Article:

Fotios, S. (2016) Review of a published article (Kakitsuba N. Comfortable Indoor Lighting Conditions Evaluated from Psychological and Physiological Responses.). Leukos, 12 (3). pp. 173-177. ISSN 1550-2724

https://doi.org/10.1080/15502724.2016.1147293

Review of a published article

(Kakitsuba N. Comfortable Indoor Lighting Conditions Evaluated from Psychological and Physiological Responses. Leukos. Online publication doi=10.1080/15502724.2015.1061945)

Published: Leukos 2016; 12(3); 173-177.

Steve Fotios
University of Sheffield, UK.

This note is a critical review of the spatial brightness experiment reported by Kakitsuba [Kakitsuba 2015]. There are several reasons why I think the study is flawed and therefore that the results should not be considered credible. In particular, the results can largely be explained by a stimulus range bias, the ‘boundary’ illuminances tending to lie near the centre of each range of illuminances reported. Therefore the results are a product of the illuminance ranges chosen by the experimenter and do not indicate observers’ preferences for light level.

Rating Scale Uncertainty

In this work, observers used category rating to describe brightness, glare and comfort whilst seated in a small office. The office was lit by lighting from five types of lamp, these characterised by CCT, with each lamp type presented separately at a range of steps in illuminance.

Evaluations were gathered using category rating scales, purporting to measure brightness, glare and comfort, and these scales are shown in Kakitsuba’s Table 2, reproduced here as Table 1. For brightness a 7-point response range was used, with 5-point scales for glare and comfort. A first question is why were these specific scales used? This is a question that should be asked in all studies using category rating because it can affect the outcome [Bishop 1987, Dawes 2008, Moors 2008]: although some types of evaluation may not be significantly affected [Parducci and Perrett 1971] it should at least be considered. From consideration of response bias, Poulton [1989] would suggest avoiding a middle category, so the 7-point brightness scale could instead be 8-point or 6-point. For this study, a specific question is why was there a different response range for brightness than for glare and comfort? That is not a common approach and may have led to unintentional differences in response.

Each point in a scale points was given a label to describe the magnitude. These meanings of these descriptors may have caused confusion and in particular it is not clear whether the scales are bi-polar, where the middle point would be the response for a desirable environment, or a uni-polar scale where an endpoint would be the response for the desirable environment. For brightness, the scale ranged from very bright to very dark with the middle point labelled neutral. What is neutral brightness? Is neutral the desired level of brightness, and if so how does this compared with 'bright'?

Lighting design should aim for a lit environment that is apparently 'light' [Loe and others 1994] (and for light, here read '*bright'*, which was the response scale used by Loe and others] which may suggest that bright or very bright is the desirable level of light. Alternatively, did 'very bright' indicate lighting that was excessively bright, for example in the context of the onset of migraine headaches [Burks 1994]. Imagine that your current environment is the perfect (for you) level of brightness: where on the scale shown in Table 1 would you place it?

The situation for glare and comfort is no better. For comfort, what does neutral comfort mean, and is it better, worse, or the same as 'comfortable'? For glare, 'almost no glare' implies that glare was still present but that it was only just noticeable. For 'slight glare', the definition is almost identical, glare is present but only just noticeable, so how did participants choose between these labels? What level of glare is represented by the term 'neutral' – is it the complete absence of glare or some magnitude of tolerable glare? These points are raised from the viewpoint of an experimenter: we know that experimenters do not always agree in their understanding of even commonly-used terms and we should therefore expect greater confusion between an experimenter and naïve test participants [Fotios and Atli 2012]. As stated over 30 years ago at least, " … *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], and an example of this is given by Houser and Tiller [2003]. To improve understanding of how the scales were used an experimenter might define the scales beforehand (e.g. visual demonstrations or verbal definitions of variations in brightness, glare and comfort) or alternatively the experimenter might subsequently ask participants to describe in their own words what they understood by the response items and response scale labels. Neither approach is mentioned in the article.

A related question is that of scale polarity; the 5-point glare and comfort scales used by Kakitsuba may have reversed polarities, with 5 being the ideal rating for glare and 1 the ideal rating for comfort, but that is not stated to have been a purposeful decision. Consideration of response scale direction is important. Reversal might be a strategy for countering pattern response ticking (although that is not a certain advantage because failure of the respondents to read the scales may increase the likelihood of error). It may counter the primacy effect of choosing the first option encountered that feels about right, demonstrated to be a significant bias in studies where magnitude direction has been reversed [Chan 1991]. Finally, repeating questions but with response scales of reversed direction offers a method for checking response reliability [Litwin 1995].

If asked to state desirable conditions using the scales in Table 1, I would like my office to be very bright (1), with no glare (5) and comfortable (1). Is that what the experimenter intended? I do not think it is because the middle point (neutral) of each scale was used to define boundary levels (see below). How did test participants tend to use the scales shown in Table 1? We cannot know and this leads to further uncertainty when interpreting the results.

I suspect that the original study was carried out using Japanese language, translated to English for publication, and that this translation may have introduced uncertainty into the scales that was not present in the original version. It was for this potential loss in translation that the issue was raised in the CIE report on spatial brightness methodology [CIE 2014].

Response point	**Brightness**	**Glare**	**Comfort**
1	Very bright	Glare	Comfortable
2	Bright	Slight glare	Slightly comfortable
3	Slightly bright	Neutral	Neutral
4	Neutral	Almost no glare	Slightly uncomfortable
5	Slightly dark	No glare	Uncomfortable
6	Dark		
7	Very dark		

Table 1. Response scales used to measure subjective evaluation of the environment.

Identifying Boundary Illuminance

The rating scales were used to define the boundary illuminances but Kakitsuba presents conflicting statements of how this was done. In section 2.1 it is stated that "The low boundary *illuminance was identified as when the subjects voted 'neutral' for brightness and comfort, and the high boundary illuminance was identified as when the subjects voted 'neutral' for the* glare and comfort sensation". According to this definition, the comfort rating was essential to estimates for both boundary illuminances. However, the note under Kakitsuba's Table 2 states that "Brightness and glare sensation preferences were recorded to determine the low and high boundaries, respectively. The comfort preference was supplementary information.” According to this definition, comfort ratings played no part in estimates for the boundary illuminances. We are left confused because the author has given two conflicting statements as to how the results were employed to identify boundary illuminances.

Stimulus Range Bias

The choice of illuminance ranges is baffling, there is certainly no obvious rationale. There were lamps of five different CCT and for each CCT there were Low (L) and High (H) conditions, these intended to reveal the upper and lower boundaries of a Kruithof-like graph of comfort. Overall this gave 9 test conditions and the ranges are shown in Table 2 (for 3000K there is a Low condition but not a high condition; this absence is not explained). There is no apparent consistency in the illuminance ranges used in each of the 9 test conditions.

Test condition*	**Range limits (lux)**		**Step increment**	**Number of steps**	**Midpoint of range**	**Mean boundary illuminance (lux)**
	Lower setting	**Upper setting**				
2700L	20	170	50	4	95	104
2700H	1500	3000	500	4	2250	2467
3000L	50	200	100	2.5	125	134
3500L	100	200	50	3	150	161
3500H	1200	1800	200	4	1500	1490
4200L	150	550	100	5	350	463
4200H	2500	4000	500	4	3250	2965
5000L	200	500	100	4	350	309
5000H	3500	5000	500	4	4250	4147

Table 2. Summary of illuminance settings used in trials and the resultant boundary illuminance.

*Test condition: the four initial digits denote CCT and the final character denotes whether this condition was used to define the lower (L) or higher (H) boundary of a Kruithof-like comfort graph.

Different ranges were used for estimating the Low and High boundaries, tending to be around 50 to 500 lux for the low range and 1500 to 5000 lux for the high range. This means test participants were forced to give different responses, a low illuminance for the Low boundary range and a high illuminance for the High boundary range. The apparent upper and lower boundary curves in Kakitsuba’s ‘belt-like’ visual comfort graph are therefore entirely false, the test participants being forced to indicate different illuminances in the trials used to estimate upper and lower boundaries that might not really exist.

Within the Low and High conditions, there is no consistency in illuminance ranges employed with the different levels of CCT. Consider the Low conditions: for 2700K the range was 20 to 170 lux, but for 5000K the range was 200 to 500 lux. The ranges are entirely different with not even a small amount of overlap. It is not possible to associate a particular response with the same illuminance in these two test conditions. Consider the High conditions: for 2700K the range was 1500 to 3000 lux, but for 5000K the range was 3500 to 5000 lux. Again it is not possible to associate a particular response point with the same illuminance in these two test conditions: if I consider 2000 lux to be my ideal illuminance for comfort I can say so in the 2700K range but cannot indicate anything lower than 3500 lux in the 5000K range. Thus with these ranges test participants were forced to respond in a manner that indicates an interaction between CCT and illuminance.

Consideration of the haphazard selection of illuminance ranges explains why the results falsely indicate variations in preferred illuminance associated with CCT and with supposed lower and upper boundaries. It can also be shown that within a given test condition the result, the illuminance for brightness, glare and comfort associated with the 'neutral' rating, is a false impression of environmental evaluation because it can be explained by consideration of stimulus range bias.

Stimulus range bias means that the stimulus range (here the range of illuminances observed) is mapped to fit the response range (here the 5- and 7-point rating scales). In the current work the neutral category in these scales, the middle point of each range, was used to identify the boundary illuminance. Mapping the stimulus range to the response scale would mean the mean neutral illuminance is near the centre of the range, regardless of the extent of that range, a centering bias. This is indeed what appears to have happened as is shown in Figure 1. Plotting these data as a scatter graph instead of a histogram indicates a strong linear relationship (r^2=0.99, n=9). Therefore, the boundary illuminances reported by Kakitsuba and used to define the upper and lower limits of comfort do not actually represent the visual preference of his test participants, but do little more than mark the middle point of the range of light levels shown to the test participants.

The data used in Figure 1 are the middle point of each of the 9 illuminance ranges, determined from the range limits defined by Kakitsuba, and the mean of the morning and afternoon estimates of boundary illuminance. Similar conclusions were drawn when considering the morning and afternoon estimates separately or considering the low and high boundaries separately. Over the 9 combinations of CCT and Low/High boundary, the mean ‘neutral’ illuminance lies at 0.53 in the available range, where 0.50 would be the mid-point, again demonstrating centering bias.

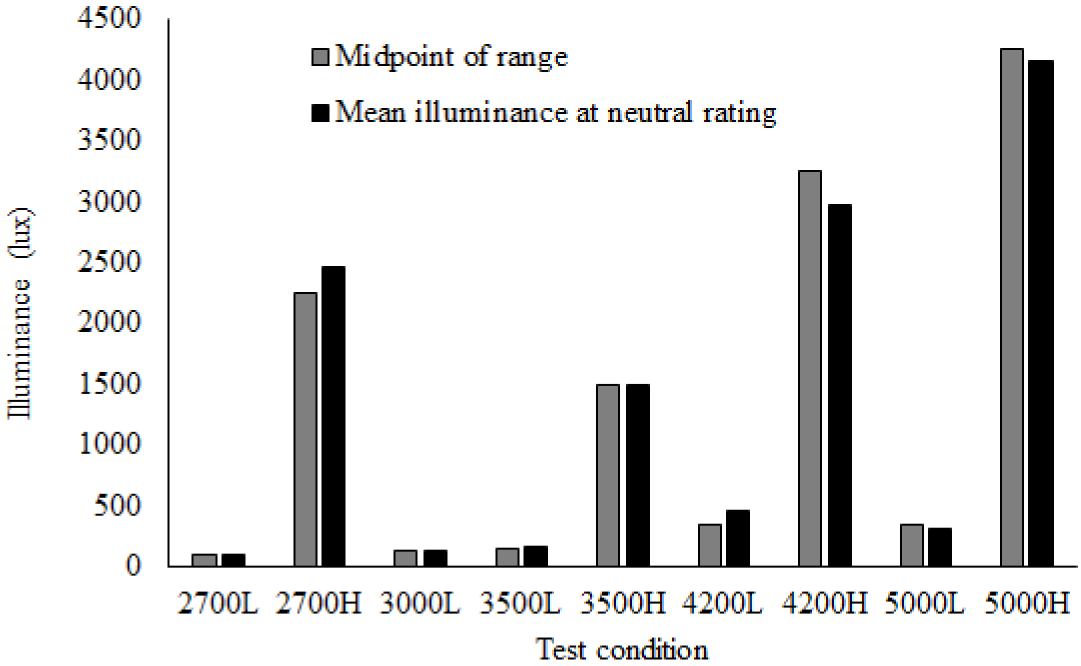

Figure 1. Demonstration of stimulus range bias: the mean ‘neutral’ illuminance in each test condition (i.e. the reported test result) is very similar to the mid-point of the range of illuminances used in that test condition. Alternatively plotting neutral illuminance against range mid-point illuminance suggests a strong linear association (r^2=0.99, n=9).

Kakitsuba reports a pilot study apparently used to inform the choice of illuminance ranges. That the pilot study itself used inconsistent ranges of illuminances, and that there is no discussion of how the pilot study results were used to guide the ranges used in the main study, means that reporting of the pilot study is meaningless.

Bias associated with stimulus range is not unexpected, having been previously demonstrated in visual evaluation experiments using an adjustment task [CIE 2014, Fotios and Cheal 2010, Logadóttir and others 2011, 2013, Uttley and others 2013].

Illuminances were presented in a stepwise procedure: starting from the lowest illuminance in a range, the illuminance was increased in fixed steps towards the highest illuminance with category ratings recorded at each step. From this point procedure was reversed, with ratings made following stepped decrease in illuminance. Imagine that you had rated the brightness at one step be 5 (slightly dark): following the increase in illuminance (Kakitsuba gives no clue that the changes were hidden from observers) the illuminance would be rated as point 4 (neutral) or above, as to keep the same rating or lower (e.g. point 6) would be nonsensical following a noticeable increase. This 'neutral' point is the one used to identify boundary illuminance: it may not be that you consider the brightness to be 'neutral', whatever that means, it just happens to be the next step up. The procedure would have been improved by randomising the order in which each illuminance step was observed.

The number of illuminance steps was not equal, being 4 steps for six conditions, but being 2.5 (an unexpected interval), 3 and 5 steps in three conditions. The size of each illuminance increment was not equal, ranging from 50 lux to 500 lux, and these did not map to regular proportions of the range, being steps of one-quarter to two-thirds of the overall range used.

There are steps that could be taken to reduce the effect of the problems described above. First, the illuminance ranges should be identical for each test condition so that observers have the opportunity to give the same response for the same illuminance in different conditions. Second, the illuminances should be observed in a random order. Third, a second procedure should be used to evaluate the same test conditions to enable the conclusions to be compared.

Further points

There are many other factors that should be considered when reviewing this experiment. The sample size (10) is too small, for example according to Cohen's rule [Field 2005] it is too small to detect a difference when expecting even a large effect size. For characterising different lamps, CCT alone is insufficient; while the paper presented graphical SPD it would have been useful to consider and report also other spectral characteristics.

Summary

This article claims to demonstrate a relationship between illuminance and CCT, updating the Kruithof curve. It does not do this: the results were derived from a confused set of response and stimulus ranges and can be explained by stimulus range bias. Such false impression of visual evaluation holds back scientific progress.

It would be interesting to repeat precisely the original Kruithof [Kruithof 1941] work to see if the same conclusions were drawn, but that is not possible since Kruithof did not describe in anything like sufficient detail what he had done. A more critical review of the literature, lacking in the Kakitsuba article, would reveal this and perhaps help to avoid further unnecessary work. The recent CIE report on spatial brightness methodology [CIE 2014] was written to promote critical consideration experimental design.

Having criticised the method, one interesting point of this work was the measurement of involuntary physiological responses in parallel to category rating. This approach deserves further attention.

References

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

Burks SL. 1994. Managing your Migraine. Humana Press, New Jersey.

Chan JC. 1991. Response-order effects in Likert-type scales. Educational and Psychological Measurement. 51: 531-540.

CIE. 2014. Guidance Towards Best Practice In Psychophysical Procedures Used When Measuring Relative Spatial Brightness. Commission Internationale De L'Éclairage. 212.

Dawes J. 2008. Do Data Characteristics Change According to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. Int. J Market Res. 50: 61–77.

Field A, 2005. Discovering Statistics Using SPSS. Sage Publications Ltd.: London. pp. 34.

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

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

Houser K, Tiller D. 2003. Measuring the subjective response to interior lighting: paired comparisons and semantic differential scaling. Lighting Research & Technology. 35(3): 183-198.

Kakitsuba N. 2015. Comfortable Indoor Lighting Conditions Evaluated from Psychological and Physiological Responses. Leukos. Online publication doi=10.1080/15502724.2015.1061945.

Kruithof AA. 1941. Tubular fluorescent lamps for general illumination. Philips Technical Review. 6(3): 65-73.

Litwin MS, 1995. How to measure survey reliability and validity. Sage Publications Ltd: London.

Loe DL, Mansfield KP, Rowlands E. 1994. Appearance of lit environment and its relevance in lighting design: Experimental study. Lighting Research & Technology. 26(3); 119-133.

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

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 Research & Application. 38(1); 46-57.

Moors G. 2008. Exploring the effect of a middle response category on response style in attitude measurement. Qual. Quant. 42: 779-794.

Parducci A, Perrett LF. 1971. Category rating scales: Effects of relative spacing and frequency of stimulus values. J Exp. Psych Monograph. 89: 427-452.

Poulton EC. 1989. Bias in quantifying judgements. Psychology Press.

Rea MS. 1982. Calibration of subjective scaling responses. Lighting Research & Technology, 14(3); 121-129.

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