Comparison of questionnaire items for discomfort glare studies in daylit spaces

Geraldine Quek PhD, **Sneha Jain** PhD, **Caroline Karmann** PhD, **Clotilde Pierson** PhD, **Jan Wienold** PhD, **Marilyne Andersen** PhD

Laboratory of Integrated Performance in Design (LIPID), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Received 5 November 2021; Revised 20 June 2023; Accepted: 8 September 2023

When studying discomfort glare, researchers tend to rely on a single questionnaire item to obtain user evaluations. It is unclear whether the choice of questionnaire item affects the distribution of user responses and leads to inconsistencies between studies. This study aims to investigate if different glare questionnaire items yield similar distributions of user discomfort in daylit environments. We conducted a comparative study of selected questionnaire items from previous glare experiments, testing them in three independent user studies with different lighting conditions and glare stimuli. We compared the resulting outputs across questionnaire items with 540 data points from 149 participants. Results indicated that ordinal questionnaire outputs show strong correlations ($0.68 < \rho < 0.85$), high internal reliability ($\alpha = 0.93$), and captured the same latent construct. Binary questionnaire items reflected different glare thresholds but still correlated well with ordinal items. The construct validity of tested questionnaire items was confirmed through responses to an open-ended question. These findings suggest that the tested questionnaire items may be used for category rating-type discomfort glare evaluations and consistently capture the same construct.

Citation:

Quek, G., Jain, S., Karmann, C., Pierson, C., Wienold, J., & Andersen, M. (2023). Comparison of questionnaire items for discomfort glare studies in daylit spaces. Lighting Research & Technology, 14771535231203564. https://doi.org/10.1177/14771535231203564

Address for correspondence:
Geraldine Quek, Architecture and Sustainable Design, Singapore University of Technology and Design, 8 Somapah Rd, Singapore 487372
Email: geraldine_quek@sutd.edu.sg

1 Introduction

In the fields of health, social and behavioural research, scales are "collections of items combined into a composite score intended to reveal levels of theoretical variables not readily observable by direct means".[1] They represent one or more latent constructs, allowing us to assess and capture a behaviour, an action, or a feeling that cannot be captured in a single variable or measured by other direct means.[2] Scales are typically composed of multiple questionnaire items that measure an underlying latent construct and protect against the influence of culture, biases, and item order, resulting in better validity in scientific investigations, but are also sometimes composed of a single questionnaire item.[3,4] Questionnaire items that make up a scale typically include a question and a response scale of response items and sometimes include definitions of keywords to aid comprehension. Examples of multi-item scales containing multiple questionnaire items are the Activity Inventory (AI) to assess low vision rehabilitation outcomes[5], the Perceived Stress Scale (PSS) to measure stress perceptions[6–8] and the Epworth Sleepiness Scale (ESS) to assess daytime sleepiness levels[9]. Discrete single-item scales such as the Visual Analogue Scale (VAS) and Verbal Numeric Rating Scale (VNRS) are used to measure the amount of pain in medical practice[10,11], or the Karolinska Sleepiness Scale (KSS) used to assess a person's level of sleepiness or drowsiness in sleep research.[12]

The avoidance of discomfort glare is one of the key factors to consider when designing indoor spaces with high comfort levels for occupants and is also acknowledged in existing standards (EN12464, EN17037).[13,14] The International Commission on Illumination (CIE) defines discomfort glare as a "condition of vision in which there is discomfort without necessarily impairing the vision of objects".[15] In discomfort glare research, whether conducted in controlled laboratory settings or field studies, researchers typically collect subjective responses from participants using questionnaire items, which include the question, the response scale, and the format in which they are presented, with or without definitions.[16] For discomfort glare, there is no consensus of a questionnaire item or scale as of time of publication. When studying the extent of discomfort glare effects, researchers tend to rely on a single questionnaire item for user evaluations of the degree of discomfort glare perceived as the main underlying latent construct. There have been numerous suggestions and criticisms about the wording of the question, the response scale and its items, the format in which they are presented, and whether accompanying definitions are included.[16–19] It is still unknown whether the type of glare questionnaire item chosen for user studies influences the response distribution due to a lack of user studies to investigate this. If there is an influence, the results of studies that use different questionnaire items may differ.

Therefore, in this paper, we aim to determine whether a selection of glare questionnaire items captures similar distributions of user discomfort from glare as resulting outputs in rating-type experiments conducted in daylit environments. To that end, a comparison study is carried out, which entails selecting which questionnaire items to test, then testing them in a randomized order in glare evaluations in three user studies of varied lighting conditions, and finally comparing their resulting outputs across questionnaire items. The findings are then presented using descriptive analysis methods and psychometric statistics, as well as tests of association, reliability, and dimensionality. The latent construct that the tested questionnaire items solicit is also checked for validity.

2 Background

Numerous critiques of some of the most commonly used questionnaire items in various glare studies have been published, and some previous literature has offered several pointers such as the comprehensibility and ordering of verbal descriptors in the response scale used.[17,18] The critiques emphasized the inconsistencies of glare questionnaire items, reflected on

whether meaningful results can be obtained through questionnaire items, and discussed their advantages and disadvantages. In ideal cases, questionnaire items should meet a few key requirements, such as being easily understood by the participant and avoiding asking about a past experience rather than the current situation they are exposed to.[11] It should only ask one question at a time, and should not mix concepts such as satisfaction, acceptance, and discomfort within the same questionnaire item, and response items should also be clear and have straightforward descriptors. Some researchers also suggested that a “no glare” or null option be included in the response items,[20,21] so that participants are not forced to report glare when they do not perceive any. However, these pointers for designing glare questionnaires, in general, may sometimes only apply to category-rating test procedures but not adjustment-type procedures which require the active interaction of the participant with the visual scene.[16] Category-rating test procedures typically expose participants to one scene or stimuli at a time and ask participants to rate pre-defined variables through questionnaire items, while adjustment-type procedures usually expose participants to a starting scene or stimuli and ask them to adjust the parameters of the stimuli to fit described levels of stimuli such as the multiple criterion method[22].

Other suggestions include giving the participant a layman's definition of the key variable in question, presenting response items in a logical and relevant order, and including a “don’t know” option to capture participants who do not understand or know what they are perceiving.[18] A balanced number of response items on each side of the neutral point should also be maintained for bipolar or semantic differential scales.[18,23,24] If numbers are used in addition to verbal descriptors as response items, they should correspond in increasing order of intensity, for example, "0" should correspond to "Not at all" and "10" should correspond to "Very much" on a scale of 0 to 10.[25] Additionally, instead of asking about the degree of discomfort from glare, some researchers proposed using a positively worded statement in conjunction with a Likert agreement scale to pose the question more optimistically.[20] Other suggestions include a fixed equal distance between response items, language consistency, and a greater number of response items on the scale than glare stimuli levels for sufficient resolution.[26] A recent proposal for questionnaire standardization is a two-step skip-sequencing method for evaluating discomfort from glare,[19] suggesting first to ask the participant if they are experiencing discomfort from glare. If the participant answers yes, then they are asked a second question on a 6-point scale labelled 1 (Very small amount) to 6 (Very large amount). If the participant answers no, the second question will be skipped.

However, these critiques on questionnaire items have not been studied using objective measures so far, and multiple variations of questionnaire item types have been used to solicit evaluations on the degree of glare in past user studies, from which discomfort glare prediction models have been developed. The multiple-criterion method for subjective glare appraisals in an adjustment-type experiment procedure was first proposed by Hopkinson[27] and participants were asked to adjust a lighting variable based on a criterion of discomfort glare on the multiple criterion scale.[22] The 4-point multiple criterion scale originally published in 1940 consisted of four degrees of discomfort glare as follows: “A: Just intolerable, B: Just uncomfortable, C: Satisfactory, and D: Just not perceptible”. Petherbridge and Hopkinson developed the British Research Station (BRS) glare index[28] to describe discomfort glare from electric lighting fittings in 1950, using a semantic variation of this response scale with C and D changed to “C: Just acceptable, D: Just imperceptible”. MacGowan then posed the question of whether these response items might have been better understood at the time they were proposed to trained observers rather than new observers.[29] In 1960, Hopkinson and Bradley developed the ‘Cornell formula’ which used another semantic variation of the scale with criterion D changed to “D: Just perceptible” to study discomfort glare from large windows simulated by electric lighting apparatus, using adjustment procedures.[30,31] In 1962, the Illuminating Engineering Society (IES) glare index was established by modifying the BRS glare index.[32] For adjustment-type experiment protocols, Kent *et al.* found significant differences in the luminances adjusted

by the participants when criteria on the multiple criterion scale were presented in ascending order, compared to when the presentation order of the criteria was randomized.[33]

Using a category-rating test procedure, Chauvel et al. modified the 'Cornell formula' through user assessments for discomfort glare studies in daylit buildings, asking observers to assess the level of discomfort in the scene presented to them.[34] Their England study used the multiple criterion scale and a five-point response scale in the France study: "1 – not uncomfortable, 2 – slightly uncomfortable, 3 – rather uncomfortable, 4 – very uncomfortable, and 5 – extremely uncomfortable".[31] Iwata et al. developed the Glare Sensation Vote (GSV) model in 1992 in daylight conditions with user assessment procedures using the 4-point response scale similar to Hopkinson and Bradley.[35] In 1995, the International Commission on Illumination (CIE) proposed the Unified Glare Rating (UGR)[36] where Sorensen developed UGR[37] using Petherbridge and Hopkinson's dataset. The UGR formula incorporated the IES glare index as well as mathematical corrections proposed by Einhorn for the CIE glare index (CGI).[38,39]

Fisekis et al. used a 7-point response scale similar to the multiple criterion scale without the "Just intolerable" criterion in their user experiments resulting in the modified Daylight Glare Index (DGI_{mod}) and the experimental Unified Glare Rating (UGR_{exp}),[40] adapted for daylight glare from windows. Wienold and Christoffersen developed the Daylight Glare Probability (DGP) in 2006[41] through user studies using a 4-point response scale with "Imperceptible, Noticeable, Disturbing, Intolerable" introduced by Osterhaus and Bailey[42] which is also similar to Hopkinson's multiple criterion scale. Here, the word "just" was omitted in the response scale items as Hopkinson's multiple criterion scale was originally meant for adjustment procedures where the borderlines of comfort and discomfort were pertinent. In 2014, Hirning et al. adapted UGR for daylight conditions in deep open-plan offices, resulting in Unified Glare Probability (UGP).[43] He collected glare responses using a glare indication diagram, which asks participants to indicate on a diagram where a glare source, if any, is in their field of view. Any marking on the glare indication diagram is interpreted as indicating that the participant experienced uncomfortable glare in that scene.

As described, we can observe that beyond the type of test procedure used (adjustment, or category-rating), there have been multiple variations to the questionnaire types and the response scales used by researchers when studying discomfort glare. An overview of variations of questionnaire items that have been used in user studies that culminated in glare model development can be found in Section I of the supplementary material. It is not yet known whether the usage of different questionnaire items in rating-type experiments may produce varying glare response results and could therefore also bias the results, such as in the development of discomfort glare models.

3 Method

As a step to study the consistency of results from different questionnaire items, we want to investigate whether the choice of questionnaire items for user studies affects the distribution of glare responses from participants. To develop the methodology, we referred to best practices recommended in the psychometry field for scale development.[44] However, in this case, instead of scale development, we are looking to compare the outputs of several glare questionnaire items. Hence, we omit factor extraction since we are only interested in one dimension (or factor) which is the extent of discomfort glare experienced if any. As a result, we chose four steps for the relevance of comparing questionnaire items - we use a process of item generation, survey administration, and tests of dimensionality and reliability in our workflow as shown in Figure 1.

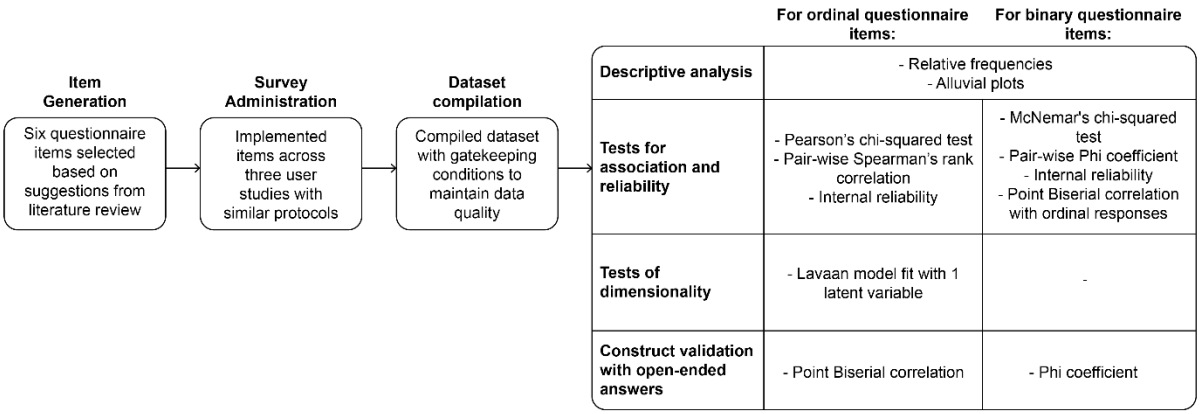


Figure 1 Methodological workflow to analyze questionnaire items for discomfort glare studies.

To execute this four-step process toward comparing the output of questionnaire items, we tested the selected questionnaire items across three independent user studies that evaluated discomfort glare from daylight each with unique research objectives and therefore covering a wide range of glare stimuli from daylight. These three user studies resulted in a dataset that covers a wide range of vertical illuminances (from 216 to 7300 lx). This procedure was implemented as a collaborative study in which the authors, who conducted the three user studies, decided on the relevant questionnaire items they wanted to compare, and coordinated to administer them simultaneously across the three user studies. We, therefore, administered six chosen questionnaire items on discomfort glare perception in English in these three parallel user studies in indoor daylit environments and used similar experimental protocols. In the following sections, we go into detail about the questionnaire items selected for comparison, the experimental protocol for survey administration, and data compilation. Finally, a construct validation of the questionnaire outputs is carried out using open-ended answers to a general question that was asked to the participants before all six questionnaires were first seen by them.

3.1. Selecting questionnaire items

This section describes how the questionnaire items were generated and assessed for viability before being administered in this comparison study. To cover a wide range of questionnaire types used in daylight glare research and related indoor environmental quality (IEQ) studies, "Binary-YesNo", "OsterhausBailey-4point", "Likert-4point", "Interval-0-10", "Comfort-agreement", and "Glare-indication-diagram" were chosen and are shown in Figure 2. These six questionnaire items were chosen for the following reasons.

First, "OsterhausBailey-4point", "Binary-YesNo" and "Glare-indication-diagram" were chosen as they are commonly used in the field[42,43,45,46], while the other three questionnaire items were chosen to add variety to the selection while being viable candidates according to the suggestions from the literature. The "Glare-indication-diagram" which was not a typical questionnaire item and requires participants to mark on a diagram if they experience discomfort glare, was still included because it was used to develop a glare model from surveys in open plan offices, namely Unified Glare Probability (UGP).[43]. Similar to Hirning *et al.*, we interpreted the results of the Glare-indication-diagram by converting any marking on the diagram to "Yes" and none to "No". "Likert-4point" was chosen as it has a simple Likert format, and has easy-to-understand, incremental response items. The main difference to the 4-point scale suggested in ISO10551:2019[47] was that we used the term “Moderately” in “Likert-4point” as the third response item rather than only “Discomfort” in order to use clearly incremental response items.[1] "Interval-0-10" is an 11-point numerical scale with labels ranging from "Not at all" to "Very much" at the extremes, similar to "Likert-4point". "Comfort-agreement" was chosen to serve as a positively worded question to the selection.[20]. Last, these questionnaire

items were chosen specifically for applicability in rating-type studies, in which participants are exposed to a lighting condition and asked to rate their discomfort by answering a questionnaire. For example, we believe Hopkinson's questionnaire multiple criterion scale[22] is inappropriate for our user studies because it was originally developed to be used in adjustment-type studies and not rating-type studies. For the participants to understand the definition of glare without knowing technical terms, a layman's definition of glare was derived by referencing the CIE's definition of glare[15]: “Glare is the sensation of visual discomfort caused by differences between light and dark areas, or by excessive brightness in your field of view.” This was shown to the participants alongside each questionnaire item[2].

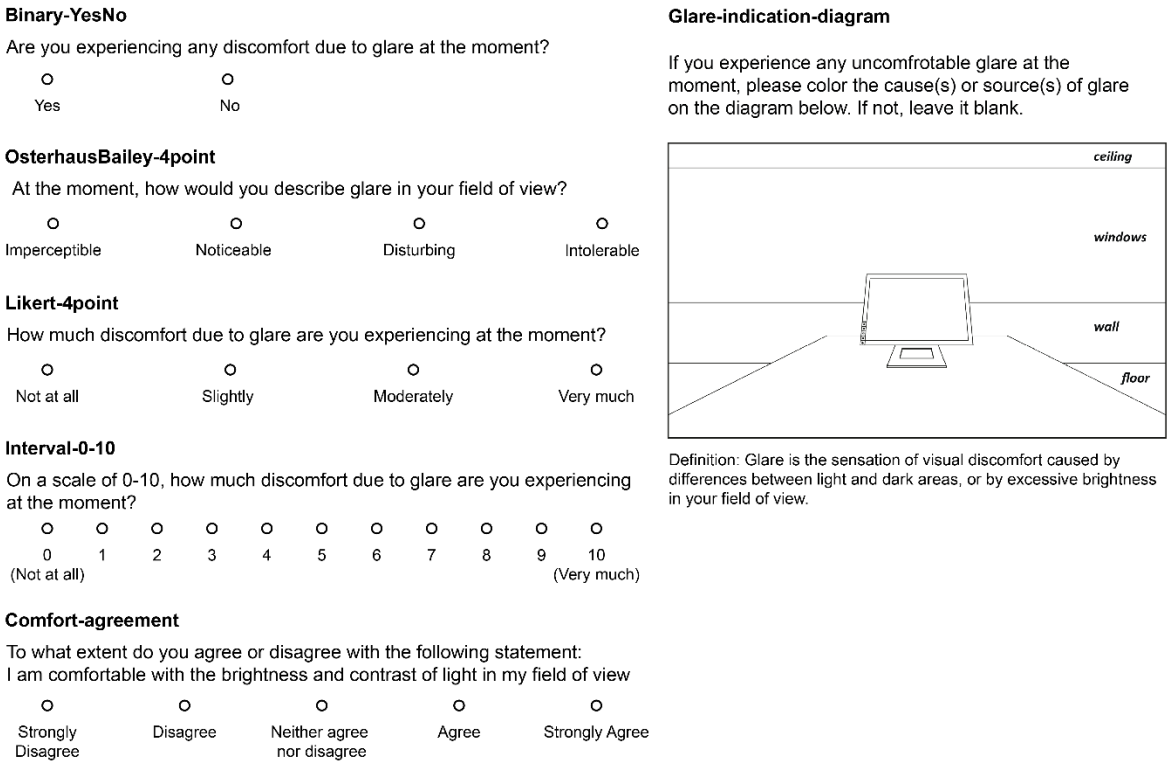


Figure 2 Six questionnaire items were selected for comparison and implemented in a random order in three user studies.

Additionally, to ensure that the questionnaire items selected for this study are viable, we compared them to a compiled list of suggestions in past literature as shown in Table 1. These suggestions were argued in the respective cited publications for more meaningful and informative glare questionnaire designs. If the pointer does not apply to the questionnaire item, it is simply not evaluated and labelled "N.A.". As can be seen, it is nearly impossible for any questionnaire item to satisfy all suggestions at the same time. For example, the nature of ordinal response scales is that the orders of the values are known, but the distance between them is not, and hence will not meet the requirement of "equal distance between items on response scale". Similarly, the pointer suggesting a positively phrased question can only be fulfilled if the question does not inquire about the users' perception of discomfort. As one may observe, all six questionnaire items selected fulfilled at least half of the applicable suggestions. Hence, we went ahead with administering all of them in a random order of presentation in all three user studies.

Table 1 Six glare questionnaire items were evaluated alongside a list of suggested pointers from previous literature regarding questionnaire design.

Suggestions from past literature	Reference	"Binary-YesNo"	"OsterhausBailey-4point"	"Likert-4point"	"Interval-0-10"	"Comfort-agreement"	"Glare-indication-diagram"
Comprehensible question	[48]	1	1	1	1	1	1
Clear descriptors on the scale	[18,23]	1	0	1	1	1	N.A.
No mixing of concepts	[18,23,26]	1	0	1	1	1	1
Include null option	[20,21]	1	1	1	1	1	1
Avoid asking about a past experience	[48]	1	1	1	1	1	1
Layman's definition of key term in question	[18]	1	1	1	1	N.A.	1
Response items in relevant order	[18]	N.A.	1	1	1	1	N.A.
Include "Don't know" option	[18]	0	0	0	0	0	0
Balanced response items on each side of the neutral point (for bipolar/semantic differential scales)	[18,23,24]	N.A.	N.A.	N.A.	N.A.	1	N.A.
Correspondence between number and verbal descriptors		N.A.	N.A.	N.A.	1	N.A.	N.A.
Include a positively worded question	[25]	0	0	0	0	1	0
Equal distance between items on response scale		N.A.	0	0	1	0	N.A.
Language consistency (English, or validated translation)	[20]	1	1	1	1	1	1
Scale representing stimuli range	[26]	N.A.	1	1	1	1	N.A.

To check for the construct validity of the questionnaire items, a generic open-ended question, Binary-Open, was asked to participants at the beginning of each evaluation for every scene presented: “Is there anything about the physical environment that disturbs you at this moment? (Answer "No", if you are not disturbed by anything.)”. We processed their answers by categorizing them into two bins, ‘Yes’ and ‘No’. Any mention of glare, bright sources of light caused by the sun, façade, or reflections is categorized as ‘Yes’, and if none of these are mentioned, the answers were categorized as ‘No’. Only the first data point from every participant was used for this analysis, such that they would not have been exposed to glare questionnaire items before answering the open-ended question and that Binary-Open would have been presented for the first time to them. This resulted in a total of 137 data points for validating the latent construct.

3.2. Survey administration and data compilation

Following their selection, these six questionnaire items were implemented concurrently in three different user studies by the authors, each with a different setup but producing glare stimuli from daylight and all following a similar experimental protocol in office-like conditions. All three user studies were held in the same test facility, DEMONA (East), on the EPFL campus in Lausanne, Switzerland, between September 2020 to October 2021. DEMONA (East) is a single-room facility approximately 3 by 7 meters in dimension and has thermal room conditioning capabilities with radiative walls for heating and cooling. As the main focus variable in all three setups was discomfort glare, other indoor environmental aspects such as thermal and acoustic qualities were maintained comfortable as much as possible and participants were also asked about their perceived indoor environment other than their visual comfort. Their specific data collection periods are shown in Table 2. The first study[49] focused on contrast-dominant glare in low photopic ranges with a mean vertical illuminance of 759 lx, and the second focused on discomfort glare through shading fabric with contrast-dominant glare and high photopic range with a mean vertical illuminance of 1834 lx.[50] The third study[51] focused on discomfort glare with direct sun in the field of view (FOV) through low transmittance, color-neutral glazing with a mean vertical illuminance of 3129 lx. To investigate the influence of questionnaire items on glare responses, we combined the three collected sets of data into one consolidated dataset which embodied a large range of daylighting conditions where the selected questionnaire items were administered. We verified the sample size using the G*Power 3.1.9.6 calculation tool for repeated measures, between- factor testing three groups and four measures, assuming an effect size of 0.30, an alpha error probability of 0.05 and a power of 0.95. This calculation yielded a required sample size of 111 participants. Following the data filtering protocol described below, 63 data points were removed from the original sample, leaving 540 data points from 149 distinct participants in the compiled dataset for analysis. This exceeds the required sample size as calculated.

The three user studies followed a category-rating procedure, rating four luminous scenes in a randomized order. Participants were exposed to one lighting scene at each time and asked to complete a typing task for at least five minutes to allow their eyes to adjust to the lit environment. After the typing task, they were then asked to complete an on-screen survey administered by the online survey platform Alchemer. The generic open-ended question, Binary-Open, was asked to each participant first. After this, all six questionnaire items in all three studies were administered to each participant in a randomized order, in a one-question-per-page format in English along with a layman's definition of glare as mentioned earlier. To test the understanding of the semantics of the response items as-is, we did not give additional definitions or explanations of the response items on the scale to the participants, even though for example, the original implementation of the OsterhausBailey-4point questionnaire item gave time-based explanations.[42]

Table 2 Breakdown of the number of participants and data points included in each study, when they were conducted, and their corresponding ranges of vertical illuminances.

Dataset	Data collection period	N	n	Vertical illuminance range (lx) Total	Vertical illuminance range (lx) Scene 1 (Lowest)	Vertical illuminance range (lx) Scene 2	Vertical illuminance range (lx) Scene 3	Vertical illuminance range (lx) Scene 4 (Highest)
1. User study in contrast dominant discomfort glare in dim daylit conditions [49]	September to October 2020, March to April 2021	62	234	Min: 216 Mean: 759 Max: 2080 SD: 370	Scene: 1_panel-low Min: 216 Mean: 423 Max: 700 SD: 112	Scene: 1-panel_high Min: 334 Mean: 642 Max: 1000 SD: 166	Scene: 2-panel_low Min: 355 Mean: 770 Max: 1230 SD: 219	Scene: 2-panel_high Min: 704 Mean: 1183 Max: 2080 SD: 378
2. User study of discomfort glare from shading fabrics [50]	December 2020 to March 2021, October 2021	32	109	Min: 290 Mean: 1834 Max: 4960 SD: 1348	Scene: B2 Min: 290 Mean: 718 Max: 1010 SD: 209	Scene: B1 Min: 260 Mean:1136 Max: 3950 SD: 633	Scene: G2 Min: 430 Mean: 1613 Max: 4520 SD: 762	Scene: B7 Min: 620 Mean: 3575 Max: 4960 SD: 1076
3. User study with direct sun as a glare source (only data from color-neutral glazing are used) [52]	October 2020 to March 2021	55	200	Min: 830 Mean: 3128 Max: 7300 SD: 1341	Scene: N1 Min: 830 Mean: 2274 Max: 5520 SD: 1173	Scene: N2 Min: 940 Mean: 2394 Max: 4420 SD: 645	Scene: N3 Min: 1070 Mean: 3496 Max: 5830 SD: 947	Scene: N4 Min: 1630 Mean: 4502 Max: 7300 SD: 1140
	Total	149	540					

Table 2 contains a breakdown of the three datasets and their corresponding ranges of vertical illuminances. The first study varied the luminance and size of glare sources created using different combinations of diffuse films and low-transmittance color-neutral films attached to the window. The second study varied fabric blinds with different openness factors with direct sun in the field of view, while the third study varied the luminance of the direct sun disk with color-neutral films of different low transmittances. Figure 3 shows the four scenes of each of the studies where the questionnaire items were administered. In the second study, there were a total of five scenes evaluated where the participant was asked to adjust the blinds in the fifth scene. However, we only considered the first four evaluated scenes of each user study when compiling the sets of collected data from the three user studies. There were no other critical differences in their experimental protocol other than the luminous ranges focused on by each study.

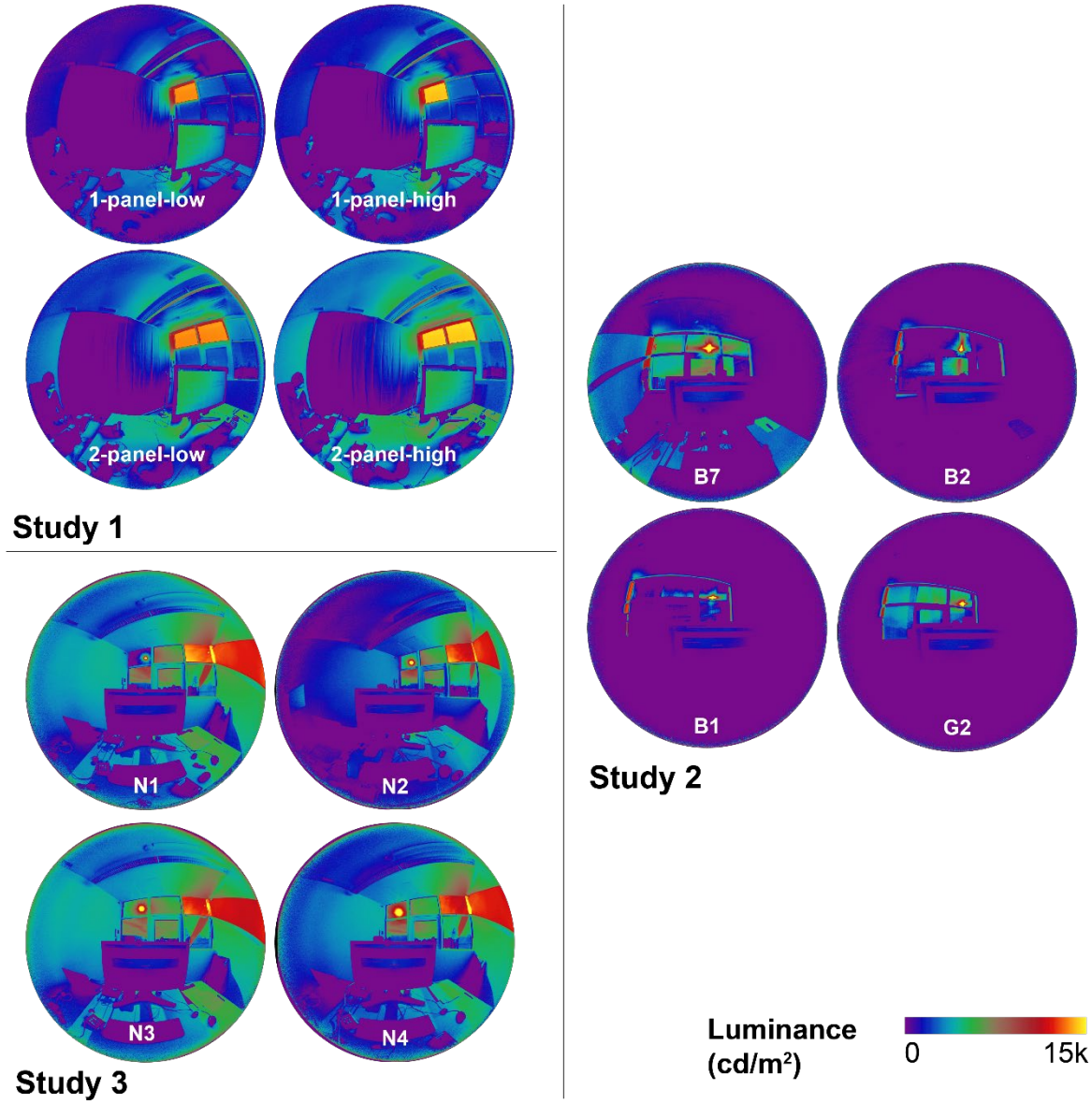


Figure 3 Example HDR images of the four scenes evaluated in each of the three studies where the questionnaire items were administered. (Available in colour in online version)

Participants in all three studies were also recruited such that they have at least C1 English proficiency according to the Common European Framework of Reference for Languages

(CEFR), are not studying or working in the built environment sector, do not have any eye related pathologies, have normal color vision, are between the ages of 18 and 35, and present in good health on the day of their scheduled participation in the user studies. The three user studies, in which all six questionnaire items were administered, were conducted in Lausanne, Switzerland, with each participant engaging only once, in one of the three studies. There were no additional recruitment criteria based on cultural background because it had previously been discovered that cultural background has no significant effect on glare perception.[53] Each data point consists of one participant's answers to all six questionnaire items to one lighting scene, which is measured using HDR images and illuminance meters, as well as weather conditions measured either by continuous vertical illuminance (E_v) measurements indoors from the participants' point of view, or global horizontal irradiance (GHI) measurements outdoors on a rooftop of a nearby building. We filtered the data in the following way to ensure that the lighting conditions remained stable throughout the survey duration so that for an evaluated scene, all six questionnaire items were answered with the fewest variations in lighting conditions due to fluctuating weather conditions.

Where continuous vertical illuminance was available in the first study (as derived from continuous captures of HDR images indoors (every 15 seconds)), we removed data points above a 25% deviation of vertical illuminance, $(E_{v,\,max} - E_{v,\,min})/(E_{v,\,mean})$ for the duration of the participant's exposure to the scene (from typing task to the end of the survey). The 25% threshold was used for gatekeeping criteria for weather stability used in the field.[53,54] For the second and third studies, we removed data points where the GHI deviated more than 25% $(GHI_{max} - GHI_{min})/(GHI_{mean})$, as measured by an on-site outdoor pyranometer (every 1 second). We exceptionally accepted a few more data points where the E_v or GHI deviation was higher than 25% during the typing task period but not during the survey response period.

4. Results

4.1. Descriptive analysis

In this section, we use the compiled dataset to analyse how participants responded to the six questionnaire items. Here, we ran descriptive analyses using stacked bar charts to describe the relative frequencies of each response item and paired alluvial diagrams to illustrate the flow of user responses. We chose to use alluvial plots as they represent the flow of data from one state to another, and flow lines represent the percentage of respondents, which are typically colored by the variables of the first state. We can see how participants respond to one questionnaire item after another while also describing the percentages of responses for each response item. Note that the order of the questionnaire items in the alluvial plots does not reflect the order in which they were asked to participants because the questionnaire items were presented to them in a randomized order each time.

The stacked bar chart in Figure 4 depicts how the 149 participants responded to the six questionnaire items as well as Binary-Open in 540 evaluated scenes based on the compiled dataset. Comparing, "Binary-YesNo" and "Glare-indication-diagram", 48% of participants answered “Yes” to Binary-YesNo while 65% indicated a glare source on the Glare-indication-diagram. On the OsterhausBailey-4point response scale, approximately 70% of participants reported noticeable glare and above, while 25% reported disturbing glare and above. The response distribution of the other questionnaire with four response items, Likert-4point, was similar, with 68% of participants reporting slight glare and above and 27% of participants reporting moderate glare and above. The Interval-0-10 questionnaire item produced a higher resolution due to its 11 response items. On the Interval-0-10 questionnaire, 28% of participants rated 6 or higher (beyond the middle point of 5). For the positively worded Comfort-agreement questionnaire, 30% disagreed that the brightness and contrast in their field of view were comfortable.

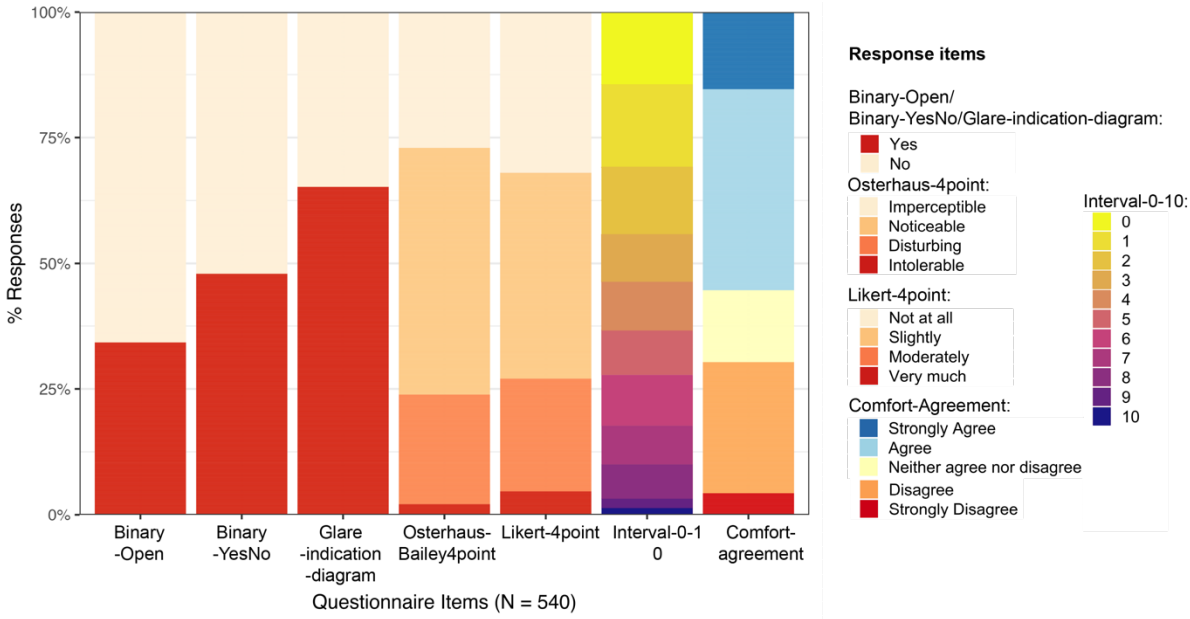


Figure 4 Relative frequencies of responses across the six questionnaire items as well as Binary-Open, which will be used for checking for construct validity. (Available in colour in online version)

For the binary questionnaire items, a pairwise alluvial plot of how participants answered the Binary-YesNo question versus the Glare-indication-diagram is shown in Figure 5, revealing a difference in response distribution. The flow lines connecting "No" on the Binary-YesNo questionnaire item to "Yes" on the Glare-indication-diagram demonstrate the participants (17%) who answered "No" to the Binary-YesNo question but also indicated a glare source on the diagram. Such differences are sufficient to affect derived discomfort thresholds such as in DGP. The full set of possible pairwise alluvial plots between the six questionnaire item outputs can be found in the supplementary material. Answers from the open-ended Binary-Open (only using the first evaluated scene from each participant, n=137) will be used to check for construct validity later in the analysis.

Furthermore, the percentage of participants who reported glare on the Glare-indication-diagram corresponds to the beginning of "slightly" and "noticeable" responses on the OsterhausBailey-4point and Likert-4point questionnaire items, respectively, as shown in Figure 4. Meanwhile, the positive responses on Binary-YesNo correspond to somewhere in the middle of the "slightly" and "noticeable" responses as shown in Figure 6. Although both questionnaire items have binary output, this phenomenon could imply that the answers to these two questions correspond to different levels of discomfort, with the Glare-indication-diagram corresponding to noticeable discomfort and the Binary-YesNo corresponding to somewhere between noticeable and disturbing discomfort. In other words, positive responses from Binary-YesNo refer to a higher threshold between "noticeable" and "disturbing" thresholds, while positive responses from Glare-indication-diagram refer to a lower threshold nearer to the "imperceptible" to "noticeable" threshold.

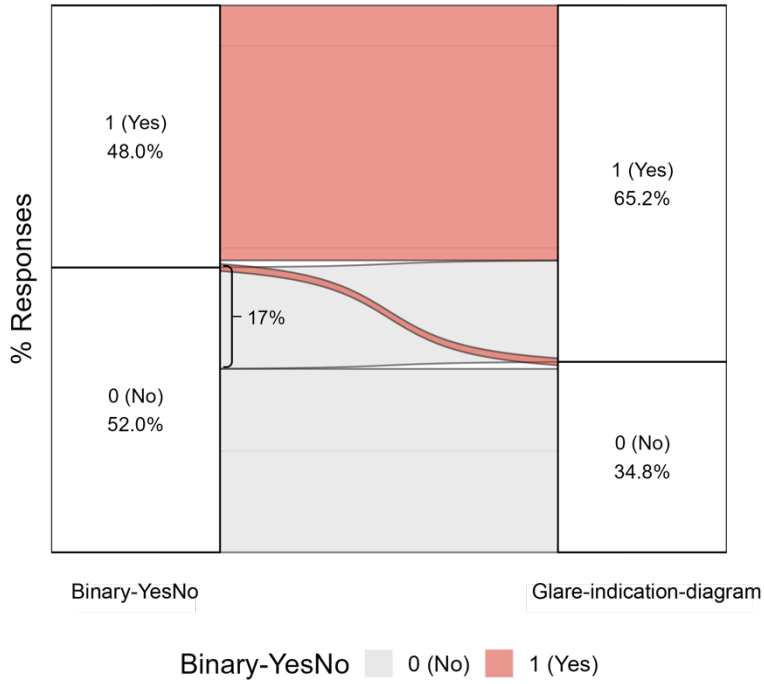
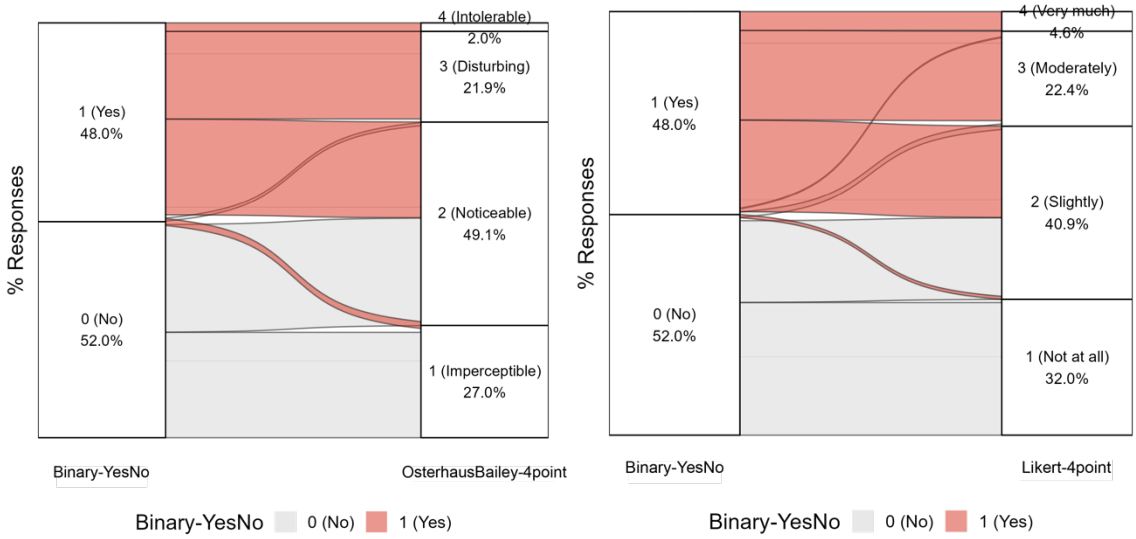


Figure 5 Flow of participants' responses between Binary-YesNo and Glare-indication-diagram.

Figure 6 depicts the pair-wise alluvial plots between Binary-YesNo and OsterhausBailey-4point, as well as of Binary-YesNo and Likert-4point respectively. Both questionnaire items have four response items and are compared to the binary questionnaire item distribution, Binary-YesNo. Surprisingly, participants who answered "No" to Binary-YesNo did not all answer "Imperceptible" or "Not at all" in both cases. 50% of those who answered "No" reported it was "Noticeable" on the OsterhausBailey-4point, and 38% reported it was "Slightly" on the Likert-4point. It can be seen that "No" in a binary questionnaire item does not always correspond to an absolute null response in other questionnaire items with a higher resolution (more than two response items). This phenomenon also occurs between Binary-YesNo and Interval-0-10 with a response scale of 11 points. Figure 7 shows the pairwise alluvial plot between them, which shows that two-thirds of the total participants who answered "No" to the binary question answered more than "0" on the Interval-0-10 response scale.



(a) Binary-YesNo and OsterhausBailey-4point (b) Binary-YesNo and Likert-4point

Figure 6 Flow of participants' responses between Binary-YesNo and between OsterhausBailey-4point and Binary-YesNo and Likert-4point.

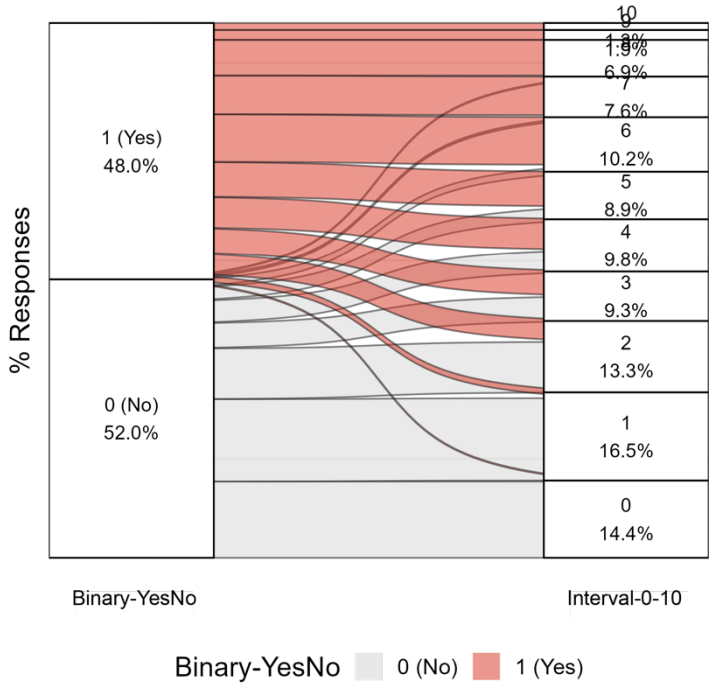


Figure 7 Flow of participants' responses between Binary-YesNo and Interval-0-10.

There is some indication that participants responded similarly between the two questionnaire items that contain four response items, OsterhausBailey-4point and Likert-4point. Interestingly, as shown by the orange flow lines in the pair-wise alluvial plot in Figure 8 coloured by the response scale of the OsterhausBailey-4point questionnaire item, participants who answered "Noticeable" to the OsterhausBailey-4point also answered "Moderately" and "Not at all" to the Likert-4point. However, overall, this observation implies that the four response items between the two questionnaire items generally correspond to each other. In the following section, we will delve deeper into the psychometric analysis that statistically confirms this indication.

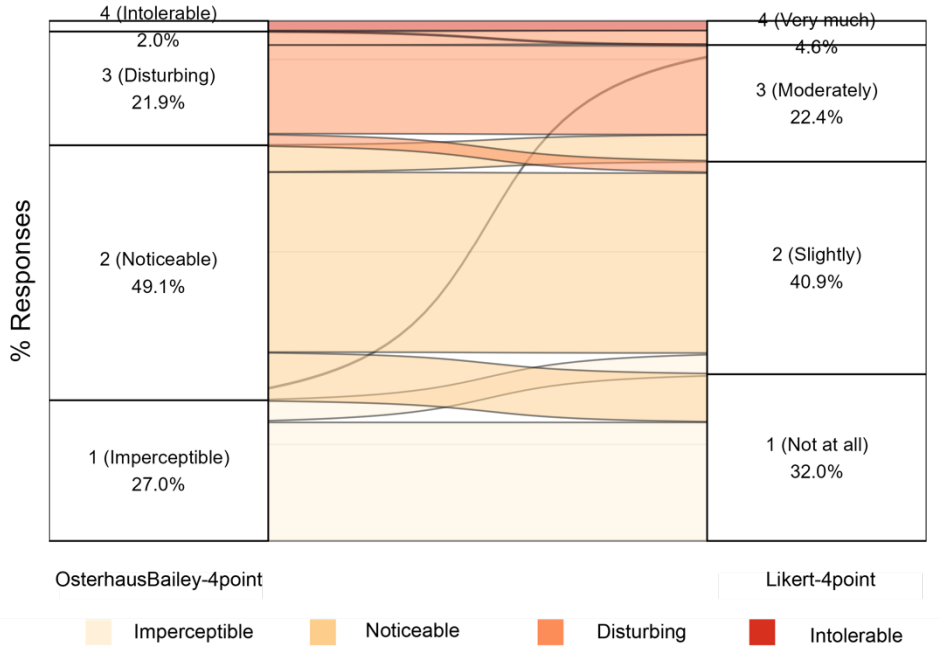


Figure 8 Flow of participants' responses between OsterhausBailey-4point and Likert-4point.

4.2. Psychometric analyses

To confirm the results of the descriptive analyses in the previous section, we will use a selection of statistical methods to determine if there is sufficient evidence that the response outputs of the questionnaire items are contradictory.

The following statistical tests were performed to test for association and reliability for the four ordinal questionnaire items with more than two response items, namely "OsterhausBailey-4point", "Likert-4point", "Interval-0-10" and "Comfort-agreement". The Pearson chi-squared test was conducted to test the relationship between the outputs of the questionnaire items. Spearman rank correlation was then calculated to check the correlational strength between the ordinal data outputs. To assess the internal reliability, Cronbach's α,[55] McDonald's omega estimate,[56] Guttman's Lambda 6 (G6), and Explained Common Variance (ECV) were calculated. Then, a test of dimensionality was conducted where the fit of a uni-dimensional Lavaan model with one latent variable was used to confirm whether the four ordinal questionnaire items point to a single variable – in this case, the amount of discomfort due to glare. A robust Weighted Least Square Mean and Variance adjusted (WLSMV) estimator for ordinal non-normally distributed variables was used for the Lavaan model.

Another set of statistical tests designed for dichotomous data was used to determine whether the two binary questionnaire items, "Binary-YesNo" and "Glare-indication-diagram" produced similar results in terms of association and reliability. First, McNemar's chi-squared test is performed to determine whether there were any significant differences in frequency between their outputs. To assess internal reliability, the Kuder-Richardson Formula 20 (KR20), which is similar to Cronbach's α but for dichotomous data, was calculated. To assess the correlation between the two outputs, the Phi coefficient is calculated instead of Spearman's rank correlation. There were no dimensionality tests performed between the two binary questionnaire items, but a Point biserial correlational test with each of the ordinal questionnaire output were then run to check if the binary questionnaires point to the same latent variable as the ordinal output. The "psych" package (version 1.9.12.31) in R (version

3.6.3) was used to perform reliability, dimensionality, and validity tests after normalizing responses (between 0 and 1) from the six questionnaire items.

The descriptive analyses in the preceding sections show that the responses to the four ordinal questionnaire items seem to agree with each other. To ascertain this, statistical tests on association and internal reliability are presented in this section, with separate sections on ordinal questionnaire items and binary questionnaire items.

4.2.1. Ordinal questionnaire items

First, a Pearson chi-squared test was performed across paired questionnaire items. The null hypothesis is that the questionnaire responses were independent and that no relationship exists between the categorical variables. The results rejected the null hypothesis with sufficient evidence, with all p-values being < 2.2e-16, at a significance value of 0.05 (p-values shown in Section III of the supplementary material). As a result, there is reason to believe that there is a significant relationship between ordinal questionnaire items.

In Table 3, pair-wise Spearman rank correlations ρ between questionnaire responses are shown. The output of ordinal questionnaire items generally shows strong intercorrelations, as the ρ are greater than 0.6. The strongest correlations are found between Interval-0-10 and Likert-4point, with a ρ of 0.85, and the second highest correlation is found between Likert-4point and OsterhausBailey-4point. All pairwise p-values show statistical significance of the Spearman rank correlation ρ values, rejecting the null hypothesis which is that there is zero correlation, as represented by "***" in the table.

Table 3 Spearman rank correlation rhos between Likert-4point, OsterhausBailey-4point, Interval-0-10, and Comfort-agreement questionnaire responses with ordinal data. ρ values show the strength of correlations, such as weak ($\rho < 0.4$), moderate ($0.4 \leqslant \rho < 0.6$), and strong correlations ($\rho > 0.6$).[57] "***" indicates p-value < Bonferroni-corrected significance level of 0.0083 ($\alpha = 0.05/6$) for six comparisons.

Questionnaire item	OsterhausBailey-4point	Interval-0-10	Comfort-agreement
Likert-4point	0.80***	0.85***	0.71***
OsterhausBailey-4point	-	0.79***	0.68***
Interval-0-10	-	-	0.76***

Following the analysis of pair-wise correlations, psychometric statistics testing for the internal reliability of the ordinal questionnaire items was conducted and the results are shown in Table 4. The Cronbach's α, Guttman's Lambda 6 (G6), Omega total, and Explained Common Variance (ECV) are all greater than 0.9, indicating a high level of internal consistency among the four questionnaire items.[58] A Cronbach's α above 0.7 shows acceptable internal reliability but one must keep in mind that α increases with the number of items tested and average item intercorrelation.[59] To this end, we found that the internal reliability does not increase more than 0.93 when any of the questionnaire items are removed from the group as shown in Table 5. This demonstrates that none of the questionnaire items reduces the internal consistency of the four items and that they have overall high consistency with each other.

Table 4 Psychometric statistics for internal reliability for Likert-4point, OsterhausBailey-4point, Interval-0-10, and Comfort-agreement questionnaire responses.

Index	Value
Cronbach's α	0.93
Guttman's Lambda 6 (G6)	0.92
Omega Total	0.94
Explained Common Variance (ECV)	0.95

Table 5 Results of Cronbach's α, if each questionnaire item is removed.

Item dropped	Cronbach's α
Likert-4point	0.90
OsterhausBailey-4point	0.91
Interval-0-10	0.89
Comfort-agreement	0.93

A test for dimensionality was performed for the ordinal questionnaire items to see if the outputs point to a single variable. From the Lavaan model fit with 1 latent variable, the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) are both greater than 0.95 (Table 6), indicating that responses from the four ordinal questionnaire items point to a singular variable.[60] The root mean square error of approximation (RMSEA) is 0.094, greater than 0.08, which has been proposed as a marginally acceptable minimum threshold for other Lavaan model estimators[61] although a specific RMSEA threshold for the Lavaan model specifically with the WLSMV estimator is not established yet.[62] Hence, the confirmation of the Lavaan model fit with 1 latent variable indicates that the four ordinal questionnaire items, OsterhausBailey-4point, Likert-4point, Interval-0-10, and Comfort-agreement, all describe one variable.

Table 6 Lavaan unidimensional model results fit indexes, using the robust weighted least square mean and variance adjusted (WLSMV) estimator.

Index	*Value*
Comparative Fit Index (CFI)	0.989
Tucker-Lewis Index (TLI)	0.967
Root Mean Square Error of Approximation (RMSEA)	0.094

4.2.2. Binary questionnaire items

McNemar's chi-squared test for paired dichotomous data was used to test the association between binary questionnaire items, where the null hypothesis is that the two outcomes are the same. In this case, we refer to the response outputs of Binary-YesNo and Glare-indication-diagram. Using a significance level of 0.05, there is sufficient evidence to reject the null hypothesis, with a chi-squared value of 79.1 and p-value of < 2.2e-16. As a result, the alternative hypothesis (that there is a significant difference between these two outputs) is accepted. This demonstrates that the response distributions from the two binary questionnaire items differ significantly.

Then, instead of Spearman rank correlation, the Phi coefficient is used to determine the strength of association between dichotomous data from the binary questions. According to the Phi test coefficient of 0.65, the correlation between the outcomes of the two ordinal questionnaire items is considered strong according to the same criteria[57] used for Spearman rank correlation ρ. However, the correlation is not as strong as that between OsterhausBailey-

4point and Likert-4point, which both have the same number of response items and have a paired Spearman rank rho of 0.80, as shown previously in Table 3.

The internal reliability between the two binary outputs was tested. The KR20 value is found to be 0.79 which is lower than the Cronbach α of the ordinal questionnaire items. In contrast to Cronbach's α found between the ordinal questionnaire items (which was 0.93), the outputs between binary questionnaire items point to different thresholds of glare. Nevertheless, the outputs from the binary questionnaire items still significantly correlated with that of the ordinal questionnaire items, as shown from Point biserial correlational test results shown in Table 7. This indicates that their output still corresponds well to that of the ordinal questionnaire items.

Table 7 Point biserial correlation ρ between the output of binary questionnaires to that of ordinal questionnaire items. ρ values show the strength of correlations, such as weak ($\rho < 0.4$), moderate ($0.4 \leqslant \rho < 0.6$), and strong correlations ($\rho > 0.6$).[57] "***" indicates p-value < Bonferroni-corrected significance level of 0.000625 ($\alpha = 0.05/8$) for eight comparisons.

Questionnaire item	Likert-4point	OsterhausBailey-4point	Interval-0-10	Comfort-agreement
Binary-YesNo	0.69***	0.64***	0.70***	0.62***
Glare-indication-diagram	0.66***	0.63***	0.63***	0.53***

4.2.3. Construct validity check with an open-ended question

To confirm the construct which the six questionnaire items solicit is indeed about discomfort glare, their outputs were tested for correlation against the results of the open-ended question, Binary-Open for the very first scene evaluated per participant (n=137). The distribution of categorized responses from Binary-Open is shown in Figure 4. As shown by the Phi coefficient and Point biserial correlation results between Binary-Open and the respective questionnaire item outputs in Table 8, significant correlations with moderate strength were found except for the Glare-indication-diagram. This suggests that the latent construct that was solicited in the survey questionnaire items was indeed regarding discomfort glare. The results from Binary-Open also show that participants reported glare or uncomfortable lighting in the open-ended question even before being asked specifically about glare in the questionnaire.

Table 8 Phi coefficient and Point biserial correlation ρ between the output of the open-ended question (Binary-Open) to that binary and ordinal questionnaire items, respectively. ρ values show the strength of correlations, such as weak ($\rho < 0.4$), moderate ($0.4 \leqslant \rho < 0.6$), and strong correlations ($\rho > 0.6$).[57] "***" indicates p-value < Bonferroni-corrected significance level of 0.0083 ($\alpha = 0.05/8$) for eight comparisons.

	Phi coefficient		Point biserial correlation ρ			
Questionnaire item	Binary-YesNo	Glare-indication-diagram	Likert-4point	OsterhausBailey-4point	Interval-0-10	Comfort-agreement
Binary-Open (n = 137)	0.53***	0.36***	0.52***	0.42***	0.60***	0.50***

5. Discussion

From the conducted comparability study, it appears that the outputs of the four questionnaire items with multiple-point response scales, namely OsterhausBailey-4point, Likert-4point, Interval-0-10, and Comfort-agreement, are inter-dependent, correlate with each other, have high internal reliability, and describe the same latent variable. This means that the distributions

of their results are comparable and assess the same construct but still differ in terms of the level of resolution and semantic interpretations of their response items. Meanwhile, it also revealed that the outputs of two binary questionnaire items, such as Binary-YesNo and Glare-indication-diagram, also point to the same latent variable but seem to solicit different thresholds of glare.

5.1. Interpretations of response items in questionnaires

Despite the high correlation between the outputs of ordinal questionnaire items, there still exist slight nuances and differences between them. Although the OsterhausBailey-4point and Likert-4point both have 4-point response scales that produce similar results, the semantics used in the response items in OsterhausBailey-4point may point to the noticeability instead of the intensity of discomfort glare despite being somewhat in increasing intensity order. For example, some participants may select “Noticeable” glare on the OsterhausBailey-4point indicating that they visually noticed a bright glare source, but that glare source may not generate discomfort for them as they simultaneously also select "Not at all" on the Likert-4point (Figure 8).

The findings of this study also begin to demonstrate the corresponding relationships between the response outputs of these six questionnaire items and the flow of responses between them. For example, while a “6” and above on the Interval-0-10 scale may not have had a clear meaning tied to it so far, this study shows that it may correspond to “moderate” glare and above on the Likert-4point, as seen in Figure 4. As the binary questionnaire items show results in a lower resolution because there are only two response items to choose from, we may compare their output to corresponding response items on ordinal scales as well. For example, from the Binary-YesNo question, the distribution of "Yes" responses corresponds to the distribution of half of the "Noticeable" responses plus the "Disturbing" and "Intolerable" responses from the OsterhausBailey-4point question.

Although we do not believe that the diagrammatic method is better or worse than questionnaire items, the Glare-indication-diagram seems to ask for an additional qualitative concept of where the glare source is located instead of just the reporting of discomfort due to glare. Although it is asked as a single question, it still has a conditional structure in which it asks if the participants feel discomfort from a glare source and if so, to color the location on the diagram. Hence, this questionnaire item may be useful for qualitatively identifying the sources of glare from the user’s perspective and to identify whether participants were attentive and understood the survey if the locations of the indications are not random. For this study, we manually parsed the responses to the Glare-indication-diagram and made sure that responses are logical before entering the data for analysis. Interestingly, for specific conditions, some participants indicated the darker areas of their field of view as sources of glare, implying that they associated this with the effect of contrast instead. Such responses may thus provide interesting spatial feedback on whether the source of discomfort is due to excessive brightness (saturation effect) or contrast. On another note, we found that the descriptor threshold for an indication on the Glare-indication-diagram is approximately corresponding to "Noticeable" on the OsterhausBailey-4point. This means a marking on the diagram represents a glare source noticed by the participant, even if it is just slightly bothersome. On the other hand, a positive response to Binary-YesNo corresponded to a slightly higher degree of glare (between “Noticeable” and “Disturbing” on the OsterhausBailey-4point response scale. This could also be due to a different understanding of a slightly nuanced wording in these two binary questionnaires leading to dissimilar thresholds of glare that solicits a “Yes” in each method: the Binary-YesNo question asks if *discomfort* due to glare is experienced while the Glare-indication-diagram instead asks if *uncomfortable* glare is experienced.

5.2. Simulating a skip sequencing method

The usage of a two-step skip sequencing method has been suggested in recent publications on evaluating discomfort glare.[19,46] The method involved asking a binary question if the participant is experiencing discomfort glare first, then if the answer is “No”, no subsequent question is asked. If the answer is “Yes”, the participant is asked to evaluate the amount of discomfort from glare on a 6-point numerical response scale from 1 to 6, with 1 labeled “Very small amount” and 6 labeled “Very large amount”. As shown in the results of this study, we found that participants who answer “No” to the Binary-YesNo question do not always directly correspond to “0” on the Interval-0-10 response scale, nor to the null response item “Not at all” in the Likert-4point. They also answer “Slightly” or “Noticeable” to other questionnaire items like Likert-4point and OsterhausBailey-4point, as shown in Figure 6. This might be because the participants have more options in the ordinal response scales and can choose a better fitting response for the degree of discomfort glare they experience, than in the binary response scale. This hypothesis is also supported by several past studies indicating that that too few response labels can lead to forced grouping[63] and therefore the number of response labels should be similar to, or greater than, the number of scenes to be evaluated[16,64] which is four in our case.

Using skip sequencing may include 'non-response' and 'response' errors in the second question due to item response errors in the initial question.[65] Hence, while the Binary-YesNo question may reduce the duration of the experiment by not asking more than 1 question when not necessary, it can cause non-response errors in the subsequent question. 'Non-response' errors occur when participants answer "No" to the first question and hence do not get to respond to the second question. As a result, this may change the distribution of the responses to the second question. For example, in this case, around a quarter of the participants answered "No" to Binary-YesNo but answered "Noticeable" to OsterhausBailey-4point.

To check for the significance of such non-response errors, we simulated the skip-sequencing method by comparing two groups formed using our data. The first group contained data points where participants answered “No” to Binary-YesNo are forcefully mapped to the null option of the response scale (e.g., “Imperceptible” in the OsterhausBailey-4point), and the second group has all data points kept regardless of the Binary-YesNo output. Hence, to see if the distribution of results in the two groups is significantly affected by the simulated skip sequencing, a Wilcoxon signed rank test was used, where the null hypothesis is that there is no significant difference between two groups of ordinal data, as shown in Table 9. Since the p-values are less than the Bonferroni-corrected significance level of 0.0125 (α = 0.05/4) for four comparisons, we can reject the null hypothesis for all four tests. This simulation shows that there can be a significant impact of skip-sequencing on the distribution of responses in an ordinal response scale in the second question due to potential non-response errors, as illustrated in Figure 9. However, we observed that the percentage of “disturbing and above” responses on OsterhausBailey-4point question, which is typically used in glare studies, remains similar in both cases, with or without skip-sequencing. This is similar for Likert-4point (“Moderately and above”), Comfort-agreement (“Disagree and above”) and Interval-0-10 (6 and above). Nevertheless, the simulated skip-sequencing showed significant differences to the resulting distribution of lower response items on the glare scales tested. However, to ascertain these simulated results, future research should aim to implement and test skip-sequencing vs. no skip-sequencing protocols (only second question is asked separately) in glare evaluations.

Table 9 p-values of the Wilcoxon signed rank tests run between data with simulated skip-sequencing method and without. "***" indicates p-value < Bonferroni-corrected significance level of 0.0125 (α = 0.05/4) for four comparisons.

Questionnaire item	Likert-4point	OsterhausBailey-4point	Interval-0-10	Comfort-agreement
Effect of simulated skip sequencing	0.00013***	3.14e-07***	5.98e-06***	< 2.2e-16***

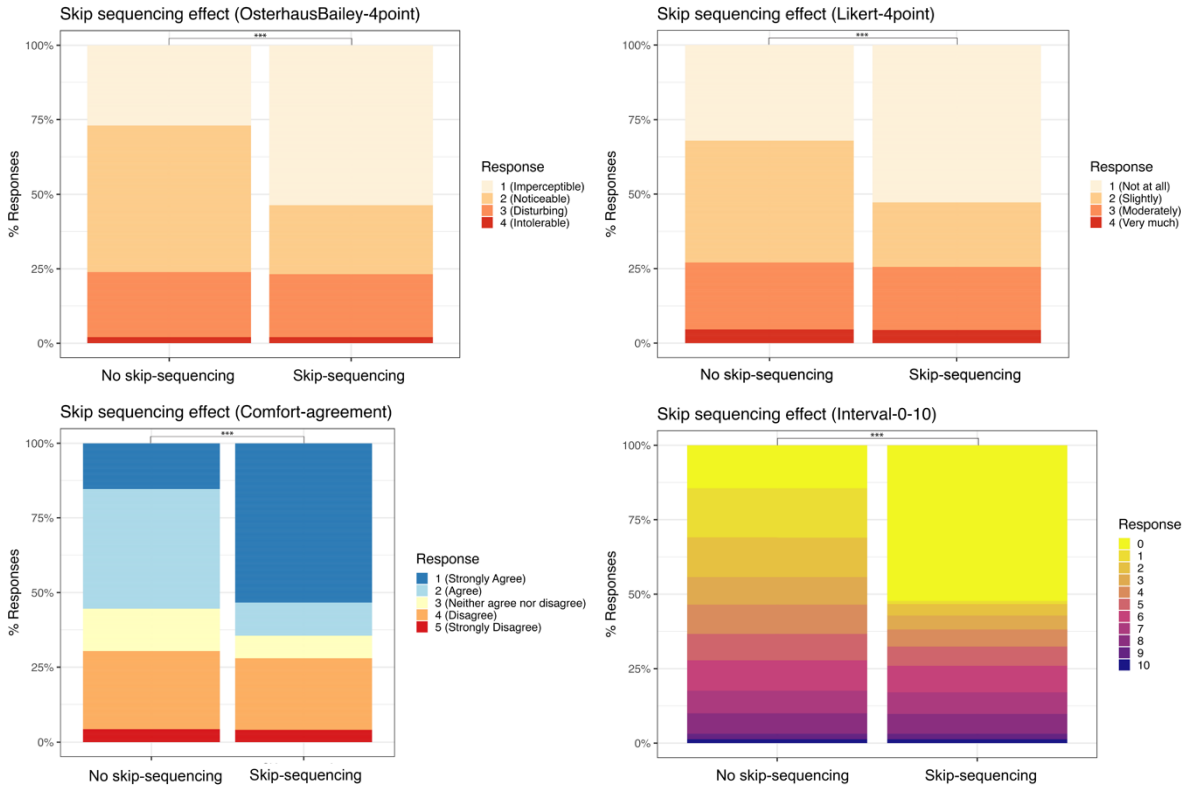


Figure 9 Effects of a simulated skip-sequencing method on the distributions of participants' responses in the four ordinal questionnaire items (without the null response item) and without a skip-sequencing method. Significant effects are labeled "***" based on a Wilcoxon signed rank test. (Available in colour in online version)

One could argue, that participants who answered "No" on the Binary-YesNo question, but went on to answer a non-null response on other questions can be considered as "unreliable" participants. In general, unreliable participants would add noise to the data and therefore result in a lower internal reliability between the questionnaire items. Therefore, to check if the inclusion of the responses of such participants lowers the reliability, we conducted the following test: we removed "unreliable" participants (who answered "No" then answered a non-null response) and the null responses (i.e., participants answering "No" on the Binary-YesNo question). We call this *dataset1*. To create *dataset2,* we only removed the null responses from the ordinal questionnaires. The null answers from both datasets had to be removed for two reasons: 1) the skip-sequenced data would contain a large amount of identical data (100% of the null responses would be, per definition, exactly the same for all scales) which would bias the result and 2) only a potential difference in the non-null distribution matters for this analysis. For both datasets, the internal reliability tests showed they have similar internal reliability as shown in Table 10, indicating that these participants are not "unreliable" because the internal reliability between ordinal questionnaire items did not change significantly – If they were indeed unreliable, the internal reliability should have increased significantly in *dataset1*.

Table 10 Psychometric statistics, to test if all participants are reliable, from Likert-4point, OsterhausBailey-4point, Interval-0-10, and Comfort-agreement.

Psychometric Statistics	dataset1 (Removed "unreliable" *participant*s, n = 240)	dataset2 (n = 330)
Cronbach's α	0.88	0.89
Guttman's Lambda 6 (G6)	0.85	0.87
Omega Total	0.91	0.92
Explained Common Variance (ECV)	0.85	0.95

5.3. Limitations

Some limitations of this study include, first, that the questionnaires were only tested in English and no other translations were tested, which means that the results of this study are limited to questionnaires administered in English. Future research may want to discuss how to effectively translate across several languages and to test if the relationships between the original English questionnaire items and translated ones stay true.

Although there are some devices used in clinical determinations of individuals' discomfort glare through electromyograms,[66] or to measure an individual's sensitivity to discomfort glare,[67] these objective methods are usually considered invasive with the attachments of electrodes around the eye in the former or cover the entire field of view in the latter. Physiological and ocular data such as pupil diameter, pupil unrest index, and eye fixation rate were also found to correlate with glare stimuli[67] but no thresholds to describe glare degree were derived. Artificial intelligence has also been trained to predict if the occupant experiences discomfort glare.[68] However, in most of these studies, the "ground truth" of the degree of discomfort glare perceived is still solicited from subjective evaluations through the choice of a single questionnaire item. Even though objective measures exist, as previously stated, there is still a strong reliance on questionnaires, as they are used to derive semantic meaning for the degree of glare even for objective measurements. Hence, the construct validity of questionnaire items may be proven through the associative relationship with physiological and ocular markers that correlate with glare stimuli, but these objective measures may not convey semantic meanings of the degree of glare perceived (criterion-related validity).

As the three user studies cover different ranges of glare, one may question if there is a range effect, where the within-study analysis would output different results from that of the consolidated dataset. We checked that the conclusions made in the overall study do correspond to that of the individual datasets, indicating no effects of range bias (see detailed results in Section V of the supplementary material). In addition, to check for stimulus range bias such as those found in adjustment-type studies,[69] we re-ran the analysis for only the first scene evaluated by each participant, such that there is no range of stimuli to bias the results of the evaluations. Similarly, the conclusions did not deviate from that of the full dataset, suggesting that there was no significant stimuli range bias in the glare evaluations of the underlying experiments.

The small selection of questionnaire items or more specifically the number of tested items is a limitation since this research is not a representative testing of all so far used glare scales. Considering experimental constraints regarding increased duration of experimental phases when adding more questionnaire items and avoiding annoyance of the participants when asking too many questions in the same direction, we had to limit the number of questions to a reasonable number, which is six. The skip-sequencing effect was only simulated in this analysis using both the Binary-YesNo and the other scales which were individually

administered. More research using the actual skip-sequencing protocol will be needed to validate the preliminary results shown in Section 5b.

Although the response items in the selected questionnaire items do have different semantics, the study of semantics is currently outside the scope of this investigation. In future discussions on glare questionnaire design or standardization such as in the International Standardization Organization[47], as well as when determining the standard questionnaire item to apply for visual discomfort assessments.

It may also be worth noting the necessary range, meaning, and informativeness of semantical categorizations that are required for visual comfort criteria in spaces. The semantics of the response items should ultimately depend on which questionnaire item may provide sufficient differentiated "levels" of discomfort from glare that will be useful for its purpose. For example, considerations of semantics are needed when researching temporal aspects of annual glare requirements - current recommendations of EN17037 recommend that DGP should not report "disturbing" or "intolerable" glare for more than 5% of the occupied time annually based on the OsterhausBailey-4point scale[13]. In addition, the semantic biases that may occur with translation processes between languages may lose original meanings or may not exist in different languages altogether.[16]

6. Conclusion

To evaluate if the type of questionnaire item captures corresponding or contradictory distributions of glare responses, we selected and compared six questionnaire items for evaluating discomfort glare in rating-type experiments. They were subsequently administered to each participant in a randomized order when implemented in three user studies and resulted in a diverse dataset of lighting conditions with 540 user assessment data points from 149 individuals. The outputs of the six questionnaire items were examined pairwise descriptively and then tested for association, reliability, and dimensionality.

The first finding of the study is that the outputs of ordinal questionnaire items tested were found to have strong correlations with each other, have excellent internal reliability, and point to the same latent variable. We make the reasonable assumption that this variable refers to the degree of discomfort caused by glare, backed up by the construct validity check compared to the open-ended question. This finding signifies that the four tested ordinal questionnaire items are interchangeable to some extent. This means that their results distributions are comparable and assess the same construct, but they differ in terms of informativeness (the level of resolution and semantic interpretations of their response items). We also confirmed the validity of the latent construct using responses from Binary-Open only for the first evaluation per participant, an open-ended question asked before the six questionnaire items were administered.

The response outputs of the Binary-YesNo question and Glare-indication-diagram correlate well with those of ordinal ones, as well as responses from the open-ended question, Binary-Open. This means they solicit about the same latent construct, which is the degree of glare experienced. Results also show that these two tested binary questionnaire items where users should indicate "any discomfort" may point correspond to slightly higher thresholds of the degrees of glare compared to scales with a finer resolution (around 25% of the participants that indicated "No" still reported a slight discomfort or higher on other scales). Therefore, for research questions requiring information about lower levels of glare, we recommend using scales with a finer resolution.

Our findings also indicate that participants may interpret each of the binary items differently; some may indicate "Yes" when there is slight discomfort while for others only when there is moderate discomfort. This finding let us also hypothesize that the usage of the 3rd response

item "discomfort" from Annex C of ISO10551:2019[47] might be interpreted differently by different participants and therefore a modification should be considered.

Overall, this study provides a scientific basis for future psychometric research and discussions about the usability and applicability of questionnaire items for collecting user evaluations of discomfort from glare in rating-type studies in daylight. The selected six questionnaire items from previous glare research do point to the same latent variable and therefore are valid for use in daylight glare studies that use similar rating-type procedures. They mainly differ in the granularity and the levels and thresholds of glare they solicit and therefore researchers should select them depending on their research question. Nonetheless, we recommend researchers to use at least two types of questionnaire items to ensure that participants understand the questions, especially if items are translated or if people with diverse backgrounds participate in the studies. Our findings hope to support future discussions on glare questionnaire standardizations and as such, this study does not intend to specifically recommend any of the tested glare questionnaire items. We believe that more psychometric research is also needed to ascertain these findings for adjustment-type studies commonly conducted for evaluating glare from electric light sources.

The authors declare no potential conflicts of interest with respect to the research, authorship and/or publication of this paper.

Funding
This research was funded by the Swiss National Science Foundation (SNSF) as part of the ongoing research project, “Visual comfort without borders: Interactions on discomfort glare” (SNSF #182151). Geraldine Quek is a recipient of the Graduate Merit Scholarship from the Singapore University of Technology and Design (SUTD).

Acknowledgments
We thank Andrew S. Dissanayake of the Neurosciences and Mental Health Program, The Hospital for Sick Children, Toronto, ON, Canada, for his advice in psychometric statistics.

ORCiD

Quek, Geraldine	0000-0003-2864-3860
Jain, Sneha	0000-0001-7981-2754
Karmann, Caroline	0000-0003-3328-7936
Pierson, Clotilde	0000-0001-7847-6568
Wienold, Jan	0000-0002-3723-0323
Andersen, Marilyne	0000-0001-8813-1184

References

1. DeVellis RF, Thorpe CT. *Scale Development: Theory and Applications*. SAGE Publications, 2021.
2. Worthington RL, Whittaker TA. Scale Development Research: A Content Analysis and Recommendations for Best Practices. *The Counseling Psychologist* 2006; 34: 806–838.
3. Morrison JT. Evaluating Factor Analysis Decisions for Scale Design in Communication Research. *Communication Methods and Measures* 2009; 3: 195–215.
4. Walford G, Tucker E, Viswanathan M. *The SAGE Handbook of Measurement*. SAGE, 2010.
5. Gobeille M, Bradley C, Goldstein JE, et al. Calibration of the Activity Inventory Item Bank: A Patient-Reported Outcome Measurement Instrument for Low Vision Rehabilitation. *Translational Vision Science & Technology* 2021; 10: 12.
6. Cohen S, Kamarck T, Mermelstein R. A Global Measure of Perceived Stress. *Journal of Health and Social Behavior* 1983; 24: 385–396.
7. Cohen S. Perceived stress in a probability sample of the United States. In: *The social psychology of health*. Thousand Oaks, CA, US: Sage Publications, Inc, 1988, pp. 31–67.
8. Lee E-H. Review of the Psychometric Evidence of the Perceived Stress Scale. *Asian Nursing Research* 2012; 6: 121–127.
9. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep* 1991; 14: 540–545.
10. Holdgate A, Asha S, Craig J, Thompson J. Comparison of a verbal numeric rating scale with the visual analogue scale for the measurement of acute pain. *Emergency Medicine (Fremantle, W.A.* 2003; 15: 441–446.
11. Ohnhaus EE, Adler R. Methodological problems in the measurement of pain: a comparison between the verbal rating scale and the visual analogue scale. *Pain* 1975; 1: 379–384.
12. Kaida K, Takahashi M, Åkerstedt T, Fukasawa K. Validation of the Karolinska sleepiness scale against performance and EEG variables. *Clinical Neurophysiology* 2006; 117: 1574–1581.
13. European Committee for Standardization. *Daylight in Buildings.* European standard EN 17037: 2022-05. Brussels: CEN, 2022
14. European Committee for Standardization. *Light and Lighting – Lighting of Work Places Part 1 : Indoor Work Places*. European standard EN 12464-1:2021. Brussels: CEN, 2021.
15. CIE. *ILV: International Lighting Vocabulary*. Standard CIE S 017/E:2011, Vienna, Austria: CIE Central Bureau.
16. Fotios S, Kent M. Measuring Discomfort from Glare: Recommendations for Good Practice. *Leukos* 2021; 17: 338–358.
17. Gellatly AW, Weintraub DJ. *User reconfigurations of the de Boer rating scale for discomfort glare*. The University of Michigan, Transportation Research Institute, 1990.
18. Fotios S. Using Category Rating to Evaluate the Lit Environment: Is a Meaningful Opinion Captured? *Leukos* 2019; 15: 127–142.
19. Hickcox KS, Fotios S, Abboushi B, et al. Correspondence: A new two-step approach for evaluating discomfort from glare. *Lighting Research and Technology* 2022; 54: 91–92.
20. Allan AC, Garcia-Hansen V, Isoardi G, Smith S. Subjective Assessments of Lighting Quality: A Measurement Review. *Leukos* 2019; 15: 115–126.
21. Fotios S. Research Note: Uncertainty in subjective evaluation of discomfort glare. *Lighting Research and Technology* 2015; 47: 379–383.

22. Hopkinson RG. The Multiple Criterion Technique of Subjective Appraisal. *Quarterly Journal of Experimental Psychology* 1950; 2: 124–131.
23. Geerdinck L. *Glare Perception in Terms of acceptance and Comfort.* Graduation Report, Eindhoven University of Technology, 2012.
24. Heise DR. Some methodological issues in semantic differential research. *Psychological Bulletin* 1969; 72: 406–422.
25. Saris WE, Gallhofer IN. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. John Wiley & Sons, 2014.
26. Iodice M. *Méthodes de caractérisation psychophysique et physiologique de l'éblouissement d'inconfort en éclairage artificiel intérieur : étude de l'influence du contenu spectral des sources lumineuses*. *Methods for psychophysical and physiological characterization of discomfort glare in artificial indoor lighting: study of the influence of the spectral content of light sources.* These de Doctorat, Université de Lyon, 2020 .
27. Hopkinson RG. Discomfort Glare in Lighted Streets. *Transactions of the Illuminating Engineering Society* 1940; 5: 1–32.
28. Petherbridge P, Hopkinson RG. Discomfort Glare and the Lighting of Buildings. *Transactions of the Illuminating Engineering Society* 1950; 15: 39–79.
29. MacGowan D. Correspondence. *Lighting Research and Technology* 2010; 42: 121–122.
30. Hopkinson RG, Bradley RC. A study of glare from very large sources. *Illuminating Engineering* 1960; 55: 288–297.
31. Hopkinson RG. A Note on the Use of Indices of Glare Discomfort for a Code of Lighting. *Lighting Research and Technology* 1960; 25: 135–138.
32. Robinson W, Bellchambers HE, Grundy JT, et al. The Development of the IES Glare Index System: Contributed by the Luminance Study Panel of the IES Technical Committee. *Transactions of the Illuminating Engineering Society* 1962; 27: 9–26.
33. Kent MG, Fotios S, Altomonte S. Order effects when using Hopkinson's multiple criterion scale of discomfort due to glare. *Building and Environment* 2018; 136: 54–61.
34. Chauvel P, Collins JB, Dogniaux R, Longmore J. Glare from windows: current views of the problem. *Lighting Research and Technology* 1982; 14: 31–46.
35. Iwata T, Tokura M, Shukuya M, Kimura K. Experimental study on discomfort glare caused by windows Part 2: Subjective response to glare from actual windows. *Journal of Architecture, Planning and Environmental Engineering (Transactions of AIJ)* 1992; 439: 19–31.
36. International Commission on Illumination (CIE). *Discomfort Glare in Interior Lighting*. 117:1995, Vienna, Austria: CIE, 1995.
37. Sörensen K. A modern glare index method. *Proceedings of 21st Session of the CIE* 1987, pp. 17–25.
38. Einhorn HD. A new method for the assessment of discomfort glare. *Lighting Research and Technology* 1969; 1: 235–247.
39. Einhorn HD. Discomfort glare: a formula to bridge differences. *Lighting Research and Technology* 1979; 11: 90–94.
40. Fisekis K, Davies M, Kolokotroni M, Langford P. Prediction of discomfort glare from windows. *Lighting Research and Technology* 2003; 35: 360–369.
41. Wienold J, Christoffersen J. Evaluation methods and development of a new glare prediction model for daylight environments with the use of CCD cameras. *Energy and Buildings* 2006; 38: 743–757.
42. Osterhaus WKE, Bailey IL. Large area glare sources and their effect on visual discomfort and visual performance at computer workstations. In: *Conference Record of the 1992 IEEE Industry Applications Society Annual Meeting*. 1992, pp. 1825–1829 vol.2.
43. Hirning MB, Isoardi GL, Cowling I. Discomfort glare in open plan green buildings. *Energy and Buildings* 2014; 70: 427–440.

44. Boateng GO, Neilands TB, Frongillo EA, et al. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*; 6: 149.
45. Hirning MB, Isoardi GL, Garcia-Hansen VR. Prediction of discomfort glare from windows under tropical skies. *Building and Environment* 2017; 113: 107–120.
46. Hamedani Z, Solgi E, Hine T, et al. Lighting for work: A study of the relationships among discomfort glare, physiological responses and visual performance. *Building and Environment* 2020; 167: 106478.
47. International Organization for Standardization. ISO 10551:2019(en), Ergonomics of the physical environment — Subjective judgement scales for assessing physical environments 2019.
48. Tourangeau R, Rips LJ, Rasinski K. *The Psychology of Survey Response*. Cambridge University Press, 2000.
49. Quek G, Wienold J, Andersen M. User evaluations of contrast-dominant discomfort glare in dim daylit scenarios: Preliminary findings. In: *CIE 2021 Midterm Meeting & Conference*. Kuala Lumpur, Malaysia, 2021.
50. Karmann C, Chinazzo G, Schüler A, et al. User assessment of fabric shading devices with a low openness factor. *Building and Environment* 2022; 109707.
51. Jain S, Wienold J, Andersen M. Effect of window glazing color and transmittance on human visual comfort. In: *PLEA 2022 SANTIAGO Will Cities Survive?* Santiago, Chile: PLEA, 2022.
52. Jain S, Wienold J, Lagier M, et al. Perceived glare from the sun behind tinted glazing: Comparing blue vs. color-neutral tints. *Building and Environment* 2023; 234: 110146.
53. Pierson C, Piderit B, Iwata T, et al. Is there a difference in how people from different socio-environmental contexts perceive discomfort due to glare from daylight? *Lighting Research and Technology* 2022; 54: 5–32.
54. Pierson C, Sarey Khanie M, Bodart M, et al. Discomfort glare cut-off values from field and laboratory studies. In: *PROCEEDINGS OF the 29th Quadrennial Session of the CIE*. Washington DC, USA: International Commission on Illumination 2019, pp. 295–305.
55. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; 16: 297–334.
56. McDonald R. *Test Theory: A Unified Treatment*. Psychology Press, 2013.
57. Dancey CP, Reidy J. *Statistics Without Maths for Psychology*. Pearson Education, 2007.
58. Taber KS. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education* 2018; 48: 1273–1296.
59. Cortina JM. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology* 1993; 78: 98–104.
60. Bentler PM. Comparative fit indexes in structural models. *Psychological Bulletin* 1990; 107: 238–246.
61. Fabrigar LR, Wegener DT, MacCallum RC, et al. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 1999; 4: 272–299.
62. Xia Y, Yang Y. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods* 2019; 51: 409–428.
63. Matell MS, Jacoby J. Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology* 1972; 56: 506–509.
64. Wang J, Wang Z, de Dear R, et al. The uncertainty of subjective thermal comfort measurement. *Energy and Buildings* 2018; 181: 38–49.
65. Manski CF, Molinari F. Skip Sequencing: A decision problem in questionnaire design. *The annals of applied statistics* 2008; 2: 264–285.

66. Murray I, Plainis S, Carden D. The ocular stress monitor: a new device for measuring discomfort glare. *Lighting Research and Technology* 2002; 34: 231–239.
67. Montés-Micó R, Cerviño A, Martínez-Albert N, et al. Performance of a new device for the clinical determination of light discomfort. *Expert Review of Medical Devices* 2020; 17: 1221–1230.
68. Johra H, Gade R, Poulsen MØ, et al. Artificial Intelligence for Detecting Indoor Visual Discomfort from Facial Analysis of Building Occupants. *Journal of Physics: Conference Series* 2021; 2042: 012008.

69. Kent MG, Fotios S, Cheung T. Stimulus range bias leads to different settings when using luminance adjustment to evaluate discomfort due to glare. *Building and Environment* 2019; 153: 281–287.