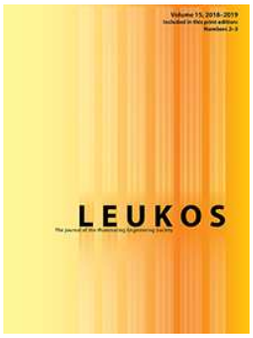

Article:
Fotios, S. orcid.org/0000-0002-2410-7641 (2019) Using category rating to evaluate the lit environment: Is a meaningful opinion captured? LEUKOS, 15 (2-3). pp. 127-142. ISSN 1550-2724

https://doi.org/10.1080/15502724.2018.1500181



LEUKOS

The Journal of the Illuminating Engineering Society

ISSN: 1550-2724 (Print) 1550-2716 (Online) Journal homepage: https://www.tandfonline.com/loi/ulks20

Using Category Rating to Evaluate the Lit Environment: Is a Meaningful Opinion Captured?

Steve Fotios

To cite this article: Steve Fotios (2019) Using Category Rating to Evaluate the Lit Environment: Is a Meaningful Opinion Captured?, LEUKOS, 15:2-3, 127-142, DOI: 10.1080/15502724.2018.1500181

To link to this article: https://doi.org/10.1080/15502724.2018.1500181

Published online: 26 Nov 2018.

LEUKOS
2019, VOL. 15, NOS. 2–3, 127–142
https://doi.org/10.1080/15502724.2018.1500181

Using Category Rating to Evaluate the Lit Environment: Is a Meaningful Opinion Captured?

Steve Fotios

School of Architecture, University of Sheffield, Sheffield, UK

ABSTRACT

Do responses gained using category rating accurately reflect respondents' true evaluations of an item? "True" in this sense means that they have a real opinion about the issue, rather than being compelled by the survey to speculate an opinion, and that the strength of that opinion is faithfully captured. This article describes some common issues that suggest that it should not be simply assumed that a response gained using category rating reflects a true evaluation. That assumption requires an experiment to have been carefully designed and interpreted, and examples are shown where this is not the case. The article offers recommendations for good practice.

ARTICLE HISTORY
Received 19 October 2017
Revised 10 July 2018
Accepted 11 July 2018

KEYWORDS
Category rating; lighting; subjective evaluation

1. Introduction

Category rating is one of several procedures that might be used to provide a quantitative subjective evaluation, alongside matching, discrimination, and adjustment (CIE 2014). Observers are given a limited set of categories for their responses, usually arranged in order of magnitude. The categories are given either a numeric (for example, 1, 2, . . ., 6, 7 for a seven-point scale) or descriptive label (for example, very small, . . ., very large), and in some cases only the two extreme categories are labeled. Though these response categories are clearly ordered (for example, a progressive increase in the degree of discomfort for a discomfort glare evaluation) and thus might be more precisely defined as ordinal scales, they are widely described in the literature as category rating scales.

Figure 1 shows an example of results gained using category rating: these are undergraduate students' evaluations of the quality of lectures delivered by the author. Five items were evaluated, including *clarity of presentation*, and responses were given using a five-point scale that ranged from *very poor* to *very good*. It can be seen, for example, that clarity of presentation was evaluated as good by 30 respondents, as average by 15 respondents, with five each for very good and poor. Data such as these are used by lecturers to evaluate changes in their methods of lecture delivery and used by management to inform their opinion of a lecturer's progress and their salary review. Such data are therefore important.

Category rating is widely used in studies related to lighting, including glare (De Boer and Schreuder 1967), brightness (Boyce and Cuttle 1990), the alleged Kruithof effect (Fotios 2017a; Kakitsuba 2016), the overall impression of road lighting (Simons et al. 1987), visual clarity (Flynn and Spencer 1977; Vrabel et al. 1998), and perceived safety (Boyce et al. 2000). In a recent review of 70 studies investigating how changes in lamp spectral power distribution (SPD) affect spatial brightness, 30 used category rating; the remainder used either matching, adjustment, or discrimination procedures (Fotios et al. 2015a).

One reason for the prevalence of category rating questionnaires in research is that they are relatively cheap and apparently simple to conduct: "subjective, or pencil and paper, methods often present a cheaper alternative to those involving instrumentation, be it sensors for physiological recording or video for recording behaviour" (Annett 2002). The collection of a large amount of data may give the potentially incorrect assumption that these are good data. Inappropriate questionnaire design and lack of

CONTACT Steve Fotios steve.fotios@sheffield.ac.uk School of Architecture, University of Sheffield, The Arts Tower, Western Bank, Sheffield S10 2TN, UK.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ULKS.
Open Access for this manuscript was supported by the Illuminating Engineering Society and International Commission on Illumination.

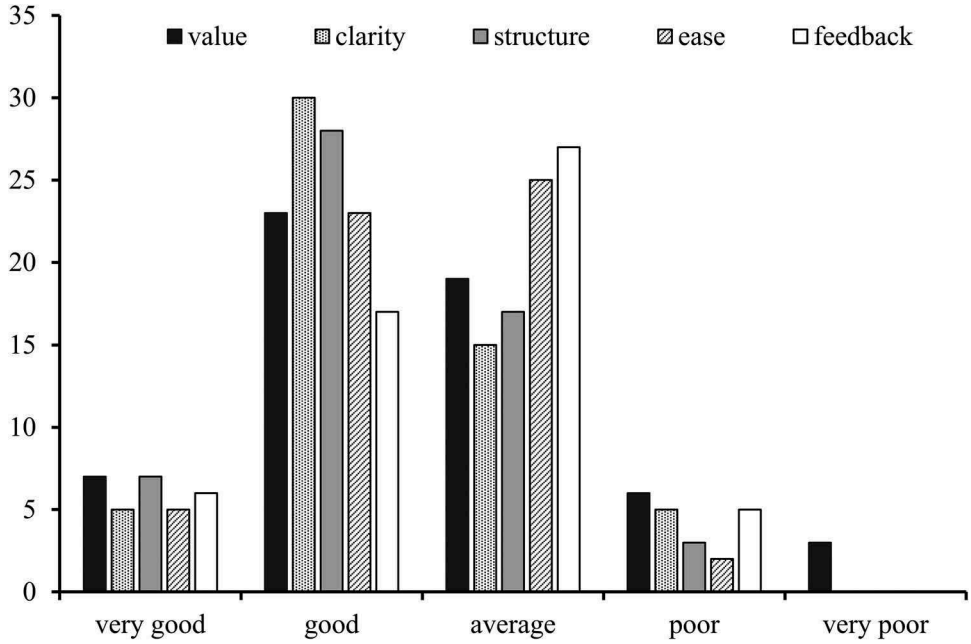


Fig. 1. An example of category rating: Undergraduate students' evaluations of a lecture series. The five items evaluated were value of lectures, clarity of presentation, structuring of course content, ease of understanding, and feedback.

rigor, however, can lead to poor data quality, misleading conclusions, and incorrect recommendations: "Anybody can write down a list of questions and photocopy it, but producing worthwhile and generalisable data from questionnaires needs careful planning and imaginative design" (Boynton and Greenhalgh 2004).

An awareness of good practice in experimental methods can help to make progress. Returning to the review of spatial brightness: at first sight there is no consensus—some studies suggest a significant effect of SPD on brightness (Berman et al. (1990); Boyce and Cuttle (1990), experiment 2), whereas others do not suggest the effect to be significant (Boyce and Cuttle (1990), experiment 1; Davis and Ginthner (1990)). Review of methods was used to establish credible evidence; only 19 of the 70 studies were considered to provide credible evidence (Fotios et al. 2015a); the reminder did not meet criteria for robust experimental design (CIE 2014), and these 19 gave consistent evidence that SPD affects spatial brightness.

One model of the survey response process proposes four sequential (or at least partially overlapping) components: comprehension, retrieval, judgment, and response (Tourangeau et al. 2000). The mental processes within these components include comprehension of the question, retrieval from memory of relevant information, and matching the internally generated answer to one of the available response categories (Lietz 2010; Tourangeau et al. 2000). Response effects may arise at each step: a respondent may misinterpret the question, forget crucial information, make an erroneous inference, or map his or her response to the wrong response category (Tourangeau et al. 2000). Retrieval from memory introduces noise; for example, a tendency to recall previously encountered physical stimuli as lower (for example, shorter in length, or less bright) than the original stimuli (LaBoeuf and Shafir 2006). Though this may be reduced by presenting a reference stimulus along with the test stimulus (Alfonso-Reese 2001), it is not common practice in laboratory studies of lighting and may be impractical in field studies. The step from retrieval to matching is subject to influences on the respondent's ability, motivation, and preparedness to be truthful. For example, in fear-of-crime studies, male respondents may be inclined to underreport their degree of fear, an example of socially desirable responding (Sutton and Farrall 2005; Weisberg 2005).

For visual evaluations, responses are informed by both cognitive and sensory factors (Gescheider 1988). It is not only the stimulus that matters but also the way in which the question is asked. In the context of category rating, "error" refers to the difference between an obtained value and the

true value (Weisberg 2005). Measurement error occurs when the measure obtained is not an accurate measure of what was to be measured. In other words, whether respondents are giving the answers they should, given the experimenter's intentions, and this is often a matter of how well the question was designed. Even minor variations in the design and formulation of questions and response scales can have major effects on the responses obtained and hence the conclusions drawn (Lietz 2010).

The aim of this article is to draw attention to some of the factors that, if not properly considered during experimental design, may cast doubt on the credibility of data gathered using category rating and, in doing so, to extend previous discussion (Fotios et al. 2015a; Fotios and Houser 2009) toward establishing procedural improvements for lighting research. This is done through consideration of whether evaluated items and response options are consistently defined by experimenters and respondents, whether the opinion expressed is meaningful, and the influence of range equalizing bias.

2. What is being evaluated?

When a questionnaire asks for an item to be evaluated, it is anticipated that respondents define the item in the same manner as the experimenter. Many studies have examined brightness (Fotios et al. 2015a). Though there is probably a consistent understanding of brightness, this should not be assumed: Note, for example, the article "Perceived Brightness and Classroom Interactions" (Armstrong et al. 1979) included within the CIE classified bibliography on brightness and luminance relations (CIE 1988). This was, however, an erroneous inclusion because the brightness it alludes to is the perceived intelligence of schoolchildren, not a perceived amount of light. Brightness is also a quality of the acoustic environment (Kato et al. 2010; Song and Kim 2009).

Within lighting, brightness is a well-defined term. It is defined by the CIE as "the attribute of a visual perception according to which an area appears to emit, or reflect, more or less light" (CIE 2016). The focus of many past studies associated with interior lighting (Fotios et al. 2015a) is more precisely named *spatial* brightness. An early definition of spatial brightness was as follows:

> Spatial brightness describes a visual sensation to the magnitude of the ambient lighting within an environment, such as a room or lighted street. Generally the ambient lighting creates atmosphere and facilitates larger visual tasks such as safe circulation and visual communication. This brightness percept encompasses the overall sensation based on the response of a large part of the visual field extending beyond the fovea. It may be sensed or perceived while immersed within a space or when a space is observed remotely but fills a large part of the visual field. Spatial brightness does not necessarily relate to the brightness of any individual objects or surfaces in the environment, but may be influenced by the brightness of these individual items. (Fotios and Atli 2012)

The recent revision of the CIE lighting vocabulary presented a more concise definition (CIE 2016): "Attribute of a visual perception according to which a luminous environment appears to contain more or less light." In each of these definitions, brightness is associated with the perceived magnitude of "how much light."

Though several studies have investigated visual clarity (Aston and Bellchambers 1969; Bellchambers and Godby 1972; DeLaney et al. 1978; Hashimoto and Nayatani 1994; Thornton and Chen 1978; Vrabel et al. 1998; Worthey 1985), there does not appear to be an agreed-upon definition for this perception. According to some, visual clarity is simply an alternative term for spatial brightness (Flynn et al. 1973; Hashimoto and Nayatani 1994), the latter term not being defined until more recently and specifically with regard to the effect on spatial brightness of changes in SPD (Lynes 1996). According to others, however, visual clarity is instead associated with contrast and distinctness of detail (Aston and Bellchambers 1969; Hashimoto et al. 2000; Thornton and Chen 1978; Vrabel et al. 1998; Worthey 1985).

Given that researchers do not agree on what visual clarity means, we should not expect naïve test participants to have a consistent understanding or for their understanding to match that held by the experimenter. In this situation, an experimenter may be drawing conclusions that were not intended by test participants. One way to encourage consistent understanding when asking for evaluations would be for visual clarity (and other items) to be explained or illustrated prior to seeking a response. This,

however, appears to be done only rarely (Vrabel et al. 1998).

It is likely that naïve observers are familiar with the term *visual clarity*: when Boyce and Cuttle (1990) asked test participants to describe in their own words the lighting of a room, brightness and clarity were the terms most frequently used. As to what these respondents meant by visual clarity, however, we do not know. What is known is that responses given to evaluations of visual clarity tend to be similar to those for spatial brightness, when using a matching procedure (Fotios and Atli 2012; Fotios and Gado 2005) or a category rating procedure (Fig. 2) (Fotios and Atli 2012). This may be an example of respondents using visual clarity as a metaphor for something simpler and more familiar; that is, brightness (Tiller and Rea 1992).

Participants in an experiment may provide evaluations even if the question or response scale is nonsensical. In one study, 20 concepts (including "boulder," "lady," and "fraud") were evaluated using a series of seven-point semantic differential scales including sweet–sour, hot–cold, pungent–bland, fragrant–foul, and bright–dark (Osgood et al. 1975). Despite the semantic differential scales not usually being associated with these particular concepts, responses were given to all items by all 100 respondents. In common with many studies, there was no clear option to ignore a question where the question or response scale was not understood or if no strong opinion was held. In this particular study, a purposeful academic study rather than a field study, it may be that the respondents, students in Introductory Psychology, responded because they were instructed to. That tendency may be common in many studies, the outcome being that conclusions are drawn from false assumptions of meaningfulness. Again, in this particular study, it may have been the case that responses were placed in the central category to indicate that neither of the semantic differential labels fit the evaluated concept.

Respondents can misinterpret questions, even apparently well-formulated questions, and when that happens a respondent may not be answering the question the experimenter assumed they had asked (Tourangeau et al. 2000). In lighting research this may be particularly applicable to the parameters of a visual scene the test participants are asked to evaluate: just because the test instructions requested judgment of a certain parameter does not mean that the results gained from test participants are for the same visual phenomena the experimenter intended. In other words, "an investigator's intended meaning for scales like brightness, spaciousness or comfort may not be interpreted in the same way by the subjects" (Rea 1982). There are two approaches

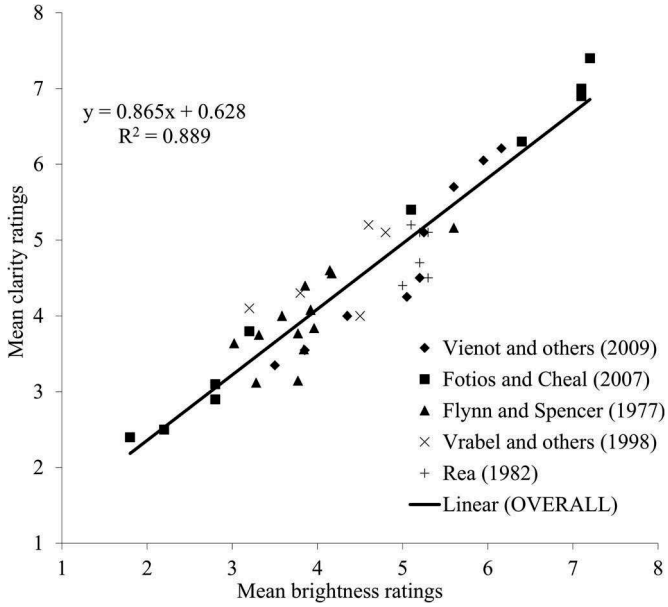


Fig. 2. Mean ratings of brightness plotted against mean ratings of visual clarity from past studies (Flynn and Spencer 1977; Fotios and Cheal 2007a; Rea 1982; Vienot et al. 2009; Vrabel et al. 1998) after Fotios and Atli (2012).

an experimenter might use to mitigate this problem. One approach is to define to test participants the nature of the evaluated items (Fotios and Houser 2009; Houser and Tiller 2003; Tiller and Rea 1992), and this definition may be a written one (Vrabel et al. 1998) or a visual demonstration of visual scenes exemplifying levels of a particular item. The second approach is to ask test participants to describe the items in their own words (or, similarly, by using focus groups to probe understanding of a draft questionnaire) (Tourangeau et al. 2000); the experimenter would subsequently consider whether these descriptions matched the experimenter's own definitions. These approaches are not, however, perfect solutions. The former approach may unintentionally lead toward an expected answer, in particular when participants are being asked to evaluate more abstract items such as diffuse rather than more fundamental items such as brightness. The latter approach may demand a degree of experimenter interpretation where the descriptions given by participants include unclear terms.

3. Is the expressed opinion meaningful?

With a moment of reflection you may be able to recall expressing an opinion about a subject simply because you were asked, not because it was a deeply held or significant or educated opinion. Fig. 1 provides an example of this. These ratings of a lecture series were sought independently of the lecturer and, on this occasion, were sought in error *before* the lectures had commenced: when the students provided these responses they had not had any lectures on the particular subject or by the particular lecturer, but they responded to the rating questions anyway. Only one respondent raised this error. This is not a unique situation: another lecturer has reported receiving an average grade of 4.2/5.0 for being good at explaining things before the lectures were delivered, increasing to 4.7/5.0 after the lectures had been delivered (Foulsham 2018).

Figure 3 shows the responses obtained *after* the lecture series. These might be considered more credible than those shown in Fig. 1, given that the students had at least experienced the lectures they were asked to evaluate. A comparison of Figs. 1 and 3, however, suggests otherwise: the two graphs reveal very similar distributions of responses, both displaying an apparent central tendency. It may be coincidence that the naïve ratings of Fig. 1 match the experience-based ratings of Fig. 3. Alternatively, it may be the case that both sets of responses were influenced by the same biases.

Studies investigating road lighting have used rating scales to evaluate fear of crime under different lit conditions. Rating scales are usually administered within a questionnaire that contains a number of questions: this may prompt respondents to give an opinion about an issue for which they would otherwise not wish to express an opinion. This can be seen in two studies. Ramsay and Newton (1991) refer to an unpublished study of lighting in Deptford, at the time a high-crime inner-city area of London: when asked to list the three main disadvantages of their location, only 8% of respondents mentioned poor lighting, but when asked specifically whether better lighting would decrease fear of crime, 80% agreed. Acuña-Rivera et al. (2011) used qualitative and quantitative evaluations of signs of incivility on simulated residential roads. In the qualitative method, participants were asked to provide at least five words that best expressed their impression of the scene: though participants referred to physical disorder, only a few mentioned crime and safety. The quantitative method used a series of rating scales applied to the same scenes, and these suggested that safety was an issue.

These examples reveal a problem of the category rating questionnaire. Respondents may express an opinion about an issue for which they have no basis of opinion or would not have chosen to evaluate or considered to be relevant if they had not been prompted to respond by the questionnaire.

When used in field studies, rating scales give the respondent the opportunity to complain. Consider the online survey of perceived safety and road lighting carried out by The Suzy Lamplugh Trust and Neighbourhood Watch, two not-for-profit organizations concerned with crime and safety, which received responses from 15,786 people across England (Neighbourhood Watch and Suzy Lamplugh Trust 2013). Two questions asked "How safe do you feel when walking in . . ." either a well-

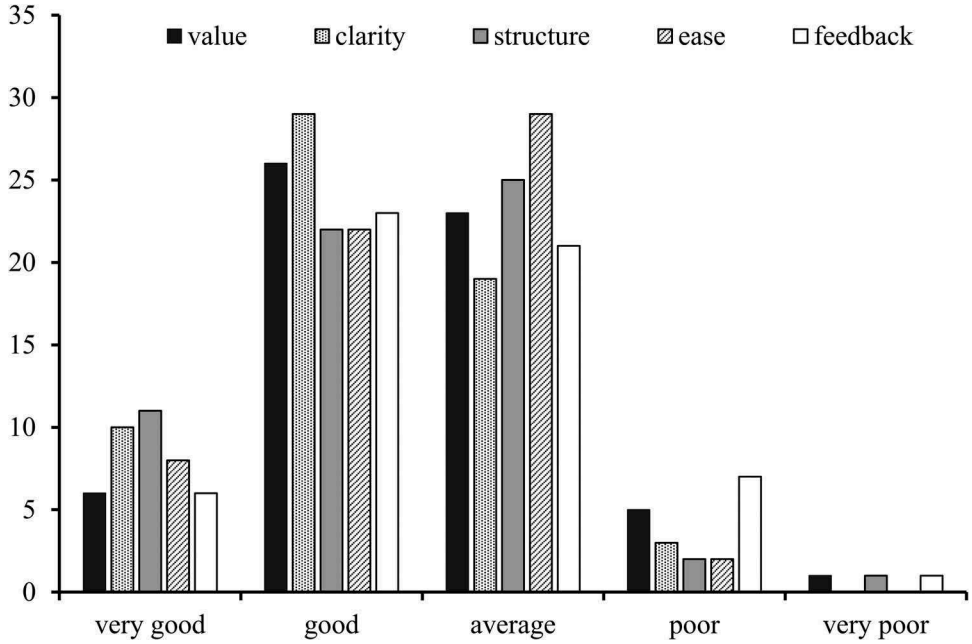


Fig. 3. Undergraduate students' evaluations of a lecturer captured *after* the lecture series; those given in Fig. 1 were from the same student group but mistakenly sought *before* the lecture series.

lit, or an unlit/badly lit, neighborhood; the four response categories were *very safe, fairly safe, a bit unsafe*, and *very unsafe*. The original report indicated a general trend for people to think that they would feel safer in the well-lit area. A subsequent independent analysis (Fotios 2016a) assigned numeric values to the four categories (*very safe* = 4 to *very unsafe* = 1) and calculated for each person the difference between the well-lit and badly lit scores. According to this analysis, 84% of respondents indicated that well-lit lighting would improve their feeling of safety by at least one grade on the four-point rating scale.

Finally, the respondents were asked whether they noticed any changes to street lighting in the past 3 years and, if so, whether they now felt safer, no change, or less safe. Of the 5929 respondents who had noticed a change, only 7% responded that they now felt safer, and their reasons for this included that the new lighting was brighter and/or was whiter. In contrast, 48% of respondents stated that they were less safe after the change, and 45% reported that their level of safety was about the same. The reasons given for feeling less safe included lamp posts being moved from directly outside the respondents' houses, there not being enough lamp posts on a particular street, and the use of assumed energy-saving or light emitting diode light sources: these might be seen as complaints about changes to a resident's immediate environment rather than a real decrease in perceived safety.

The analysis (Fotios 2016a) found that though a majority (84%) of respondents thought that they would feel safer in an area that they considered to be well lit rather than badly lit or unlit, when road lighting was changed in their neighborhood this translated into an increase in safety for only a minority (7%). Evaluations similar to the former ("I think I will feel safer if the lighting is improved") may influence local authorities to allocate a budget to lighting improvements. If that is an attempt by the authority to do something positive for an area, then we should question how much value should be placed on people's speculative opinions about something they have not yet experienced: the budget may be better spent elsewhere.

This section has raised the question of whether the response given is a meaningful opinion. The use of supplementary questions can help to understand this. In the first situation, respondents were asked to give an opinion about something for which they had no experience or knowledge upon which to base their response but responded anyway. Some respondents might recognize their

own lack of knowledge and would prefer not to respond but feel compelled to by the situation in which the survey is conducted. The experimenter needs to consider whether and how these responses should be filtered out (Lietz 2010). This could be addressed by adding supplementary questions that lead to the option of not using the rating scale—the "don't know" option. In the lecture review example, the question could be, "Have you attended the specific lecture course? (yes/no)" with the course evaluations being given only following a "yes." Acuña-Rivera et al. (2011) offer an alternative approach, the use of a second means of evaluation alongside the rating scale: if the outcomes of both converge on the same conclusion, then it may be considered a more robust conclusion. The second issue is that respondents may use a questionnaire as an opportunity to complain about something else that affects them; for example, if a questionnaire is administered shortly after an intervention such as a change in road lighting or the move of an office. A better understanding might be gained if participants are also given the option to state the reasons for the evaluations in their own words.

4. Response scales

Discomfort glare is the form of glare that causes visual discomfort without necessarily impairing the vision of objects (Boyce 2014). A nine-point response range is often used to quantify the magnitude of discomfort due to glare, and for evaluations associated with outdoor lighting this is often known as the de Boer scale. De Boer scales have been used in studies associated with glare from outdoor lighting (Bullough et al. 2008; Tashiro et al. 2014; Villa et al. 2017), vehicle headlamps (Christiansen et al. 2009; Lockhart et al. 2006; McLaughlin et al. 2004; Reagan et al., 2016; Schmidt-Clausen and Bindels 1974; Sivak et al. 1989; Theeuwes et al. 2002), and interior lighting (Bangali 2015, 2015; Lin et al. 2014).

Figure 4 shows one example of the de Boer scale in which the descriptors of glare magnitude range from just noticeable to unbearable. There are a number of problems with this version of the scale (Fotios 2015). The minimum discomfort that can be recorded is "just noticeable," which implies that

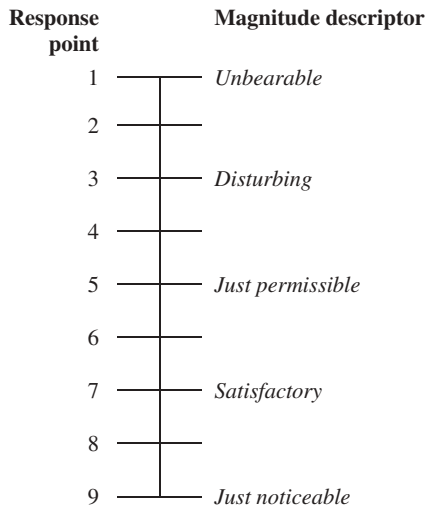


Fig. 4. The de Boer rating scale for evaluating discomfort glare.

it is just possible to perceive glare, with the remaining responses being further increasing magnitudes of discomfort: the observer is not given the option of stating that glare is not at all apparent. This situation may unintentionally force respondents to report a stimulus to be uncomfortable when that is not the case. In turn, this may lead to a datum that forces the magnitude of discomfort with other stimuli to be overestimated.

A second problem is that the magnitude descriptors are not sufficiently precise. Point 7 in Fig. 4 is labeled "satisfactory": if the stimulus is causing discomfort due to glare, then what aspect of this is satisfactory? Furthermore, what is the distinction between satisfactory and "just permissible" discomfort? Satisfactory means fulfilling expectations, acceptable but, though not outstanding or perfect, good enough for a particular purpose: permissible means something that can be permitted or allowed. These are not clear descriptors of the magnitude discomfort. Similarly, past studies have commented on the difficulty in understanding the criteria of "just perceptible" glare (Kent et al. 2016) and difficulty in distinguishing between "just imperceptible" and "perceptible" (Akashi et al. 1996).

In a study where 23 naïve test participants were required to arrange in order of magnitude five descriptors of a de Boer scale (unbearable; disturbing; just admissible; satisfactory; and unnoticeable), only seven placed satisfactory in the same location as did de Boer (that is, one step more

discomforting that just noticeable) and 15 people assumed it to be a lower level of discomfort (Gellatly and Weintraub 1990). Alleged experts performed no better, with just one of the 14 test participants matching the de Boer order. These do not suggest a consistent understanding of satisfactory glare.

A key output from discomfort glare studies is the borderline between comfort and discomfort (BCD). Hopkinson (1940) used an adjustment task, with test participants instructed to adjust the light level (of the glare source or the background) to each of the four criteria in turn: just intolerable, just uncomfortable, satisfactory, and just not perceptible. According to Hopkinson, the BCD falls between points 2 and 3; that is, the central point. In later work using a four-point rating scale with Hopkinson-like labels, the BCD is defined as point 2, not the center (Adrian and Schreuder 1970). These two studies reveal differences in experimenters' approaches: given this, we might also expect differences to exist between naïve respondents, which might influence the way in which the response scale was used.

Assumed threshold was investigated in a study of thermal comfort (Schweiker et al. 2017). The authors examined a seven-point response scale commonly used to investigate thermal sensation, with end points labeled "cold" and "hot" and the middle point labeled "neutral" (Fig. 5). Test participants were asked to indicate which of these categories encapsulated conditions that would be considered comfortable, to test the assumption that the three central categories (slightly cool, neutral, and slightly warm) denote the comfort band.

Only 12% of participants agreed with this standard assumption. The remainder suggested a variety of comfort ranges, including (as the extreme responses given) only the central point (neutral) and the five central points (cool to warm). Only the extreme categories (cold and hot) were not indicated to be comfortable by any respondent.

A related assumption within semantic differential scales such as that shown in Fig. 5 is that two contrasting adjectives are about equidistant from the neutral center point: where this assumption is not valid, the measurement is distorted (Heise 1969). This might be the case in de Boer-type scales (Fig. 4) where point 9 is labeled "unnoticeable"

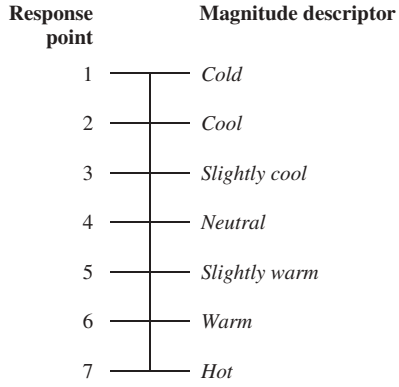


Fig. 5. A 7-point scale used to investigate thermal sensation (Schweiker and others 2016).

(Christianson et al. 2009), giving just one response point to say no discomfort but several points to express discomfort.

University students in the UK are asked to participate in an annual national survey to evaluate their learning experience (National Student Survey (NSS) 2017). The questions include the following: staff are good at explaining things; the course is intellectually stimulating; and marking and assessment has been fair. Responses are recorded using five-point scales (5 = *definitely agree*; 4 = *mostly agree*; 3 = *neither agree nor disagree*; 2 = *mostly disagree*; 1 = *definitely agree*). It appears, however, that when the responses are analyzed, the middle (neutral) category is summated as a negative opinion, despite that not being evident on the questionnaire and hence possibly not being the intention of the respondents. Such grouping may therefore lead to an underestimate of satisfaction with teaching quality.

In summary, experimenters and their respondents can differently define the category labels of a response scale, in which case there must be some doubt about the experimenter's interpretation of the findings. Such confusion may be alleviated by providing descriptions of the meaning of the response scale categories, and there are examples from discomfort glare studies where this has been done. Osterhaus and Bailey (1992) used a four-point rating scale with points labeled "imperceptible," "noticeable," "disturbing," and "intolerable." Their participants were told that the borderline between imperceptible and noticeable was the changeover point where glare discomfort would be first noticed, and they defined this criterion as equivalent to a very slight experience of discomfort

that could be tolerated for approximately one day when placed at someone else's workstation. The borderline between noticeable and disturbing glare was defined as discomfort that could be tolerated for 15 to 30 min but that would require a change in lighting conditions for any longer period. The borderline between disturbing and intolerable glare was defined as the turning point where one would no longer be able to tolerate the lighting conditions. Ngai and Boyce (2000) also gave extended descriptions to clarify the meaning of their seven-point scale. For example, just perceptible (point 2) was described as "I am aware that there is something overhead but cannot tell what it is" and uncomfortable (point 5) as "I am aware of a luminaire overhead and I would complain to my supervisor about it."

5. Checking for inattentive responding

Inattentive responding, otherwise known as *content nonresponsivity*, is responding without regard to item content (Meade and Craig 2012). It may be particularly prevalent with long and involuntary surveys, such as an office worker filling in yet another extensive questionnaire when he or she would rather be getting on with the day's work load. Inattentive respondents include the unmotivated person who ticks the same category for consecutive items or who ticks categories at random without reading the question (Huang et al. 2012). There is an example of respondents ticking categories along response scales even when there was no associated question (Piferi and Jobe 2003). Random answers increase response variance and hence a decrease in internal consistency (Fronczyk 2014) and decrease the apparent correlation between variables.

Respondents who provide the same response to a series of consecutive questions can be identified using long-string analysis, although this requires estimation of when a string does indeed become long, and that may be somewhat arbitrary (Huang et al. 2012). An alternative approach for highlighting inattentive responding is to include bogus questions within the survey. These are questions targeting a predictable response if the question is properly read, with obvious correct and incorrect responses. If the respondent chooses an incorrect response, there is little doubt that he or she is responding carelessly or dishonestly (Meade and Craig 2012). Examples of bogus questions used in a recent survey of outdoor lighting were "I have been to every country in the world" and "I am wearing clothes" (Fotios et al. 2018). Alternatively, a direct instruction can be given instead of a question, such as "For this question, please tick category 2."

Bogus questions may reveal a problem but do not reveal how to deal with that problem. Consider that analysis of a survey reveals an incorrect response to the bogus question: should that person's whole set of responses be discarded? Given that this may be a difficult decision for an experimenter, it is possible to let respondents inform the decision by asking after completion, "In your honest opinion, should we use your data in our analyses in this study? (yes/no)" (Meade and Craig 2012).

6. Range equalizing bias

Range equalizing bias describes what happens when respondents make quantitative judgemnts without knowing how responses should be mapped onto the stimuli: respondents will tend to use most of the range of responses regardless the size of the range of stimuli (Gescheider 1988; Poulton 1989). Four studies associated with lighting are used here to illustrate range bias (Fotios and Castleton 2016; Kakitsuba 2016; Simons et al. 1987; Teller et al. 2003).

Teller et al. (2003) asked test participants to evaluate brightness of a small target (2°) against the background (white computer screen, 42°). Specifically, participants had to report whether the target was brighter or dimmer than the surround. The background did not change during trials. The target changed luminance, with three ranges of luminance (high, middle, and low, observed in separate blocks) each including typically 11 target luminances (observed in random order). Critically, there was a slight overlap between the ranges such that (for example) the targets of highest luminance in the low range of luminances were also the lowest luminances of the middle. Seven test participants were used, with each making 20 discrimination judgments per condition.

Consider a particular target. When evaluated within the luminance range for which it was at the lower end, it was considered by almost 100% of participants to be dimmer than the surround.

However, when this same target was evaluated with the luminance range for which it was at the upper end, it was now considered by almost 100% of participants to be brighter than the surround. Neither the target nor background varied; only the range of other target luminances within which it was evaluated did. These results suggest that participants were not following the instruction of comparing the target with the background but judging the target relative to other targets in that range.

The target and background would be considered equally bright when 50% of responses were that the target was the brighter. The three luminance ranges gave different target luminances for this equal brightness conclusion. By using three ranges of luminance, Teller et al. (2003) reveal the effect of range bias: studies using only one range (the common approach) and not otherwise considering the effect of range bias may be providing misleading conclusions. This is discussed further elsewhere (Fotios and Houser 2013).

The second demonstration of range bias concerns evaluations of perceived safety, a focus of outdoor lighting research. Fotios and Castleton (2016) evaluated perceived safety using a five-point rating scale (1 = *very unsafe*; 2 = *somewhat unsafe*; 3 = *neutral*; 4 = *somewhat safe*; 5 = *very safe*). The scenes evaluated were the 100 photographs used in a previous study (Van Rijswijk 2016) from which they could be placed into a rank order according to the mean ratings of safety. To demonstrate range bias, the 100 photographs were divided into two subsets of 55. Set A contained the 45 photos given the higher ratings of safety and set B included the 45 photos given the lower ratings of safety. Both sets also included the 10 photographs in the center of the rank order of all 100: these 10 photographs were therefore the most-safe scenes in set B and the least-safe scenes in set A. In trials, the 55 photographs within a set were observed individually and in a random order, with a different sample of observers for the two sets. The differences in ratings between sets A and B were significant for eight of the 10 scenes (Mann-Whitney, $P < 0.05$). The common scenes received higher ratings (that is, were considered safer) when observed within set B (that is, alongside the 45 scenes considered the least safe) than when observed within set A (that is, alongside the 45 scenes considered the most safe), a median increase of 1.0.

Simons et al. (1987) carried out field surveys of road lighting in 24 residential roads, with average horizontal illuminances ranging from about 1.0 lux to 12.0 lux. A nine-point rating scale was used to rate their "overall impression" of the lighting with points labeled *very poor* (1), *poor* (3), *adequate* (5), *good* (7), and *very good* (9). Twenty years beforehand, De Boer (1961) had also carried out a field study using a similar nine-point rating scale, asking for a "general appraisal" with response points labeled *bad* (1), *inadequate* (3), *fair* (5), *good* (7), and *excellent* (9). In this study, the illuminances ranged from approximately 1.0 lux to 71 lux (assuming an average luminance coefficient of $q_0 = 0.07$ to determine illuminance from the reported luminances, which ranged from approximately 0.06 cd/m^2 to 5.0 cd/m^2), a larger stimulus range than was examined by Simons et al. (1987). Range equalizing bias is evident in both studies, with the roads of low illuminance receiving ratings near the low end of the rating scale and the roads of high illuminance rated toward the top of the rating scale.

Kakitsuba (2016) sought to validate the Kruithof effect (Kruithof 1941). As shown in Fig. 6, the Kruithof effect is alleged to identify pleasing combinations of correlated color temperature (CCT) and illuminance (the unshaded region); specifically, it suggests using lower CCT at lower illuminances and higher CCT at higher illuminances. Kakitsuba's (2016) study (among many others) was carried out despite previous studies demonstrating that the relationship did not exist (Boyce and Cuttle 1990; Davis and Ginthner 1990), despite Kruithof providing almost no information about how the relationship was established (Fotios 2017a), and despite Kruithof himself being apparently "appalled at how his tiny little thought experiment (which included Kruithof and his wife as the only two subjects) had been so widely accepted without further investigation" (Fotios 2017b).

Kakitsuba (2016) used category rating to evaluate brightness, glare, and comfort in a small office. The combinations of illuminance and CCT evaluated were chosen to define the upper and lower borders of the comfortable (unshaded) region of

the Kruithof curve (Fig. 6). At four intervals of CCT (2700 K, 3500 K, 4200 K, and 5000 K), a range of illuminances was set around the illuminance of the lower border and around the illuminance of the upper border. For example, at 4200 K, an illuminance range of 150 lux to 550 lux was used to establish the lower border and an illuminance range of 2500 lux to 4000 lux for the upper border (see Fig. 6). Within each range there were (typically) four incremental steps of illuminance, each step was evaluated using category rating, and the boundary illuminance was determined as that which would lead to a neutral rating (that is, the midpoint of the five- and seven-point response scales that were used).

This approach is unlikely to do anything but validate Kruithof (Fotios 2016b). The use of separate illuminance ranges for the upper and lower borders with no overlap guaranteed that two distinct borders would be found. The illuminance ranges for different CCT also varied (for example, for the upper border, 1500 lux to 3000 lux at 2700 K; 3500 lux to 5000 lux at 5000 K) in a manner that was destined to show a preference for higher illuminance at higher CCT. Range equalizing bias is clearly evident in these results for all combinations of illuminance and CCT (Fotios 2016b).

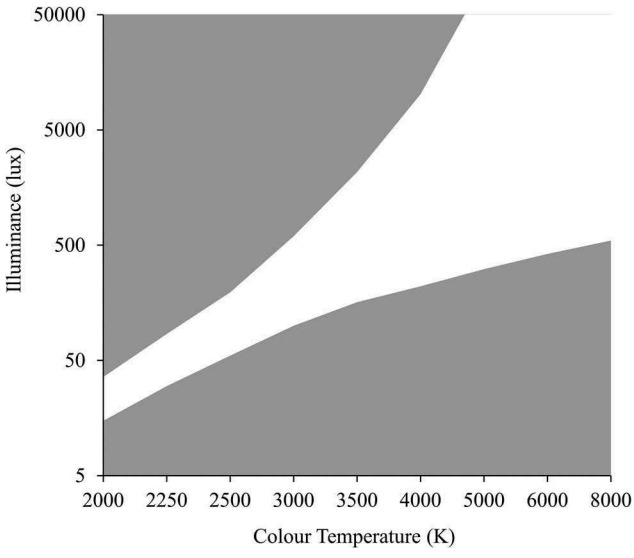


Fig. 6. The Kruithof graph (adapted from fig. 10 of Kruithof (1941)). Note: The two vertical lines at 4200 K show the range of illuminances used by Kakitsuba (2016) to establish lower and upper illuminance borders for the pleasant region.

Ratings are clearly relative to the range of stimuli. Consider two studies using seven-point semantic differential rating scales to capture spatial brightness responses: Davis and Ginthner (1990), who used three illuminances, 269, 592, and 1345 lux, and Vienot et al. (2009), who used three illuminances but of a lower level, 150, 300, and 600 lux. In these studies, a rating of 1 was a low brightness (a dim or dark environment) and 7 a high brightness. For evaluations of lighting of CCT 2700 K, a brightness rating of 4 equates to an illuminance of 200 lux in the study using the lower range of illuminances (Vienot et al. 2009) and 1200 lux in the study using a higher range of illuminances (Davis and Ginthner 1990). Within any given study, the rating responses reveal the relative magnitude of response to different stimuli: without supplementary data, this should not be extended to an absolute evaluation.

Though the results of separate evaluations using category rating are prone to range bias, this does not mean that category rating should not be used but rather that some thought is required before doing so. Recall first the influence of range bias on ratings of perceived safety. The evaluation of a scene is made relative to the other scenes evaluated, a relative judgment and not the absolute judgment implied by the test instruction. This means that roads with higher illuminances tend to be rated as safer than roads of lower illuminance, regardless of what these higher and lower illuminances are (Fotios 2016c; Fotios and Castleton 2016). This is a trivial outcome because regardless of the illuminance recommended by one study to provide for good safety, a second study using a still higher illuminance will now recommend that to be the optimum: these data do not converge towards an asymptote.

It may be possible to overcome range bias; an example of this is the day–dark approach first used by Boyce et al. (2000). Though the standard approach to evaluating road lighting is to carry out the evaluations after dark, Boyce et al. (2000) also evaluated the same locations at daytime and used the difference between the daytime and after-dark ratings in analysis of the effect of changes in illuminance. This approach did lead toward an asymptote. Boyce considered this study to be one of the two projects of which he was most proud during his career (Bright Lights 2016).

To show the impact of this, consider one recent study (Peña-García et al. 2015) in which evaluation of road lighting was carried out only after dark; it was concluded that lighting on the roads lit to 50 lux horizontal illuminance was perceived as safer than the roads lit to 15 lux or 25 lux, hinting at the need for higher light levels than the range (2 lux to 15 lux) currently recommended for pedestrians (CIE 2010). In contrast, two independent studies carried out using the day–dark approach found that average horizontal illuminances of 7 lux (Sheffield, UK) and 10 lux (Rome, Italy) were associated with a day–dark difference of 0.5 units of a 1–6 rating scale (Fotios et al. 2017).

7. Discussion

An experimenter needs to uncover the clues that lead to a test participant's response, not what he or she *thinks* the clues are (Feynman 1974). The study of methods is an effort to discover all of the things one has to do to discover something about (in this case lighting). Methods research is, however, often overlooked because these studies do not say anything about lighting; they do not reveal exciting new effects of lighting, the findings that lead to prestige and funding, but instead might show that apparent exciting new effects of lighting are false.

This article has discussed category rating, a procedure in which test participants give subjective estimates of the magnitude of a parameter. In many cases there are reasons to suspect that the evaluations given are speculation about something of which the respondent has no experience or no interest; are not responding to the question the experimenter thinks they have posed; and are influenced by context (range equalizing bias). According to Poulton (1977), quantitative subjective assessments are almost always biased, sometimes completely misleading. Category rating may be particularly so: in Poulton's (1989) order of preference, category rating was placed below discriminating and matching judgments (although above magnitude judgments). Boynton and Greenhalgh (2004) suggest that "no single method has been so abused" and, furthermore, that "inappropriate instruments and lack of rigour inevitably lead to poor quality data, misleading conclusions, and woolly recommendations." Note also the opinion of others involved in lighting research: "Semantic differential scaling experiments are meaningless by themselves, but can serve as the critical first step in developing reasonable hypotheses about proposed higher order phenomena" (Tiller and Rea 1992).

Where subjective evaluations are sought it is recommended that two different procedures be used (CIE 2014)—if the results converge, this gives some confidence that the data are robust. Though it is not commonly reported in the lighting literature, some studies do so, including studies of spatial brightness (Boyce 1977; Fotios and Cheal 2007a, 2011; Fotios et al. 2015b; Han and Boyce 2003; Houser et al. 2003; Houser and Tiller 2003; Vrabel et al. 1998) and experiments related to discomfort and distraction associated with glare (Ngai and Boyce 2000; Osterhaus and Bailey 1992; Ramasoot and Fotios 2012).

Though this article has raised criticism of category rating, that does not mean that category rating should be avoided but rather that researchers should take care when designing experiments and should be sceptical when considering conclusions drawn from category rating. Such scepticism should also be applied to the conclusions drawn from experiments using matching (Fotios 2001; Fotios and Cheal 2007b; Fotios, Houser and Cheal 2008), adjustment (Fotios and Cheal 2010; Kent et al. 2017; Logadóttir et al. 2011, 2013; Uttley et al. 2013) and discrimination procedures (Fotios and Houser 2013; Teller et al. 2003). Note also that this article does not claim to provide an exhaustive review of the category rating procedure: the literature raises many other questions (Annett 2002; Brink et al. 2016; Gescheider 1988; Gohardoust Monfared 2012; Heise 1969; Hyvärine 2015; Lietz 2010; Tourangeau et al. 2000; Weisberg 2005).

To highlight the impact of experimental design, consider the results of two experiments. It was suggested (above) that the road lighting evaluations conducted by Simons et al. (1987) were affected by range equalizing bias. The data of Simons and others are of interest because they were the basis for the light levels of the three lighting classes in the 1992 issuance of BS5489 (British Standards Institution 1992): Horizontal illuminances of 10.0, 5.0, and 2.5 lux were proposed, because these corresponded to ratings of *good* (7), *adequate* (5), and *poor to adequate* (4),

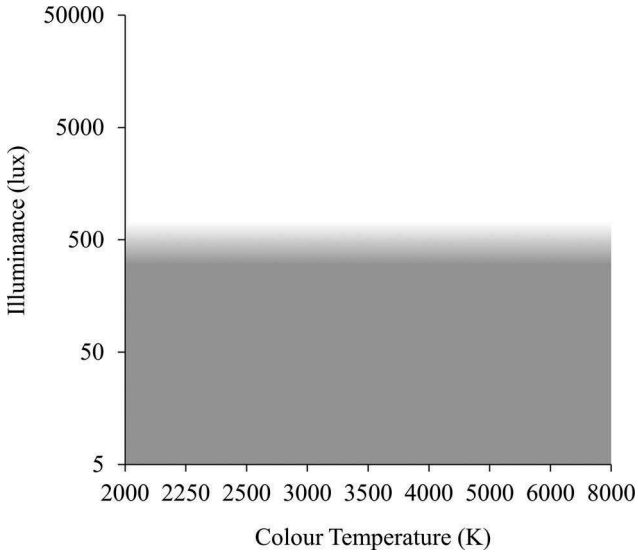


Fig. 7. The Kruithof graph revised according to the results of credible studies (Fotios 2017a). The shaded region represents conditions likely to be considered unpleasant, and the clear region above suggests conditions likely to be acceptable. The credible studies evaluated suggest that the transition from likely acceptable to likely unpleasant conditions is in the range of (approximately) 300 lux to 500 lux. Low illuminances (less than approximately 300 lux) may be perceived as unpleasant; an illuminance of 500 lux is sufficient to provide a pleasant environment and a further increase in illuminance above 500 lux is of little benefit.

respectively. Had instead De Boer's (1961) results been used to establish suitable illuminances, then good lighting would have been set at approximately 20 lux and a different illuminance would have been recommended in the British standard. Next consider the Kruithof curve (Fig. 6). It was suggested (above) that Kakitsuba's (2016) experiment did not provide a meaningful validation of Kruithof. A subsequent review (Fotios 2017a) of studies investigating the pleasantness of lighting under different combinations of illuminance and CCT, which rejected those studies not meeting recommendations for best practice (CIE 2014), suggested an relationship entirely different from that proposed by Kruithof as shown in Fig. 7.

8. Conclusion

When reviewing an experiment, as a reader of other studies or one's own work, a question that should be asked is, "Do I believe these results?" A bias-free experiment is unlikely, but with care the direction and magnitude of bias may be better understood. This article has discussed the category rating procedure. For category rating, the precautions that deserve consideration include defining and/or checking understanding of the meaning of question items and response scale labels. For example, if a question has asked for visual clarity to be evaluated, then ask also for a definition of visual clarity. Respondents should have the option to state that they have no opinion about an item. Null condition trials should be included to test whether the experiment yields an unexpected difference, and similarly extreme test conditions might be included to check that the experiment reveals a likely effect. If the outcome sought is an absolute magnitude, such as the luminance associated with a specific level of discomfort from glare, then it would be useful to confirm that the same absolute magnitude would be gained by repeating the experiment with a different range of test luminances.

Acknowledgment

This article benefited from constructive comments on a draft from Kynthia Chamilothori of the Laboratory of Integrated Performance in Design (LIPiD) group at EPFL Lausanne.

Disclosure statement

The author reports no conflict of interest.

Funding

The authors reported no funding.

ORCID

Steve Fotios http://orcid.org/0000-0002-2410-7641

References

[CIE] Commission Internationale de l'Eclairage. 1988. Brightness luminance relations: Classified bibliography. Vienna (Austria): CIE. Publication No: CIE 078-1998.

[CIE] Commission Internationale de l'Eclairage. 2010. Lighting of roads for motor and pedestrian traffic. Vienna (Austria): CIE. Publication No: CIE 115:2010.

[CIE] Commission Internationale de l'Eclairage. 2014. Guidance towards best practice in psychophysical procedures used when measuring relative spatial brightness. Vienna (Austria): CIE. Publication No: CIE 212:2014.

[CIE] Commission Internationale de l'Eclairage. 2016. ILV: International lighting vocabulary, 2nd edition. Vienna (Austria): CIE. Publication No: CIE DIS 017/E:2016.

Acuña-Rivera M, Uzzell D, Brown J. 2011. Perceptions of disorder, risk and safety: The method and framing effects. Fundación Infancia y Aprendizaje Psyecology: revista Bilingüe de Psicología Ambiental. Bilingual J Environ Psychol. 2(2):167–77.

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

Akashi Y, Muramatsu R, Kanaya S. 1996. Unified glare rating (UGR) and subjective appraisal of discomfort glare. Lighting Res Technol. 28(4):199–206.

Alfonso-Reese LA. 2001. Technique for estimating perceptual noise in categorization tasks. Behavior research methods. Instr Comput. 33(4):489–95.

Annett J. 2002. Subjective rating scales: science or art? Ergonomics. 45(14):966–87.

Armstrong SW, Algozzine B, Sherry L. 1979. Perceived brightness and classroom interactions. Educ Res Q. 4:54–60.

Aston SM, Bellchambers HE. 1969. Illumination, colour rendering and visual clarity. Lighting Res Technol. 1(4):259–61.

Bangali J. 2015. Evaluation of discomfort glare by using lighting simulation software for optimal designing of indoor illumination systems. Int J Emerging Eng Res Technol. 3 (12):173–78.

Bellchambers HE, Godby AC. 1972. Illuminance, colour rendering and visual clarity. Lighting Res Technol. 4(2):104–06.

Berman SM, Jewett DL, Fein G, Saika G, Ashford F. 1990. Photopic luminance does not always predict perceived room brightness. Lighting Res Technol. 22(1):37–41.

Boyce PR. 1977. Investigations of the subjective balance between illuminance and lamp colour properties. Lighting Res Technol. 9:11–24.

Boyce PR. 2014. Human factors in lighting. 3rd ed. Boca Raton (FL): CRC Press.

Boyce PR, Cuttle C. 1990. Effect of correlated colour temperature on the perception of interiors and colour discrimination. Lighting Res Technol. 22(1):19–36.

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

Boynton PM, Greenhalgh T. 2004. Selecting, designing, and developing your questionnaire. BMJ. 328:1312–15.

Bright Lights. 2016. Discussion one: Kynthia and Peter (22/09/2016). [accessed 2017 Jan 6]. http://brightlights-research.blogspot.co.uk/.

Brink M, Schreckenberg D, Vienneau D, Cajochen C, Wunderli JM, Probst-Hensch N, Röösli M. 2016. Effects of scale, question location, order of response alternatives, and season on self-reported noise annoyance using ICBEN scales: A field experiment. Int J Environ Res Public Health. 13(11):1163.

British Standards Institution. 1992. BS5489, road lighting: Part 3, code of practice for lighting for subsidiary roads and associated pedestrian areas. London (UK): BSI.

Bullough J, Brons JA, Qi R, Rea MS. 2008. Predicting discomfort glare from outdoor lighting installations. Lighting Res Technol. 40:225–42.

Christianson KB, Greenhouse DS, Barton JE, Chow C 2009. Methods to Address Headlight Glare. California PATH Research Report UCB-ITS-PRR-2009-20. California Partners For Advanced Transit And Highways.

Davis RG, Ginthner DN. 1990. Correlated color temperature, illuminance level and the Kruithof curve. J Illuminating Eng Soc Winter. 19:27–38.

De Boer JB. 1961. The application of sodium lamps to public lighting. Illum Eng. 56(4):293–312.

De Boer JB, Schreuder DA. 1967. Glare as a criterion for quality in street lighting. Trans Illuminating Eng Soc. 32 (2):117–35.

DeLaney WB, Hughes PC, McNelis JF, Sarver JF, Soules TF. 1978. An examination of visual clarity with high colour rendering fluorescent light sources. J Illuminating Eng Soc. 7(2):74–84.

Feynman RP. 1974. Cargo cult science: some remarks on science, pseudoscience, and learning how to not fool yourself. Caltech's 1974 Commencement address. Eng Sci. June:10–13.

Flynn JE, Spencer TJ. 1977. The effects of light source colour on user impression and satisfaction. J Illuminating Eng Soc. 6:167–79.

Flynn JE, Spencer TJ, Martyniuk O, Hendrick C. 1973. Interim study of procedures for investigating the effect of light on impression and behaviour. J Illuminating Eng Soc. 3(1):87–94.

Fotios S. 2001. An error in brightness matching associated with the application of dimming. Lighting Res Technol. 33 (4):223–31.

Fotios S. 2015. Research Note: Uncertainty in subjective evaluation of discomfort glare. Lighting Res Technol. 47 (3):379–83.

Fotios S. 2016a. The road less travelled. Lighting J. 81(5):20–23. May.

Fotios S. 2016b. Review of a published article (Kakitsuba N. Comfortable Indoor Lighting Conditions Evaluated from Psychological and Physiological Responses). Leukos. 12 (3):173–77.

Fotios S. 2016c. Comment on of empirical evidence for the design of public lighting. Saf Sci. 86:88–91.

Fotios S. 2017a. A revised Kruithof graph based on empirical data. Leukos. 13(1):3–17.

Fotios S. 2017b. Response to comments from K. Cuttle. Leukos. 13(1):19–22.

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):261–81.

Fotios S, Atli D, Cheal C, Hara N. 2015b. Lamp spectrum and spatial brightness at photopic levels: investigating

prediction using S/P ratio and gamut area. Lighting Res Technol. 47(5):595–612.
Fotios S, Atli D, Cheal C, Houser K, Logadóttir A. 2015a. 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):235–43.
Fotios S, Cheal C. 2007a. Lighting for subsidiary streets: investigation of lamps of different SPD. Part 2 – brightness. Lighting Res Technol. 39(3):233–52.
Fotios S, Cheal C. 2007b. Evidence for response contraction bias in side-by-side matching tasks. Lighting Res Technol. 39(2):159–69.
Fotios S, Cheal C. 2010. Stimulus range bias explains the outcome of preferred-illuminance adjustments. Lighting Res Technol. 42(4):433–47.
Fotios S, Cheal C. 2011. Predicting lamp spectrum effects at mesopic levels. part 1: spatial brightness. Lighting Res Technol. 43(2):143–57.
Fotios S, Gado T. 2005. A comparison of visual objectives used in side-by-side matching tests. Lighting Res Technol. 37(2):117–31.
Fotios S, Houser K. 2013. Using forced choice discrimination to measure the perceptual response to light of different characteristics. Leukos. 9(4):245–59.
Fotios S, Houser KW. 2009. Research methods to avoid bias in categorical ratings of brightness. Leukos. 5(3):167–81.
Fotios SA, Houser KW, Cheal C. 2008. Counterbalancing needed to avoid bias in side-by-side brightness matching tasks. Leukos. 4(4):207–23.
Fotios S, Liachenko Monteiro A, Uttley J. 2018. Evaluation of pedestrian reassurance gained by higher illuminances in residential streets using the day-dark approach. Lighting Res Technol. Online First. doi:10.1177/1477153518775464.
Fotios S, Uttley J, Liachenko-Monteiro A, Mattoni B, Bisegna F 2017. Field surveys of reassurance in two European cities using Boyce's day-dark approach. Proceedings of CIE conference; Jeju Island (Korea), October 2017.
Foulsham T. 2018. Tweet posted 09/ 01/2018. https://twitter.com/tomfoulsh/status/950717308829388800.
Fronczyk K. 2014. The identification of random or careless responding in questionnaires: the example of the neo-ffi22. Roczniki Psychologiczne/Ann Psychol. XVII(2):457–73.
Gellatly and Weintraub. 1990. User reconfigurations of the de boer rating scale for discomfort glare. Ann Arbor (MI): The University of Michigan Transportation Research Institute.
Gescheider GA. 1988. Psychophysical scaling. Annu Rev Psychol. 39:169–200.
Gohardoust Monfared I 2012. Importance of scale format, respondents attitude, and temporal effects in post-occupancy evaluation surveys. PhD thesis, University of Sheffield, the School of Architecture.
Han S, Boyce PR. 2003. Illuminance, CCT, décor and the Kruithof curve. 25th Session of the CIE, San Diego, 25 June to 2 July 2003 1(2):D3 282–285. (see also Han S. 2002. Effect of illuminance, CCT and décor on the perception of lighting. MS in Lighting thesis, Rensselaer Polytechnic Institute, Troy, NY).
Hashimoto K, Nayatani Y. 1994. Visual clarity and feeling of contrast. Color Res Appl. 19(3):171–85.
Hashimoto K, Yano T, Nayatani Y. 2000. Proposal of practical method for calculating and indexing feeling of contrast for light source. J Illuminating Eng Inst Japan. 84 (11):843–50.
Heise DR. 1969. Some methodological issues in semantic differential research. Psychol Bull. 72(6):406–22.
Hopkinson RG. 1940. Discomfort glare in lighted streets. Trans Illuminating Eng Soc. 5:1–32.
Houser KW, Tiller DK. 2003. Measuring the subjective response to interior lighting: paired comparisons and semantic differential scaling. Lighting Res Technol. 35(3):183–98.
Houser KW, Tiller DK, Hu X 2003. Prototype demonstration of vision-tuned fluorescent lamps, EISG Final Report for CEC.
Huang JL, Curran PG, Keeney J, Poposki EM, DeShon RP. 2012. Detecting and deterring insufficient effort responding to surveys. J Bus Psychol. 27(1):99–114.
Hyvärine M 2015. Methodological questions in lighting acceptance and preference studies. Doctoral dissertation, Aalto University, Department of Electrical Engineering and Automation.
Kakitsuba N. 2016. Comfortable indoor lighting conditions evaluated from psychological and physiological responses. Leukos. 12:163–72.
Kato K, Nagao T, Yamanaka T, Kawai K, Sakakibara K-I 2010. Study on effect of room acoustics on timbral brightness of clarinet tones. Part II: An acoustic interpretation and synthesis of analytical results. Proceedings of 20th International Congress on Acoustics, ICA 2010 23-27 August 2010; Sydney (Australia).
Kent M, Fotios S, Altomonte S. 2017. Discomfort glare evaluation: The influence of anchor bias in luminance adjustments. Lighting Res Technol. First Published October 13, 2017. doi:10.1177/1477153517734280.
Kent MG, Altomonte S, Tregenza PR, Wilson R. 2016. Discomfort glare and time of day. Lighting Res Technol. 47(6):641–57.
Kruithof AA. 1941. Tubular fluorescent lamps for general illumination. Philips Tech Rev. 6(3):65–73.
LaBoeuf RA, Shafir E. 2006. The long and short of it: physical anchoring effects. J Behavioural Decis Making. 19:393–406.
Lietz P. 2010. Research into questionnaire design: A summary of the literature. Int J Market Res. 52(2):249–72.
Lin Y, Fotios S, Wei M, Liu Y, Guo W, Sun Y. 2015. Eye movement and pupil size constriction under discomfort glare. Investigative Ophthalmol Visual Sci. 56(3):1649–56.
Lin Y, Liu Y, Sun Y, Zhu X, Lai J, Heynderickx I. 2014. Model predicting discomfort glare caused by LED road lights. Opt Express. 22:18056–71.
Lockhart TE, Atsumi B, Ghosh A, Mekaroonreung H, Spaulding J. 2006. Effects of planar and non-planar driver-side mirrors on age-related discomfort-glare responses. Saf Sci. 44(3):187–95.

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):403–22.

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. Colour Res Appl. 38(1):46–57.

Lynes JA. 1996. Daylight and photometric anomalies. Lighting Res Technol. 28(2):63–67.

McLaughlin S, Hankey J, Green CA, Larsen M. 2004. Discomfort glare ratings of swivelling HID headlamps. SAE Int. 2004-01-2257.

Meade AW, Craig SB. 2012. Identifying careless responses in survey data. Psychol Met. 17(3):437–55.

National Student Survey (NSS). 2017. (accessed 2017 Aug 2). http://www.thestudentsurvey.com/.

Neighbourhood Watch and the Suzy Lamplugh Trust (NW and SLT). 2013. Street Lighting and Perceptions Of Safety Survey: results And Analysis. November. (accessed 2017 Aug 2). http://www.ourwatch.org.uk/uploads/pub_res/Perceptions_of_Safety_survey_FINAL.pdf.

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

Osgood CE, Suci GJ, Tannenbaum PH. 1975. The measurement of meaning. 9th printing. Originally published in 1957. Chicago (IL): University of Illinois Press. ISBN 0-252-74539-6.

Osterhaus WKE, Bailey IL 1992. Large area glare sources and their effect on discomfort and visual performance at computer work stations. Proceedings of the 1992 IEEE Industry Applications Society Annual Meeting, 4-9 October 1992; Houston (TX).

Peña-García A, Hurtado A, Aguilar-Luzón MC. 2015. Impact of public lighting on pedestrians' perception of safety and well-being. Saf Sci. 78:142–48.

Piferi RL, Jobe RL. 2003. Identifying careless responses in questionnaire research: a serendipitous finding. Psychol Rep. 93:320–22.

Poulton EC. 1977. Quantitative subjective assessments are almost always biased, sometimes completely misleading. Br J Psychol. 68:409–25.

Poulton EC. 1989. Bias in quantifying judgements. London (UK): Lawrence Erlbaum Associates Ltd.

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

Ramsay M, Newton R 1991. The effect of better street lighting on crime and fear: A review. Crime Prevention Unit paper No. 29. London (UK): The Home Office.

Rea MS. 1982. Calibration of subjective scaling response. Lighting Res Technol. 14(3):121–29.

Reagan IJ, Frischmann T, Brumbelow ML. 2016. Test track evaluation of headlight glare associated with adaptive curve HID, fixed HID, and fixed halogen low beam headlights. Ergonomics. 59(12):1586–95.

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

Schweiker M, Fuchs X, Becker S, Shukuya M, Dovjak M, Hawighorst MKolarik J. 2017. Challenging the assumptions for thermal sensation scales. Building Research & Information. 45(5):572–89.

Simons RH, Hargroves RA, Pollard NE, Simpson MD. 1987. Lighting criteria for residential roads and areas. Venice (Austria): CIE; p. 274–77.

Sivak M, Flannagan M, Ensing M, Simmons CJ. 1989. Discomfort glare is task dependent. UMTRI-89-27. Ann Arbor (MI): University of Michigan Transportation Research Institute.

Song M, Kim Y-H. 2009. Mathematical characterization of the acoustic brightness and contrast control problem in complex vector space. J Acoust Soc Am. 125:2538.

Sutton RM, Farrall S. 2005. Gender, socially desirable responding and the fear of crime. Br J Criminol. 45:212–24.

Tashiro T, Kawanobe S, Kimura-Minoda T, Kohko S, Ishikawa T, Ayama M. 2014. Discomfort glare for white LED light sources with different spatial arrangements. Lighting Res Technol. Published online 24 April 2014. doi:10.1177/1477153514532122.

Teller DY, Pereverzeva M, Civan AL. 2003. Adult brightness vs. luminance as models of infant photometry: variability, biasability, and spectral characteristics for two age groups favour the luminance model. J Vis. 3:333–46.

Theeuwes J, Alferdinck JWAW, Perel M. 2002. Relation between glare and driving performance. Hum Factors. 44:95–107.

Thornton WA, Chen E. 1978. What is visual clarity? J Illuminating Eng Soc. 7:85–94.

Tiller DK, Rea MS. 1992. Semantic differential scaling: prospects in lighting research. Lighting Res Technol. 24(1):43–52.

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

Uttley J, Fotios S, Cheal C. 2013. Satisfaction and illuminances set with user-controlled lighting. Archit Sci Rev. 56(4):306–14.

Van Rijswijk L 2016. Shedding light on safety perceptions: Environmental information processing and the role of lighting. PhD thesis, Eindhoven University of Technology.

Vienot F, Durand M-L, Mahler E. 2009. Kruithof 's rule revisited using LED illumination. J Mod Opt. 56 (13):1433–46.

Villa C, Bremond R, Saint-Jacques E. 2017. Assessment of pedestrian discomfort glare from urban LED lighting. Lighting Res Technol. 49(2):945–63.

Vrabel PL, Bernecker CA, Mistrick RG. 1998. Visual performance and visual clarity under electric light sources: part II - Visual Clarity. J Illuminating Eng Soc. 27(1):29–41.

Weisberg HF. 2005. The Total Survey Error Approach. London (UK): The University of Chicago Press.

Worthey JA. 1985. An analytical visual clarity experiment. J Illuminating Eng Soc. 15(1):239–51.