly identified gaps from Sivak et al [1989]: when plotted to show mean results in each test condition (left) or mean results for each participant (right) the data indicate a greater degree of discomfort as the percentage of correct responses decreases, i.e. as the task became more difficult. As commented by Sivak et al [1989], these data suggest that their participants were rating perceived disability rather than discomfort.

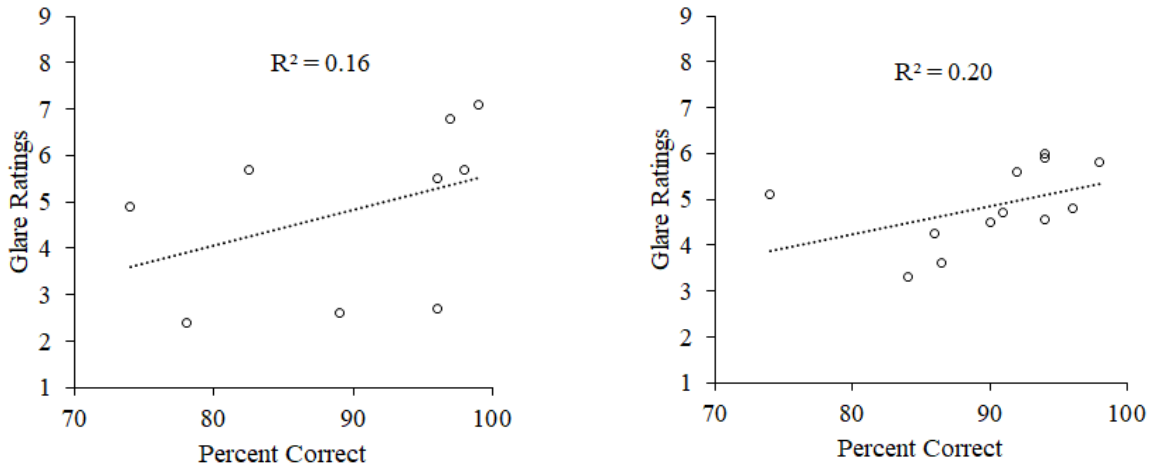

Figure 2. Mean glare ratings plotted against the percentage of correctly identified gaps. *Left*: results for each of the 9 test conditions; *Right*: results for each of the 12 test participants. Note: (1) glare ratings ranged from 1 = unbearable to 9 = just noticeable; (2) data points interpolated from Sivak et al 1989: *Left*, data from their Figs 3 and 4; *Right*; data from their Fig 5.

Tuaycharoen and Tregenza [2005, 2007] varied scene interest rather than task difficulty and found this had a significant effect on glare evaluations. Specifically, they found greater glare tolerance (i.e. glare perceived to be less discomforting) to images of scenes rated to be interesting than neutral scenes (i.e. a blank screen) of the same mean luminance (experiment 1), and a greater glare tolerance to natural scenes than urban scenes (experiment 2). In an anecdotal but common situation, people frequently sit for hours in front of a television set by free choice even though it is likely to be, according to the relevant discomfort model, producing intolerable glare [Markus 1974].

These data regarding task difficulty (and scene interest) show that the task conducted by a test participant whilst evaluating discomfort from glare can affect that evaluation. What we do not yet know is whether the concurrent tasks affect discomfort or instead the process of evaluating discomfort. In further work, the context in which a discomfort from glare is evaluated should resemble the context to which the findings are applied.

4. Experimental bias: Category rating

4.1 Stimulus range bias and order effect

In a category rating experiment, a series of scenes are evaluated which differ in magnitude of one or more independent variables. Consider changes in glare source luminance. Stimulus range refers to the minimum and maximum glare source luminances. While it is expected that stimulus range would bias evaluations of discomfort, having been demonstrated in investigations of reassurance and brightness [Fotios and Castleton, 2016, Fotios 2016, 2019], as yet there do not appear to be data investigating this in the context of discomfort from glare: further work is required.

Order refers to the sequence in which the different glare source luminances are evaluated. In a repeated measures experiment it is expected that the observation and evaluation of one scene would affect evaluation of the next [Staddon et al 1980]. While an order effect has been demonstrated in discomfort from glare studies using an adjustment procedure (see section 3.3), there do not appear to be any studies demonstrating this in discomfort from glare studies using category rating.

4.2 Pre-trial demonstration

For the category rating procedure, anchors are the visual and memory references held before the first trial, and, for a repeated measures design, the previously evaluated stimuli. These anchors are unknown to the experimenter and will vary between participants. It has been suggested that the categorical response range should be anchored to the stimulus range using Pre-Trial Demonstration (PTD) to demonstrate to observers the meaning of the upper and lower limits of a rating scale [Fotios and Houser 2009, Houser and Tiller 2003, Tiller and Rea 1992]. However, PTDs have been used rarely, if at all, in past studies of discomfort from glare.

Kent and Fotios [2019] compared evaluations of discomfort from glare using category rating without and without a PTD, and found higher glare ratings (a greater degree of discomfort for the same glare luminance) in those trials using a PTD. The effect, however, was small, and it is not clear how this should be implemented in practice. Specifically, which condition (i.e., with-PTD or no-PTD) provides a closer approximation to the degree of discomfort experienced in a natural setting.

4.3 Response scale design

Tourangeau et al [2000] described four steps in the category rating response procedure: comprehension, retrieval, judgement and response. A respondent is required to comprehend the question, retrieve from memory the relevant information, and match the internally generated answer to one of the available response categories. Errors are introduced at each step.

4.3.1 Number of response categories

Response scales in category rating studies have two or more categories describing the degree of discomfort. The number of categories used in a particular study is rarely, if at all, questioned. In outdoor lighting there is a common tendency to use a 9-point scale and describe this as a de Boer scale: this is sometimes done with the apparent assumption that a de Boer scale is somehow validated but that is not the case. The labeling of the numerical categories is addressed in Section 4.3.3.

It may be questioned why de Boer and Schreuder [1967] used a 9-point scale rather than the 4-point scale previously introduced by Hopkinson [Hopkinson 1940]. While Hopkinson initially used the four points as targets for adjustment settings, they have also been used as response categories in category rating procedures [Adrian & Schreuder 1970, Berman et al 1994, Fischer 1972]. One possible answer is that the 9-point and 4-point response categories are drawn from the same scale. This can be seen in Hopkinson [1940, his Figure 9], a series of glare descriptors, of which the relevant details are shown in Table 5. This suggests that Hopkinson intended for his four descriptors to be the borderlines between absolute levels of glare, for example, just intolerable being the border between uncomfortable glare and intolerable glare. The descriptors of these absolute levels are similar to those used in the de Boer scale.

Borderline thresholds are particularly relevant for the adjustment procedure. Adjustments to ‘acceptable’ (for example) can include a wide range of scene conditions, but ‘just acceptable’ targets a more specific condition. Furthermore, consider adjustment to ‘uncomfortable’ from a high anchor, in the region of being intolerable: intolerable is already uncomfortable, so no adjustment action would be required, and luminance settings for the two criteria would overlap. Category rating does not need to rely so heavily on borderline thresholds.

Table 5. Discomfort magnitude descriptors of the de Boer and Schreuder [1967] and Hopkinson [1940] scales. Those marked (A) to (D) are the four settings used by Hopkinson as targets for adjustment settings.

Degree of discomfort	Hopkinson [1940]	De Boer and Schreuder [1967]
1	Intolerable	Unbearable
2	**(A) Just intolerable**	
3	Uncomfortable glare	Disturbing
4	**(B) Just uncomfortable**	
5	Distracting but not uncomfortable glare	Just admissible
6	**(C) Satisfactory**	
7	Perceptible but not distracting glare	Satisfactory
8	**(D) Just not perceptible**	
9	No glare	Unnoticeable

This leads to the question of whether there is an optimal number of rating categories for measuring discomfort from glare. Too few choices may impede respondents to find the most proper state to express their sensation, whereas too many categories may be beyond the respondent's powers of discrimination, causing confusion and enlarging intra-individual differences [Matell and Jacoby 1971, Wang et al 2018]. 'Too few' is relative to the number of items being evaluated. If there are five response categories and six items, then at least two items must be placed in the same category regardless of whether or not that was the respondents' opinion. In other words, with too few categories the response scale loses the ability to discriminate between stimuli. To counter this forced grouping, the number of response categories should be similar to, or greater than, the number of items to be evaluated.

Test participants are able to reliably distinguish between approximately seven categories but that with more than seven categories confusions become more frequent [Miller 1956; Saaty and Ozdemir 2003]. This implies the number of scenes evaluated should be restricted to about seven or less when using category rating.

An alternative to discrete categories is to use a continuous linear scale (also known as an analogue scale) with end points labeled (for example) unbearable and unnoticeable. The response scale might be a short line on a paper-based response sheet or a slider on a PC screen. Test participants are able to choose any point along that scale to define the degree of discomfort rather than being restricted to discrete categories. The advantages include data which can be used for a greater number of statistical tests and which allow the response to be indicated with a greater degree of precision [Funke and Reips 2012]. For the paper-based version, the precision is related to the method of measurement, perhaps in mm units: for the on-screen version, the precision may be related to pixel size on screen.

In some studies, the linear scale is gradated with tick marks. This approach may counter position bias: if a participant wants to bisect the line to indicate a central response they may actually mark the line further to the left than desired [Foulsham et al 2013] whereas tick marks may provide a guide to prevent this. On the other hand, the presence of tick marks may bias the response distribution, with responses anchored by the tick marks [Matejka et al 2016] compared to the more even distribution found without use of tick marks.

An analysis in the context of thermal comfort evaluations concluded that discrete categories were preferable to linear scales because the linear scales exaggerated intra-individual differences [Wang et al 2018]. In other words, the linear scale reveals too-small differences between items which could not be differentiated properly by the human mind. In contrast, Funke and Reips [2012] concluded in favor of linear scales rather than a 5-point scale. They assessed this with consideration to how often respondents

would modify their ratings, concluding that the linear scale allowed them to communicate their evaluations more precision than with the categorical scale because they made more changes with the linear scale than the categorical scale. Both scales led, however, to similar mean ratings. It is clear that there is no definitive support for either discrete categories or analogue scales over one another.

4.3.2 Number of rating scales

The ASHRAE approach to measuring thermal comfort uses multiple scales (Table 6), these measuring thermal sensation, thermal comfort, thermal preference, and thermal satisfaction, with the collective responses being used to evaluate comfort [ASHRAE 2009]. In contrast, the measurement of discomfort from glare with category rating typically uses a single response scale, with levels of discomfort ranging from little or none to unbearable. One exception to this is Iwata *et al.* [1990/91] who sought three responses, intending to separate impressions of discomfort and satisfaction: these were labeled as the glare vote and the discomfort sensation vote (Table 7) and a two-alternative acceptability response (acceptable or not acceptable). Further work is needed to confirm whether this improves the measurement of discomfort from glare.

Table 6. Response scales for measuring thermal comfort [ASHRAE 2009].

Response point	Thermal sensation	Thermal preference	Comfort	Satisfaction
1	Cold	Much cooler	Very uncomfortable	Very dissatisfied
2	Cool	Cooler	Uncomfortable	Dissatisfied
3	Slightly cool	Slightly cooler	Slightly uncomfortable	Slightly dissatisfied
4	Neutral	No change	Neutral	Neutral
5	Slightly warm	Slightly warmer	Slightly comfortable	Slightly satisfied
6	Warm	Warmer	Comfortable	Satisfied
7	Hot	Much warmer	Very comfortable	Very satisfied

Table 7. Glare and discomfort response items used by Iwata et al [1990/91].

Glare vote	Discomfort sensation vote
4 Intolerable 3 Uncomfortable 2 Acceptable 1 Perceptible 0 Imperceptible	3 Very uncomfortable 2 Uncomfortable 1 Slight uncomfortable 0 Not uncomfortable

4.3.3 Category labels

In category rating, the test participant picks one of a series of discrete categories. In studies of discomfort from glare it is common for each category to be labeled with a discomfort sensation. In the 9-point response scale commonly named a de Boer scale, magnitude descriptors are given to the odd intervals (1, 3, 5, 7, and 9). Table 8 shows the labels that have been used in some studies. It can be seen for the three lower degrees of discomfort in particular that there are inconsistencies in the labels used.

Consider the lowest degree of discomfort. While in 1967 this was labeled as unnoticeable by de Boer and Schreuder [1967] some later studies labeled this instead as noticeable [Schmidt-Clausen & Bindels

1974] or just noticeable [Bullough et al 2008]. Unnoticeable and (just) noticeable are not the same. This difference highlights an additional problem: those scales using just noticeable as the lowest degree of discomfort do not offer respondents the ability to say 'no discomfort': the minimum response they can give is to say discomfort is noticeable. To demonstrate one option for dealing with this, consider Tokura et al. [1996] who asked their observers to first rate whether they perceived any glare (i.e., yes/no). If yes, then the experimenter would provide the subject with the glare scale to measure the magnitude of discomfort sensation. Conversely, when subjects indicated no glare, the assessment of that scene would finish.

Table 8. Examples of variations in discomfort magnitude descriptors in evaluations of discomfort from glare in six studies using a 9-point category rating response scale. Note that in de Boer-like scales the tendency is to label only the odd-numbered categories.

Degree of discomfort	**Study**				
	de Boer & Schreuder 1967, Villa et al 2017	**Schmidt-Clausen & Bindels 1974**	**Mortimer & Olson 1974**	**Kimura-Minoda & Ayama 2011**	**Bullough *et al.* 2008**
High discomfort	Unbearable	Unbearable	Intolerable	Unbearable	Unbearable
	Disturbing	Disturbing	Disturbing	Disturbing	Disturbing
	Just Admissible	Just Admissible	Just Acceptable	Just acceptable	Just Permissible
	Satisfactory	Acceptable	Satisfactory	Satisfactory	Satisfactory
Low discomfort	Unnoticeable	Noticeable	Not Noticeable	Just noticeable	Just Noticeable

Consider next the second lowest discomfort label. This is labeled as satisfactory in many scales. Given that this is an evaluation of discomfort and alleged to be more uncomfortable than glare which is noticeable, it is unclear what 'satisfactory' means. Gellatly and Weintraub [1990] asked test participants to arrange into order of magnitude five de Boer-type scale descriptors (unbearable; disturbing; just admissible; satisfactory; and unnoticeable). Of the 26 naïve test participants, only 7 placed satisfactory in the same location as did de Boer (i.e. one step more discomforting that just noticeable) while 15 people assumed it to be a lower level of discomfort and 4 a greater level of discomfort. These data do not suggest a consistent understanding of satisfactory glare. When this task was repeated by 14 experts only one matched the de Boer descriptor sequence. (Experts were defined as members of the Southeast Michigan Chapter of the Human Factors Society, of whom five had some familiarity with the de Boer scale and the rating of discomfort glare.)

Next consider the middle discomfort category. In the studies listed in Table 8 this is labeled as just admissible, just acceptable or just permissible. In other studies, the middle category may be labeled uncertain, undecided, no difference, neutral or similar. The middle category provides an easy escape for respondents who are disinclined to express a definite view [Matell and Jacoby 1972; Bishop 1987]. Poulton [1989] suggests that response ranges with middle values enhance response contraction bias, the tendency to avoid using the ends of a scale such that ratings converge toward the center of the response range, and that this can reduce the apparent distinction between stimuli. People are much more likely to select a middle response alternative on an issue when it is explicitly offered to them as part of the question than when it must be spontaneously volunteered: offering respondents a middle

alternative can therefore make a substantial difference in the division of opinion on an issue [Bishop 1987; Fotios and Atli 2012; Presser and Schuman 1980].

Rarely are the category labels explained. In one exception (Fekete et al 2010, their Table 1) a general impression label is associated with each discomfort label (Table 9). In that particular study, the discomfort response scale was used only to support the degree of discomfort presented in a test scene, with the dependent variable being reaction time to the onset of a target with and without this glare. The association between discomfort and general impression shown in Table 9 remains the opinion of the authors of that work: it is unknown if that opinion is shared by other researchers or by naïve test participants. Furthermore, it does little to aid determination of threshold criteria for design – is the aim to provide conditions considered as fair, or should designers aim for excellent? (For further exceptions where the category labels are defined, see Tables 12 to 14).

Table 8 also reveals an unequal distribution of positive and negative options. Typically, one (at most, but sometimes none) of the five response labels allows a response that discomfort is not perceived while the remaining four are for various degrees of discomfort. This inequality may lead to a response frequency bias: when the frequencies are unequal, observers tend to respond as if the frequencies were more nearly equal [Fotios and Cheal 2008; Senders and Sowards 1952]. This may arise from a preconception of chance, leading an observer to expect that where a large number of responses are given, each of the permitted responses will be correct on an approximately equal number of occasions.

Table 9. Labels of 'general impression' associated with degrees of discomfort from glare according to Fekete et al [2010, their Table 1].

Index	Glare	General impression
1	Unbearable	Bad
2	-	-
3	Disturbing	Inadequate
4	-	-
5	Just admissible	Fair
6	-	-
7	Satisfactory	Good
8	-	-
9`	Unnoticeable	Excellent

Further inspection of the response scales shown in Table 8 illustrates an additional problem: the responses categories do not always map to unique magnitudes of discomfort. For example, an extremely bright source may be considered unbearable, but responses that the discomfort were disturbing and noticeable would also be correct (but not just admissible or acceptable). There is a position bias associated with response scales, in that response categories arranged as a lower-higher order of discomfort magnitude are expected to elicit a different response to the same categories but arranged higher-lower [Friedman et al 1994]. If respondents pick the first suitable category, then category direction will affect the results [Keusch and Yan 2018].

Similar questions can be raised about the four levels of the multiple criterion scale widely used in experiments using luminance adjustment. Hopkinson [1940] included four degrees of discomfort: just not perceptible, satisfactory, just uncomfortable and just intolerable. Tuaycharoen and Tregenza [2005] reduced this to three levels (just noticeable, just uncomfortable and just intolerable) following a pilot study in which they found no difference in understanding of the two lower levels of discomfort [Tuaycharoen 2006]. Stone and Harker [1973] used a four-level MCM adjustment procedure, with

discomfort targets just perceptible, just distracting, just uncomfortable and just intolerable. They changed the second criterion from the commonly used 'just acceptable' to 'just distracting' to make the progression more consistent in their opinion. Stone and Harker is one of few adjustment studies to define the meaning of the four discomfort criteria.

4.3.4 Common understanding

Experimenters tend to assume that their own definition of rating items and category labels is shared by the participants of their experiment, a *common understanding*. If the understanding of the meaning of terms is not common it may lead to two problems: there may be increased response variance if different respondents had different understanding, and the experimenter may incorrectly interpret the results. It may not be wise to assume a common understanding; note, for example, disagreement between researchers as to the meaning of visual clarity [Fotios and Atli 2012] and the disagreement between naïve respondents about the magnitude order of labels in a de Boer-like rating scale [Gellatly and Weintraub, 1990].

One approach to targeting a consistent understanding is to better define the meaning of the response categories. This was reported to have been done in only a few cases [Huang et al, 2018, Ngai and Boyce 2000; Osterhaus and Bailey 1992]. In these, definition of the degree of discomfort is associated with a likely reaction to the discomfort or to the duration the discomfort might be tolerated before taking action. Further work is needed to substantiate the benefit of this approach.

Table 10 shows the seven category descriptions used by Ngai and Boyce when investing discomfort from overhead glare. The descriptions describe likely reactions of an occupant, similar to those used by Osterhaus and Bailey [1992] (Table 11). An interesting feature of Ngai and Boyce's category labels is that they combine borderline levels (just perceptible, just uncomfortable and just intolerable) similar to Hopkinson [1940] with absolute levels (imperceptible, noticeable, uncomfortable, and intolerable) similar to de Boer and Schreuder [Schreuder 1967] (see also Table 5).

Huang *et al.* [2018] state that they used the response scale proposed by Ngai and Boyce: comparison of Table 10 with Table 12 shows that there were differences. Some of these differences may have changed how test participants responded. It may also illustrate the problem of (it is assumed) a two-way translation (here, English-Chinese-English).

Table 10. Descriptions of response categories used by Ngai and Boyce [2000].

Category	Name	Description of reaction
1	Imperceptible	I am not aware of anything overhead
2	Just perceptible	I am aware there is something overhead but cannot tell what it is
3	Noticeable	I am aware of the presence of the luminaire overhead but it does not bother me
4	Just uncomfortable	I am aware of a luminaire overhead and I wish it was not there
5	Uncomfortable	I am aware of a luminaire overhead and I would complain to my supervisor about it
6	Just intolerable	I am aware of a luminaire overhead and if somebody doesn't do something about it I will take direct action myself
7	Intolerable	I am aware of a luminaire overhead and I will not stay here a moment longer if somebody doesn't do something about it, now

Table 11. Descriptions of response categories used by Osterhaus and Bailey [1992]

Borderline between:	Description of reaction
Imperceptible and noticeable	A very slight experience of discomfort that they could tolerate for approximately one day when placed at someone else's workstation, but which they would rather change if they were to work here for longer periods of time.
Noticeable and disturbing	A discomfort experience that would be just disturbing and could be tolerated for 15 to 30 minutes, but that would require a change in luminance setting for any longer period.
Disturbing and intolerable	The turning point where subjects would no longer be able to tolerate the lighting condition

Table 12. Descriptions of response categories used by Huang et al [2018].

Glare ratings	Observer feeling	Description of each glare rating
1	Imperceptible	I can see nothing
2	Perceptible	I am aware there is something but can't tell what it is
3	Noticeable	I can feel the light clearly but it does not make me feel uncomfortable
4	A little uncomfortable	I am aware of the luminance and I wish it was not there
5	Very uncomfortable	I am aware of the luminance and I would complain to my supervisor about it
6	A little intolerable	I am aware of the glare and want to look away from it.
7	Totally intolerable	The light makes me feel crazy

4.3.5 Language translation

While de Boer worked in the Netherlands, and probably delivered instructions to test participants in Dutch language, his widely known reports were written in English language requiring that the response label categories were translated. When scales are translated across languages there are two forms of discrepancies that may occur. First, when the original descriptors are translated there may be not a direct word linking it to the second language. The second is that, there may be semantic bias during the translation process with the result that some descriptors may lose their original meaning. Villa et al [2017] included in their report the original French language of the five discomfort labels they used, and similarly Adrian and Schreuder [1970] for work conducted in Germany, but these are rare examples. Adjustment settings within the multiple criterion method are frequently made to the 'just' thresholds. Iwata et al. [1990/91] stated "*One additional difficulty is that the Japanese language does not have a specific word for 'just'*." Studies should report test instructions in both the original and published languages; doing so allows others to check the accuracy of translation.

4.4 Statistical analysis of rating data

Categorical rating scales require respondents to select one of a series of categories which best describes the observed scene. The outcome of a series of evaluations of a specific scene are the numbers of respondents selecting each category, commonly reported as the average and variance of the integers assigned to each category and with the differences between scenes analyzed using statistical tests.

While the decision to use parametric rather than nonparametric statistical tests should follow confirmation that the data fit the assumptions of a normally distributed population, it is uncommon in

reports of discomfort from glare to see these assumptions confirmed. If the data are not normally distributed about the mean and are not at the interval scale, parametric data analysis techniques may not be appropriate and their use may lead to incorrect conclusions. It is, however, also possible to transform data so that they become normally distributed by the application of a mathematical function to each of the individual ratings (e.g. using the square root or logarithm of the original value). With large sample sizes, bootstrapping may be appropriate [Efron and Tibshirani, 1994].

It has been suggested that response scales with at least five categories may be analyzed as though they are parametric data (assuming that they meet also the general requirements for a normal distribution) but that four or fewer categories should not [Harpe 2015, Hsu and Feldt 1969]. This is an interesting threshold for studies of discomfort from glare, where, for no known reason, outdoor lighting studies have tended to adopt a de Boer-like 9-point scale and interior lighting studies have tended to adopt a 4-point Hopkinson-like response scale.

5. Improving the measurement of discomfort from glare

5.1 Planning an experiment

In any procedure used to explicitly measure the discomfort from glare, all aspects of that procedure influence the outcome. Variations in experimental design will affect the precision and/or accuracy of the results and they may affect the degree to which findings can be generalized beyond the context of the experiment. The precision and accuracy of the results are influenced by anchors, PTD limits, order effects and response scale design. Generalization is influenced by stimulus range bias, category labels, the difference between direct and indirect control and the visual task employed in parallel with the discomfort evaluation. Some factors can be accounted for; for example, the effect of anchors can be offset by using both high and low anchors, and order effects can be offset by randomizing test sequences. Other factors can be chosen to represent the conditions of a specific application, for example using direct adjustment control for investigation of single-occupant environments. It is not known whether stimulus range bias can be countered. What might be done instead is to recognize that the results show relative effects, such as whether one scene leads to a greater or lower degree of discomfort than another, and should not be interpreted to establish an absolute threshold.

The issues described in the current article are not suggested to be exhaustive. The recent special issue of Leukos focusing on research methods carried papers raising further questions regarding category rating [Fotios 2019], ethical issues and reporting [Veitch et al 2019] and statistical analysis [Uttley 2019].

Decisions such as illuminance range, number of points on a rating scale, direct or indirect adjustment are amongst the many an experimenter must make when planning an experiment, some of which may appear arbitrary. These may be considered as researcher degrees of freedom [Wicherts et al 2016]. *P-hacking* describes the situation where researchers may opportunistically use these degrees of freedom.

There is a need for cautious and careful research to counter false positives, because once they appear in the literature, they can be persistent [Simmons et al 2011, Fotios 2017]. One proposal is for experimental procedures to be registered (and possibly, but not essentially, peer reviewed) prior to an experiment being conducted [Munafò et al 2017, Wicherts et al 2016]. A pre-registered procedure would include descriptions of the procedure, the sample size and targeted make-up, the lighting conditions, the results to be analyzed and the statistical tests to be conducted. Resultant reports could then be compared against pre-registered experimental procedures to confirm that the proposed procedure was followed. Doing so would have a number of benefits. It would counter the natural

tendency of enthusiastic scientists to be misled by a tendency to see structure in randomness [Munafò et al 2017], leading to a false positive, which is the incorrect rejection of a null hypothesis [Simmons et al 2011]. It would require the results from all conditions and procedures to be reported, rather than the selective reporting of favorable findings.

5.2 Promoting robust data: Null conditions and counterbalancing

Experiments should include some means by which to counter alternative explanations for the findings and to promote confidence that the conclusions drawn are warranted [Veitch et al 2019].

Consider a hypothesis that a change in glare source SPD (X) leads to a change in the evaluated magnitude of discomfort from glare (Y). Null conditions are trials in which there is no change in X and hence no expected change in Y: if a change in Y is found, it reveals the presence and magnitude of an unintended bias (see Table 13). In simultaneous evaluations (e.g. side-by-side comparisons) a null condition means that the two visual scenes are identical (or, intended to be identical); that is, they are lit by lamps of identical SPD, with equal luminances and spatial distributions. A significant difference between the two scenes suggests that the two fields were not identical, as intended, or that there is some asymmetry in observers' responses, such as a bias toward one position over the other. In either case, the difference suggests a bias in those trials where the scenes were purposefully different. In separate evaluations, (e.g. a series of scenes are evaluated individually, one after another) the null condition might be repeated evaluation of a particular scene with the expectation that the first and second evaluations will agree.

Table 13. Examples of null condition trials and counterbalancing that should be included as a means of exploring and countering experimental bias.

Evaluation mode		**Null condition**	**Counterbalancing**
Separate	Scenes are observed individually and evaluated before observation of next scene	The same scene is evaluated twice within the series of test scenes.	Presentation order is randomized
Sequential	Two scenes are presented in temporal alternation (1^{st}, 2^{nd}, 1^{st}, 2^{nd} …)	Evaluation conducted using two identical scenes	Interval order (1^{st} and 2^{nd}) is alternated. Stimulus pairs are presented in a randomized order.
Simultaneous	Two scenes are presented simultaneously in adjacent spatial locations (left-right, top-bottom, center-surround)	Evaluation conducted using two identical scenes	Spatial position (e.g. left and right) is alternated. Stimulus pairs are presented in a randomized order.

Consider an experiment comparing several levels of glare source CCT but not revealing a significant effect of CCT on evaluations of discomfort. There are three explanations: (1) that there is no effect of CCT on discomfort and the experiment has correctly confirmed this; (2) that there is an effect of CCT on discomfort but the experiment was not sufficient to reveal it, either through procedure or the choice of CCT levels; or (3) that CCT, being a generally insufficient proxy for variations in SPD, has confounded stimulus definitions. Including extreme levels of CCT would enable the second explanation to be countered. Extreme levels of CCT would be those which, according to previous results or theory, are expected to lead to large and significant differences in evaluated discomfort.

While null condition trials may reveal a problem, counterbalancing should be used to offset expected problems. Counterbalancing is a carefully planned schedule in which the variables are included in all possible combinations. For side-by-side evaluations, counterbalancing includes alternating the spatial location (e.g. left and right) in which scenes are observed; for separate evaluations, counterbalancing means observing and evaluating the sample of visual scenes in an order which is balanced across observers if not randomized.

5.3. Recommendations for good practice

This section presents recommended guidance for procedures used to investigate the degree of discomfort experienced under different lighting conditions using subjective (explicit) measurements. Some of these items are essential, others may be considered desirable.

Appendix 2 shows procedural steps required to promote credible data; these steps are pertinent not only to evaluations of discomfort from glare but also to a range of other psychophysical responses. For the adjustment and category rating procedures these recommendations follow from the discussions above. For the matching and discrimination procedures, where there is little empirical evidence regarding bias in the context of discomfort evaluation, these recommendations follow those made for investigation of spatial brightness [CIE 2014].

A converging operations approach is recommended where feasible. Converging operations is where the same set of stimuli are examined using different experimental procedures. If the results of two or more procedures lead toward the same conclusion then more confidence can be placed in the robustness of that conclusion. Converging operations can involve variations in research design and in the outcome measures, or both together. For example, category rating and adjustment procedures were used in parallel in three studies [Osterhaus & Bailey 1992, Ngai & Boyce 2000, Ramasoot & Fotios 2012]. A caveat to converging operations is the potential for opportunistic findings – reporting the findings of the procedure which resulted in convenient findings and ignoring those of the other procedure [Wicherts et al 2016]. That is not the intention of using multiple procedures. Rather, if the findings of different procedures do not agree, an investigation as to the cause of disagreement could lead to an improved understanding of experimental design.

Studies of discomfort from glare may be categorized as having either static or dynamic characteristics of lighting. Static characteristics are common of controlled laboratory studies, where the visual scene is set, one at a time, to a series of discrete conditions. Field studies, of either interior or exterior electric lighting, also tend to use static characteristics. Dynamic characteristics are those encountered in field studies of daylight where the characteristics of daylight, being naturally variable, are likely to change at each moment of evaluation. While the nature of the proposed recommendations (e.g. Table 13 and Appendix 2) may feel more applicable to static conditions than to dynamic conditions, that is not intended to be the case. For example, range bias is likely to persist whether or not the test conditions are static (as set by the experimenter) or dynamic (naturally variable). Instead the recommendations should prompt actions such as recording window luminances at each moment of evaluation to enable post-hoc analysis of range bias, and considering the use of additional and/or alternative procedures.

Comprehensive reporting of an experiment and its results is necessary for independent analysis of the original data and replication of the experiment [Wicherts et al 2016]. Sufficient data should be reported to enable readers to understand how the experiment was conducted and, if necessary, to repeat it. The relevant data to include is described elsewhere [CIE 212:2014, Simmons et al 2011, Veitch et al 2019, Wicherts et al 2016].

We recommend to include an objective justification of sample size such as using an analysis of statistical power. Unfortunately, this is rarely seen in studies of discomfort from glare. Failure to justify sample size is problematic because researchers' intuitions about statistical power are overly optimistic, and small sample sizes have greater potential to be influenced by research degrees of freedom [Wicherts et al 2016]. Simmons et al. [2011] propose that samples should comprise at least 20 observations: smaller samples are not usually powerful enough to detect most effects, larger samples do not necessarily lead to a lower p-value. Field [2005] suggests a minimum sample of 28 to reveal a large effect size. Rather than a power analysis, the sample may also be subject to a pragmatic limitation such as the number of occupants in a particular building. Regardless of the approach taken, authors should establish the sample before an experiment begins and report this rule in their article [Simmons et al 2011].

5.4 Further research

This article has identified some evidence regarding experimental bias in the psychophysical procedures commonly used to measure discomfort from glare. The review has also raised further questions for which further research is required.

While category rating is widely used to evaluate discomfort from glare, there remain many uncertainties. These include the influence of stimulus range bias and order effects; how (if at all) to use PTDs; and response scale design (number of response points, category labeling, category numbering, discomfort definitions, single versus multiple response scales). On the other hand, for the adjustment procedure, there has been much work recently conducted to explore the effect of changes in experimental design [Kent et al 2018, 2019a, 2019b, 2019c, Kent and Fotios 2019], although this is not proposed to be an exhaustive investigation.

To date, the commonly used methods have failed to reach a consensus regarding the effects of glare on discomfort. Significant development may require the introduction of new approaches. This might be a subtle change in the procedure: rather than asking for glare source luminance to be adjusted to a particular level of discomfort, ask instead for test participants to set the luminance as high as possible but in which they could still work [Rohles 2007, his figure 1]. Further possibilities have also been described by Fotios [2018].

All subjective evaluations are likely to be biased in some way [Poulton 1977]. A participant's response may be influenced by the nature of the question, the nature of the response mechanism, and by their preconceived notions as to what the correct response should be or the response they believe the experimenter desires. These problems are reduced if implicit measurements are used rather than explicit measurements. In the context of discomfort from glare evaluations, implicit measures include involuntary physiological responses and coping strategies – changes made by occupants to their environment to alleviate discomfort. Table 14 summarizes the measurements used in past studies of discomfort from glare.

Table 14. Examples of discomfort glare studies in which methods other than subjective psychophysical procedures were used to measure discomfort.

Method	Examples of studies using the method
Pupil response (size; change in size; hippus)	Fry & King 1975; Hopkinson 1956; Howarth et al 1993; Lin et al 2015; Stringham et al 2011; Tyukhova and Waters 2019
Electrograms (EMG, EOG etc)*	Berman et al 1994; Murray et al 2002; Lin et al 2015
Degree of eye opening	Yamín Garretón et al 2015
Brain activity (fMRI)	Bargary et al. 2015
Gaze behavior	Sarey Khanie et al 2015
Shutting window blinds or changing seating position	O'Neil 2015; Jakubiac and Reinhart 2012

*See Reilly and Lee [2010] for definitions of electrogram measurements.

6. Conclusions

This article has discussed investigation of discomfort from glare. It has focused on explicit measurement – subjective evaluations of the degree of discomfort – as this is the most common approach, and in particular the use of luminance adjustment and category rating. Evidence is presented to demonstrate that some aspects of these procedures, such as the range of glare source luminances available in an adjustment procedure, influence the resulting evaluation. Evidence by omission suggests that these procedures are given little, if any, attention in previous studies, leading to variance between studies.

The aim of this article is to raise awareness of undesirable bias and approaches that may be employed to counter it. It is, however, likely that bias will persist, albeit reduced. There may be a benefit in employing implicit methods, such as physiological measurement and behavioral observation. While these methods have been used, there are far fewer studies than those using explicit measurement, there are insufficient data to enable analysis of experimental bias, and as yet they do not appear to be feeding into models of discomfort from glare.

Funding

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

Disclosure statement

No potential competing interest was reported by the authors.

References

Adrian W, Schreuder DA. 1970. A simple method for the appraisal of glare in street lighting. Lighting Res Technol. 2(2); 61-73.

ASHRAE. 2009. ASHRAE Handbook: Fundamentals. Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

Bargary G, Furlan M, Raynham PJ, Barbur JL and Smith AT. 2015. Cortical hyperexcitability and sensitivity to discomfort glare. Neuropsychologia. 69: 194-200.

Bargary G, Jia Y and Barbur JL. 2014. Mechanisms for discomfort glare in central vision. Invest Ophth Vis Sci. 56(1): 464-471.

Berman SM, Bullimore MA, Jacobs RJ, Bailey IL, Gandhi N. 1994. An objective measure of discomfort glare. J Illum Eng Soc. 23 (2): 40–49.

Bishop GF, 1987. Experiments with the middle response alternative in survey questions. Public Opin Q. 51; 220-232.

De Boer JB, Schreuder DA. 1967. Glare as a Criterion for Quality in Street Lighting. Trans Illum Eng Soc. 32(2); 117-135.

Boyce PR, Eklund NH, Hamilton BJ, Bruno LD. 2000. Perceptions of safety at night in different lighting conditions. Lighting Res Technol. 32: 79-91.

Bullough J, Brons JA, Qi R, Rea MS. 2008. Predicting discomfort glare from outdoor lighting installations. Lighting Res Technol. 40; 225-242.

Cambridge Dictionary. 2019a. Cambridge (UK): Cambridge University Press; [accessed 2019 Oct 16]. https://dictionary.cambridge.org/dictionary/english/comfort

Cambridge Dictionary. 2019b. Cambridge (UK): Cambridge University Press; [accessed 2019 Oct 16]. https://dictionary.cambridge.org/dictionary/english/discomfort

Chapman GB, Johnson EJ. Anchoring, activation, and the construction of values. Organizational, Behaviour and Human Decision Processes 1999; 79(2): 115-153.

[CIE] Commission Internationale de l'Éclairage. 2019a. CIE e-ILV Term 17-492 Glare. Vienna(Austria): CIE; [accessed 2019 Oct 16]. http://eilv.cie.co.at/term/492

[CIE] Commission Internationale de l'Éclairage. 2019b. CIE e-ILV Term 17-333 Discomfort Glare. Vienna(Austria): CIE; [accessed 2019 Oct 16]. http://eilv.cie.co.at/term/333

[CIE] Commission Internationale de l'Éclairage. CIE 212:2014. Guidance Towards Best Practice In Psychophysical Procedures Used When Measuring Relative Spatial Brightness. Commission Internationale De L'Éclairage, Vienna, 2014.

Collins WM. 1962. The determination of the minimum identifiable glare sensation interval using a pair-comparison method. Trans Illum Eng Soc. 27(1): 27-34.

Durlak JA. 2009. How to select, calculate, and interpret effect sizes. J Pediatr Psychol. 34(9): 917–928.

Efron B and Tibshirani RJ. 1994. An introduction to the Bootstrap. CRC Press. Boca Raton.

Fekete J, Sik-Lanyi C, Schanda J. 2010. Spectral discomfort glare sensitivity investigations. Ophthal Physl Opt. 30(2): 182-187.

Ferguson CJ. 2009. An effect size primer: A guide for clinicians and researchers. Prof Psychol-Res Pr. 40(5): 532-538.

Field A. 2005. Discovering statistics using SPSS. Sage publications; London.

Fischer D. 1972. The European glare limiting method. Lighting Res Technol. 4(2); 97-100.

Flannagan MJ, Weintraub DJ, Sivak M. 1990. Context Effects on Discomfort Glare : Task and Stimulus Factors (UMTRI-90-35). Ann Arbor, USA.

Fotios S. 2019. Using category rating to evaluate the lit environment: Is a meaningful opinion captured? Leukos. 15(2-3): 127-142.

Fotios S. 2018. Correspondence: New methods for the evaluation of discomfort glare. Lighting Res Technol. 50(3): 489-491.

Fotios S. 2017. A revised Kruithof graph based on empirical data. Leukos. 13(1); 3-17.

Fotios S. 2016. Comment on of empirical evidence for the design of public lighting. Safety Sci. 86; 88-91.

Fotios S, Atli D. 2012. Comparing Judgements of Visual Clarity and Spatial Brightness Through an Analysis of Studies Using the Category Rating Procedure. Leukos. 8(4); 261-281.

Fotios S, Atli D, Cheal C, Houser K, Logadóttir A. 2015. Lamp spectrum and spatial brightness at photopic levels: A basis for developing a metric. Lighting Res Technol. 47(1); 80-102.

Fotios S, Castleton H. 2016. Specifying enough light to feel reassured on pedestrian footpaths. Leukos. 12(4): 235-243.

Fotios SA, Cheal C. 2010. Stimulus range bias explains the outcome of preferred-illuminance adjustments. Lighting Res Technol. 42(4); 433-447.

Fotios SA, Cheal C. 2008. The effect of a stimulus frequency bias in side-by-side brightness ranking tests. Lighting Res Technol. 40(1); 43-54.

Fotios SA, Houser KW. 2009. Research methods to avoid bias in categorical ratings of brightness. Leukos. 5(3); 167-181

Foulsham T, Gray A, Nasiopoulos E, Kingstone A. 2013. Leftward biases in picture scanning and line bisection: A gaze-contingent window study. Vision Res. 78: 14-25.

Friedman HH, Herksovitz PJ, Pollack S. 1994. Biasing effects of scale-checking styles on responses to a Likert scale. Proc. of the American Statistical Association Annual Conference: Survey Research Methods. 792-795.

Fry GA, King VM. 1975. The pupillary response and discomfort glare. J Illum Eng Soc. 4(4): 307-324.

Funke F, Reips U-D. 2012. Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. Field Method. 24(3): 310-327.

Gellatly & Weintraub. 1990. User Reconfigurations of the de Boer Rating Scale for Discomfort Glare. The University of Michigan Transportation Research Institute.

Gescheider GA. 1997. Psychophysics: The Fundamentals. Mahwah, NJ: Lawrence Erlbaum Associates.

Green PE, Rao VR. 1970. Rating scales and information recovery – how many scales and response categories to use? J Marketing 34(3): 33-39.

Golafshani N. 2003. Understanding reliability and validity in qualitative research. Qual. Rep. 8(4): 597–606.

Halkjelsvik T, Jørgensen M, Teigen KH. 2011. To read two pages, I need 5 minutes, but give me 5 minutes and I will read four: How to change productivity estimates by inverting the question. App Cognitive Psych. 25(2): 314-323.

Harpe SE. 2015. How to analyse Likert and other rating scale data. Currents in Pharmacy Teaching and Learning 2015; 7: 836-850.

Howarth PA, Heron G, Greenhouse DS, Bailey IL and Berman SM. 1993. Discomfort from glare: The role of pupillary hippus. Lighting Res Technol. 25(1): 37-42.

Hopkinson RG. 1940. Discomfort glare in lighted streets. Trans Illum Eng Soc. 5; 1-32.

Hopkinson RG. 1950. The multiple criterion technique of subjective appraisal. Q. J. Exp. Psychol. 2(3) 124–131.

Hopkinson RG. 1956. Glare discomfort and pupil diameter. J Opt Soc Am. 46:649–656.

Houser KW, Tiller DK. 2003. Measuring the subjective response to interior lighting: Paired comparisons and semantic differential scaling. Lighting Res. Technol. 35(3): 183-195.

Hsu T-C, Feldt LS. 1969. The effect of limitations on the number of criterion score values on the significance level of the F-test. Am Educ Res J. 6(4): 515-527.

Huang WJ, Yang Y, Luo MR. 2018. Discomfort glare caused by white LEDs having different spectral power distributions. Lighting Res Technol. 50(6): 921-936.

Iwata T, Kimura K-I, Shukuya M, Takano K. 1990/91. Discomfort caused by wide-source glare. Energ Buildings 15(3-4): 391-398.

Jakubiac JA, Reinhart CF. 2012. The 'adaptive zone' – A concept for assessing discomfort glare throughout daylit spaces. Lighting Res Technol. 44: 149-170.

Kent MG, Altomonte S, Tregenza PR, Wilson R. 2015. Discomfort glare and time of day. Lighting Res Technol. 47: 641-657.

Kent MG, Fotios S, Altomonte S. 2018. Order effects when using Hopkinson's multiple criterion scale of discomfort due to glare. Building Environ. 136: 54-61.

Kent M, Fotios S. 2019. The effect of a pre-trial range demonstration on subjective evaluations using category rating of discomfort due to glare. Leukos. Online first, 23/07/2019. doi.org/10.1080/15502724.2019.1631177.

Kent MG, Fotios S, Cheung T. 2019a. Stimulus range bias leads to different settings when using luminance adjustment to evaluate discomfort due to glare. Building Environ. 153; 281-287.

Kent M, Fotios S, Altomonte S. 2019b. Discomfort glare evaluation: The influence of anchor bias in luminance adjustments. Lighting Res Technol. 51(1): 131-146.

Kent M, Fotios S, Altomonte S. 2019c. An experimental study on the effect of visual tasks on discomfort due to peripheral glare. Leukos 51(1): 17-28.

Keusch F, Yan T. 2018. Is satisficing responsible for response order effects in rating scale questions? Survey Research Methods 2018; 12(3): 259-270.

Kim W, Kim JT. 2011. A position index formula for evaluation of glare source in the visual field. Indoor Built Environ 20(1); 47-53.

Kimura-Minoda T, Ayama M. 2011. Evaluation of discomfort glare from color LEDs and its correlation with individual variations in brightness sensitivity. Color Res Appl. 36(4): 286-294.

Lee SY, Brand JL. 2005. Effects of control over office workspace on perceptions of the work environment and work outcomes. J. Environ. Psychol. 25: 323–333.

Lin Y, Fotios S, Wei M, Liu Y, Guo W, Sun Y. 2015. Eye movement and pupil size constriction under discomfort glare. Invest Ophth Vis Sci. 56(3); 1649-1656.

Logadóttir Á, Fotios SA, Christoffersen J, Hansen SS, Corell DD, Dam Hansen C. 2013. Investigating the use of an adjustment task to set preferred colour of ambient illumination, Color Res Appl. 2013; 38(1); 46-57.

Logadóttir Á, Christoffersen J, Fotios SA. 2011. Investigating the use of an adjustment task to set preferred illuminance in a workplace environment. Lighting Res Technol. 43(4); 403-422.

Luckiesh M, Guth SK. 1949. Brightnesses in visual field at borderline between comfort and discomfort (BCD). Illum Eng 44: 650–670.

Luckiesh M, Holladay LL. 1925. Glare and visibility. Trans Illum Eng Soc. March; 221-247.

Lulla AB, Bennett CA. 1981. Discomfort glare: Range effects. J Illum Eng Soc. 10(2): 74-80.

Markus TA. 1974. The why and the how of research in real buildings. Journal of Architectural Research. 3(2): 19-23.

Matejka J, Glueck M, Grossman T, Fitzmaurice G. 2016. The effect of visual appearance on the performance of continuous sliders and visual analogue scales. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. San Jose, California, 7-12 May 2016. 5421-5432.

Matell MS, Jacoby J. 1971. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. Educ Psychol Meas. 31: 657-674.

Matell MS, Jacoby J. 1972. Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. J Appl Psychol. 56(6): 506-509.

Miller GA. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. The Psychological Review 63(2): 81-97.

Mortimer RG, Olson PL. 1974. Evaluation of meeting beams by field tests and computer simulation. Report No. UM-HSRI-HF-74-27.Ann Arbor, Michigan, USA: Highway Safety Research Institute, The University of Michigan.

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers E-J, Ware JJ, Ionnidis JPA. 2017. A manifesto for reproducible science. Nat Hum Behav. 1, Article number: 0021: 1-9.

Murray I, Plainis S, Carden D. 2002. The ocular stress monitor: a new device for measuring discomfort glare. Lighting Res Technol. 34: 231–239.

Mussweiler T, Strack F. 2001. The semantics of anchoring. Organ Behav Hum Dec. 86(2): 234-255.

Ngai P, Boyce P. 2000. The effect of overhead glare on visual discomfort. J Illum Eng Soc. 29(2): 29-38.

Nuzzo R. 2014. Statistical errors. Nature; 505: 150-152.

O'Neil SM. 2015. Quantifying adaptive behavioral responses to discomfort glare – a comparative analysis of daylit offices. Thesis for Master of Science. University of Oregon, USA.

Osterhaus WKE. 2005. Discomfort glare assessment and prevention for daylight applications in office environments. Sol Energy; 79: 140-158.

Osterhaus WKE, Bailey IL. 1992. Large area glare sources and their effect on discomfort and performance at computer workstations. Proceedings of the 1992 IEEE Industry Applications Annual Meeting, 4-9 October 1992, Houston, TX; New York: IEEE.

Oxford Dictionaries. 2019. Lexico.com. Oxford (UK): Oxford University Press; [accessed 2019 Oct 16]. https://www.lexico.com/en/definition/comfort

Parker S, Schneider B. 1994. The stimulus range effect: evidence for top-down control of sensory intensity in audition. Percept. Psychophys. 56: 1–11.

Petherbridge P, Hopkinson RG. 1950. Discomfort glare and the lighting of buildings. Trans Illum Eng Soc. 15: 39–79.

Pierson C, Wienold J, Bodart M. 2018. Review of factors influencing discomfort glare perception from daylight. Leukos. 14(3): 111-148.

Poulton EC. 1977. Quantitative subjective assessments are almost always biased, sometimes completely misleading. Brit J Psychol. 68: 409-425.

Poulton EC. 1989. Bias in Quantifying Judgements. Lawrence Erlbaum Associates: Hove, East Sussex, UK.

Presser S, Schuman H. 1980. The measurement of a middle position in attitude surveys. Public Opin Q. 44, 70-85.

Pulpitlova J, Detkova P. 1993. Impact of the cultural and social background on the visual perception in living and working perception. Proceedings of the International Symposium: Design of Amenity, 5-8 October, Fukuoka, Japan, 1993, pp. 93–95.

Ramasoot T, Fotios SA. 2012. Lighting and display screens: Models for predicting luminance limits and disturbance, Lighting Res Technol. 44(2); 197-223.

Reilly R, Lee T. 2010. Electrograms (ECG, EEG, EMG, EOG). Technol Health Care. 18:443-458.

Rodriquez RG, Pattini A. 2014. Tolerance of discomfort glare from a large area source for work on a visual display. Lighting Res Technol. 46: 157-170

Rohles FH. 2007. Temperature and temperament. A psychologist looks at comfort. ASHRAE Journal. February, 14-22.

Sarey Khanie, M.; Stoll, J.; Einhauser, W.; Wienold, J.; Andersen, M. 2016. Gaze and discomfort glare, Part 1: Development of a gaze-driven photometry. Lighting Res Technol. 49(7): 845–865.

Saaty TL, Ozdemir MS. 2003. Why the magic number seven plus or minus two. Math. Comput. Model. 38(3): 233–244.

Schmidt-Clausen HJ, Bindels JTH. 1974. Assessment of discomfort glare in motor vehicle lighting. Lighting Res Technol. 6(2); 79-88.

Senders VL, Sowards A. 1952. Analysis of response sequences in the setting of a psychophysical experiment. Amer. J. Psychol. 65(3): 358–74.

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 22(11): 1359-1366.

Sivak M, Flannagan MJ, Ensing M, Simmons CJ. (1989). Discomfort glare is task dependent (UMTRI-89-27). Ann Arbor, USA.

Staddon JER, King M, Lockhead GR. 1980. On sequential effects in absolute judgement experiments. J Exp Psychol. 6(2): 290-301

Stone PT, Harker SDP. 1973. Individual and group differences in discomfort glare responses. Lighting Res. Technol. 5(1): 41–49.

Stringham JM, Garcia PV, Smith PA, McLin LN, Foutch BK. 2011. Macular pigment and visual performance in glare: benefits for photostress recovery, disability glare, and visual discomfort. Invest Ophthalmol Visual Sci. 52: 7406–7415.

Tiller DK, Rea MS. 1992. Semantic differential scaling: Prospects in lighting research. Lighting Res. Technol. 24(1): 43-51.

Tokura M, Iwata T and Shukuya M. 1996. Experimental study on discomfort glare caused by windows part 3. Journal of Architecture, Planning and Environmental Engineering. 489: 17-25.

Tourangeau R, Rips RJ, Rasinski K. 2000. The psychology of survey response. New York: Cambridge University Press.

Tuaycharoen N, 2006. The reduction of discomfort glare from window by interesting views. PhD Thesis: University of Sheffield.

Tuaycharoen N, Tregenza PR. 2005. Discomfort glare from interesting images. Lighting Res Technol. 37: 329–341.

Tuaycharoen N, Tregenza PR. 2007. View and discomfort glare from windows. Lighting Res Technol. 39: 185–200.

Tyukhova Y, Waters CE. 2018. Discomfort glare from small, high-luminance light sources when viewed against a dark surround. Leukos 14(4): 215-230.

Tyukhova Y, Waters CE. 2019. Subjective and pupil responses to discomfort glare from small, high-luminance light sources. Lighting Res Technol. 51: 592-611.

Uttley J, Fotios S, Cheal C. 2013. Satisfaction and illuminances set with user-controlled lighting. Archit Sci Rev. 56(4); 306-314.

Uttley J. 2019. Power analysis, sample size, and assessment of statistical assumptions – Improving the evidential value of lighting research. Leukos. 15(2-3): 143-162.

Veitch JA, Fotios SA, Houser KW. 2019. Judging the scientific quality of applied lighting research. Leukos. 15(2-3): 97-114.

Villa C, Bremond R, Saint-Jacques E. 2017. Assessment of pedestrian discomfort glare from urban LED lighting. Lighting Res Technol. 49(2): 147-172.

Wang J, Wang Z, de Dear R, Luo M, Ghahramani A. 2018. The uncertainty of subjective thermal comfort measurement. Energ Buildings. 181: 38-49.

Waters CE, Mistrick RG, Bernecker CA. 1995. Discomfort glare from sources of nonuniform luminance. J Illum Eng Soc. 24(2): 73-85.

Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. Front Psychol. 7: article 1832; 1-12.

Wienold J and Christoffersen J. 2006. Evaluation methods and development of a new glare prediction model for daylight environments with the use of CCD cameras. Energ Buildings. 38(7): 743-757.

Yamín Garretón JA, Rodriguez RG, Pattini AE. 2015. Degree of eye opening: A new discomfort glare indicator. Build Environ. 88: 142-150.

Appendix 1. Further demonstration of range bias with an adjustment procedure (see section 3.1)

Lulla and Bennett [1981] examined range bias. Forty test participants were assigned to two different range conditions; twenty were exposed to a potential luminance range of 1 to 300,000 foot Lamberts (fL) and the remaining twenty were exposed to a potential luminance range of 1 to 30,000 fL. For each range, participants were exposed to six different conditions (labeled A to F, in which the number and subtended size of the glare sources were varied) and used luminance adjustment to set the BCD. The mean settings, determined from the individual setting reported by Lulla and Bennett, are shown in Table A1. The BCD occurred at higher luminance when using the higher of the two luminance ranges. Using bootstrapped Welch's (unequal variances) *t*-tests to analyze the data suggests the differences to be statistically significant in five out of the six conditions across the two ranges, with effect sizes that suggest meaningful practical significance in all conditions ($r \geq 0.20$).

Table A1. Bootstrapped Welch's *t*-tests comparing BCD luminances for two luminance ranges in six glare source configurations [Lulla and Bennett 1981]. Note: luminances reported here in foot-Lamberts following the original work.

Glare Condition	High range 1 to 300,000 fL Mean (SD)$_0$	Low range 1 to 30,000 fL Mean (SD)$_1$	$\Delta M_{(o-1)}^{NHST}$	df	*t*	*r*
A	36,793 (54,826)	3,852 (4,588)	32,942*	19.27	2.68	0.52
B	24,972 (27,227)	3,722 (3,870)	21,250**	19.78	3.46	0.61
C	26,308 (35,087)	5,511 (5,707)	20,797*	20.05	2.62	0.48
D	41,911 (61,677)	4,379 (3,271)	37,532*	19.11	2.72	0.53
E	27,991 (42,416)	4,345 (3,206)	23,646*	19.22	2.49	0.49
F	3,294 (8,730)	997 (1,500)	2,297 n.s.	20.12	1.16	0.25

*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$; n.s. = not significant

Effect size: $r < 0.20$= negligible; $0.20 \leq r < 0.50$= small; $0.50 \leq r < 0.80$= moderate; $r \geq 0.80$= large [Ferguson, 2009].

Appendix 2. Procedural steps required to promote credible data.

Stage of work	Requirements
Planning	• Establish in advance the procedure, the sample size, and the dependent, independent and control variables, and the method of statistical analysis. • Consider pre-registering these decisions. • Report all design decisions, regardless of whether they may seem arbitrary (see research degrees of freedom: Wicherts et al 2016). • Report the experimental apparatus in detail. See CIE 2014 and Veitch et al 2019 for examples of what should be included.
Procedure	• Randomize the order in which different scenes are evaluated and in which experimental variations are employed (e.g. the use of high and low anchors in adjustment trials) • Use null condition trials, extreme conditions and counterbalancing (see Table 13). • If using paired comparisons of a discrete set of independent variables, use all-possible pairs rather than comparing each against a single reference. • Use different stimulus ranges to examine the prevalence of range bias. • For procedures with active interaction (adjustment, matching) use high and low anchors (initial luminance settings); use the mean result of the two trials as best estimate. • For category rating procedures, consider carefully the number of response categories. Do not assume that a previously used response scale has validity simply because of previous use of because it has a name. Ask first whether there is any discomfort (yes/no) and evaluate the degree of discomfort only for those scenes which give discomfort. • If a matching procedure is used, counterbalance application of luminance adjustment to both stimuli in each pair. • Evaluate the same set of stimuli using more than one procedure (a converging operations approach: see text).
Analysis	• Report any results which were eliminated, the reason for elimination (e.g. extreme values), and analysis of the data with those values retained. • Report the mean and standard deviation (or median and inter-quartile range for data drawn from non-normal distributions). • Upload the raw data (e.g. as supplementary information with journal publications) to enable independent analyses by others. • Report the results of null condition trials. • Report the findings of evaluations, regardless of whether or not a significant effect was found. • Report effect sizes in addition to significance [Durlak 2009, Nuzzo 2014]. • State which statistical methods were used and whether assumptions were verified. • For correlation analyses, report the sample of observations on which a correlation coefficient is based. • Consider that the results are relative and cannot be used to establish an absolute threshold.