Research Methods to Avoid Bias in Categorical Ratings of Brightness

SA Fotios[1] PhD, BEng, CEng, MEI, MSLL, MILE

KW Houser[2] PhD, PE, LC, LEED AP

[1]School of Architecture, The University of Sheffield, UK
[2] Department of Architectural Engineering, The Pennsylvania State University, USA

Fotios SA & Houser KW, Research Methods to Avoid Bias in Categorical Ratings of Brightness, Leukos, 2009; 5(3); 167-181

Abstract

This article presents evidence of potential contraction bias in the category rating task associated with the stimulus presentation sequence, response range and response range anchors, and a grouping bias associated with the number of stimuli and response categories. These biases tend to reduce the difference between ratings given to stimuli. It is demonstrated that such bias is sufficient to hide differences in brightness under lighting from lamps of different spectral power distribution but that precautions can be taken to successfully counter the bias. Research methods that can be employed to avoid bias in categorical ratings of brightness are summarised.

Keywords: Brightness, Spectral power distribution, category rating, bias, research methods, lighting

Research Methods to Avoid Bias in Categorical Ratings of Brightness

1. Introduction

There is ongoing attempt to investigate effects of light source spectral power distribution (SPD) within the lighting community. IESNA has established the Effects of Lamp Spectral Distribution committee to investigate SPD effects on spatial brightness and visual effort and new research was presented by several groups at the 26th Session of the CIE in Beijing, 2007. This article presents evidence to aid in the identification of experimental procedures that can be expected to produce reliable data. It is a continuation of our previous work where we described methods to counteract potential biases in the side-by-side brightness matching task [Fotios, Houser & Cheal 2008].

Specifically, this article presents a critical analysis of past research on the effect of SPD on the perception on brightness focusing on studies that used category rating as the principal experimental methodology. In category rating, subjects are presented with an environment lit by a single type of light source and use rating scales to describe the appearance of the space, e.g. a semantic differential scale of brightness, along the bright-dim axis. Different stimuli (e.g. lighting of different SPD and illuminance) are presented individually, in succession, and are usually rated in isolation of other visual stimuli.

Many different items have been rated in previous work, including brightness, clarity, colourfulness, spaciousness, cool, active, soft, calm and comfort. Some of these items are influenced by variables other than the lighting [Tiller & Rea 1992] and may therefore give a false impression of the effects of lighting. The current article focuses primarily on brightness. Hesselgren suggested that brightness and colour are the main observable attributes of lighting [Hesselgren 1967]. Boyce and Cuttle asked test participants to describe the lighting in a room in their own words and found that they used mainly terms of brightness and clarity; items such as pleasantness and colourfulness were mentioned very infrequently [Boyce & Cuttle 1990].

Twenty-one studies have used category rating to compare perceptions of lighting of different SPD at photopic levels. Seventeen studies included ratings of brightness;

three of these were field studies, where test participants were located in their normal workspace and carried out their normal tasks [Akashi & Boyce 2006, Bartholomew 1975, Cockram, Collins & Langdon 1970]; seven studies sought judgements in full size laboratory rooms, typically furnished to represent an office [Baron, Rea & Daniels 1992, Boyce & Cuttle 1990, Davis & Ginthner 1990, Flynn & Spencer 1977, Knez 1995, 2001, Vrabel, Bernecker & Mistrick 1998]; one study used both field and laboratory studies [Fleischer, Krueger & Schierz 2001]; and six studies have used lighting booths [Boyce 1977, Delaney et al 1978, Ishida, Ikeyama & Toda 2007, McNelis et al 1985, Oi & Takahashi 2007, Wake et al, 1977]. Two studies did not report the rating items used [Nakamura & Oki 2002, Rubenstein & Kirschbaum 2003] and two studies did not include brightness as a specific rating item [Boray, Gifford & Rosenblood 1989, Pracejus 1967]. A further study was carried out at mesopic levels typical of street lighting at night time [Fotios & Cheal 2007]. While this list is not assumed to be exhaustive, it does extend those presented in an earlier review of the subject [Fotios 2001].

At first inspection, there is no clear conclusion available from these results. Some studies report that lamp SPD does not have an effect on judgements of brightness [e.g. Baron, Rea & Daniels 1992, Bartholomew 1975, Cockram, Collins & Langdon 1970, Davis & Ginthner 1990, Knez 2001] while others suggest a significant effect [e.g. Flynn & Spencer 1977, Fotios & Cheal 2007, Vrabel, Bernecker & Mistrick 1998].

Spaces illuminated by lamps of different SPD can appear differently bright at the same illuminance because illuminance, as defined by The CIE Standard Photopic Observer (V_λ), is derived from a different visual process to that of spatial brightness. The post-receptoral visual system is organized in three channels, one luminance channel where signals from the long- and medium-wavelength sensitive cone types are combined, and two colour channels where the differences between signals from different combinations of cone types are taken [Hunt, 1995]. The CIE Standard Photopic Observer is based on data collected primarily using flicker photometry and step-by-step brightness matching, techniques that tend to minimize activity in the colour channels; brightness perception is dependent on activity in all three channels [Lennie, Pokorny & Smith 1993, Wagner & Boynton 1972, Yaguchi & Ikeda, 1983]. Experimental studies using alternative methods to category rating have shown that lamp SPD affects judgements of spatial brightness [e.g. Berman et al 1990, Vrabel et al 1998].

Whilst spatial brightness *can* be affected by lamp SPD, it is not assumed that all studies will produce a statistically significant effect. Some variation is expected due to the choice of the independent variables, which is typically a combination of lamp SPDs, and to other aspects of the experimental design such as size of the visual stimulus and evaluation mode.

SPD is an intricate independent variable that has an infinite number of possible levels. Different commercially available lamps are often selected to be levels of the independent variable. Derived measures are often used to coarsely characterize the levels of the independent variable, such as Colour Rendering Index (CRI), Correlated Colour Temperature (CCT), or the ratio of scotopic to photopic lumens (S/P). It is less common for researchers to create custom illuminants that have spectra intentionally designed to manipulate an underlying mechanism of vision, though it has been done (e.g. Berman et al, 1990, 1992, Houser, Tiller & Hu, 2004). The actual illuminants that are selected should be expected to influence whether or not brightness differences are found.

A reported significant effect of SPD on brightness judgements may also be due to a Type I statistical error arising from experimental errors and random chance. A Type I error is also known as a false positive or an error of credulity. It means finding a statistically significant effect (e.g. rejecting the null hypothesis that both stimuli are equal in brightness), when there is none (e.g. the null hypothesis of no difference in brightness is actually true).

In consideration of the above, the current review has been carried out through consideration of methodologies, identifying where methodological features explain consistent outcomes. A first step is to assess the quality of the research. Eleven studies are considered to be of dubious value because the published work does not present sufficient information to describe the methodology or the findings [Bartholomew 1975, Cockram, Collins & Langdon 1970, Delaney et al 1978, Fleischer, Krueger & Schierz 2001, Ishida, Ikeyama & Toda 2007, McNelis et al 1985, Nakamura & Oki 2002, Oi & Takahashi 2007, Pracejus 1967, Rubenstein & Kirschbaum 2003, Wake et al 1977]. Table 1 shows some of the data missing but is not an exhaustive list. One common error is that only the mean value of a variable is reported and statistical analysis is absent or inadequately described. The absence of statistical analysis means there is no way of knowing if differences between lamps

are real or chance, and the absence of variance data (or similar) prevents post-hoc statistical analysis.

The decision to consider a study reliable is based on the quality of the research process and documentation; it is not based on the specific results. Characteristics of papers that we have been deemed reliable include: thorough explanation of the independent and dependent variables, complete reporting of the experimental methodologies and the collected data, and proper use of statistical analyses in forming the conclusions (or sufficient data to permit application of statistical analysis). If a paper is lacking in one of these categories then we have placed it in the 'unreliable' category. The decision to consider a study as 'unreliable' is frequently a subjective process and it is most often the result of incomplete reporting. We welcome feedback from others where these decisions are considered to be unfair or inconsistent.

[Table 1]

There are three characteristics of the experimental design that strongly contribute to why some studies find an effect of lamp SPD on brightness and others do not:

1. The levels of the independent variable, which relates to the choice of illuminants.
2. The degree of chromatic adaptation.
3. The experimental design introduces bias which hides any effect of SPD.

These three characteristics of are mutually exclusive but their effects may be additive. It is therefore prudent to address each of them when conceptualizing and designing experiments that use category rating to evaluate the perception of brightness. The first explanation assumes that lamp SPD does affect brightness but that the range of lamps used in some studies is not sufficient to enable visual discrimination. The second explanation also assumes that lamp SPD affects brightness and that this effect is due to the chromatic contribution, which is activity in the two opponent colour channels that process the retinal signal [Guth & Lodge 1973, Yaguchi & Ikeda 1983]. With time, chromatic activity is modified by adaptation. The time course of chromatic adaptation has been measured using colour appearance judgements following a change in adaptation. The data suggest two stages of adaptation; a rapid stage, giving approximately 60% chromatic adaptation in the first five seconds, and a slower stage where approximately 90% chromatic adaptation is

reached after 60 seconds. It takes almost two minutes to approach 100% chromatic adaptation [CIE 2004, Fairchild & Reniff 1995, Shevell 2001]. If a rating of brightness is given after two or more minutes exposure to a single stimulus, activity in the opponent colour pathways will be weaker than the original sensation and the difference in brightness between two stimuli of different SPD will be less than that of immediate observations. As discussed below, there is some evidence that lamp SPD can affect brightness despite long term adaptation [Akashi & Boyce 2006].

The third explanation also assumes that lamp SPD can affect brightness, but that behavioural bias due to the experimental design is of sufficient magnitude to hide differences between the stimuli of different SPD. The category rating task is known to be prone to bias, and is hence listed below the matching and discrimination tasks in Poulton's order of preference of methods for making quantifying judgements [Poulton 1989]. Nonetheless, there are experimental situations where category rating can be expedient. Rather than dismiss the method altogether, we instead wish to provide guidance about potential pitfalls and how to avoid them, which is the focus of this article.

In the context of a psychophysical measurement such as the effect of lamp SPD on brightness, bias means an unfair assessment of the effect. It is a systematic distortion of a response that most commonly results from the experimental methodology. An experimental bias pollutes the data by being confounded with the lighting effects under study. There are many causes of bias including experimenter induced effects such as dissimilar visual fields or inaccurate physical measurement, unintentional manipulation of subjects' behaviour, shifts in the subject's psychophysical criterion, and sensory variables such as light/dark, chromatic or contrast adaptation. Category rating judgements are determined by the relationship of the stimulus to the range of contextual values and also by habits or biases governing the frequency with which different categories or parts of the rating scale are used [Parducci & Perrett 1971]. If bias changes the rating given to a particular stimulus then this leads to a misunderstanding of the effect of lamp SPD.

The easiest way to bias peoples' judgements is to ask them to judge magnitudes that do not have familiar physical units, and this is the case for brightness judgements [Poulton 1989]. Whilst naïve observers can discriminate *between* brightness magnitudes they do not have familiar physical units with which to make judgements of absolute magnitude. In a side-by-side comparison, a joint evaluation of two stimuli,

the difference in brightness and its direction are easily evaluable: in single stimulus presentations, whereby two stimuli are judged independently, the evaluation of brightness is a more difficult task [Hsee et al 1999] leading to greater variance between subjects and between trials.

This article discusses response contraction biases associated with the stimulus presentation sequence, response range and anchoring of the response range, and also a grouping bias associated with the relationship between the number of stimuli and the number of response categories. These biases are used to explain the conclusions drawn from previously published experimental data. The main objective of this article is to provide guidance about dealing with potential bias in the category rating task, while drawing attention to why some studies find a significant effect of SPD on brightness and others do not.

2 Causes of Contraction Bias

Four potential causes of contraction bias in the category rating task are identified from the experimental psychology literature, a key text being Poulton [Poulton 1989]. These biases tend to reduce the apparent difference between stimuli.

2.1 Order Effect

In category judgement experiments (and similarly with magnitude estimation) subjects' judgements are affected by both the present and previous stimuli. The response to a particular stimulus tends to be biased toward the response on the previous trial, resulting in an underestimate of the size of the difference between stimuli [Gescheider 1988, Poulton 1989, Staddon et al 1980]. Subjects are constantly making small adjustments to their internal reference during an experiment; this may be an adaptation effect or they may improve on later conditions through a learning effect or becoming less anxious. Thus the order of stimulus presentation will affect behaviour.

Empirical evidence for sequential effects can be found in the loudness judgements of Ward and Lockhead; the higher the value of the stimulus on trial N-1, the higher the average response to the stimulus on trial N [Ward & Lockhead 1970]. In their *direct comparison* tests, Flynn & Spencer sought judgements of CW lighting both before and after judgements of HPS lighting using seven semantic differential rating scales [Flynn & Spencer 1977]. In five of the scales (pleasantness, distinctness, clarity, brightness, and colourfulness), but not the likeness and beautiful scales, the CW

post-HPS received a higher rating than the CW pre-HPS, although the differences are not significant.

Practice trials are commonly employed to reduce the response variability associated with learning an unfamiliar task. While the learning-curve for a new task is steepest in the beginning, subjects will continue to learn and will continue to adjust their internal reference with repetitions. Practice trials are a feature of good experimental protocol, but they cannot be relied upon to eliminate order effects.

Poulton [Poulton 1989] suggested three procedures for avoiding sequence effects due to stimulus presentation order. One procedure is to record judgements of a stimulus immediately following judgements of an identical stimulus. The second is to ask subjects for only a single judgement, hence using independent samples to judge different stimuli. The third procedure is to present a fixed standard before each test stimulus. These procedures are not always practical, convenient or efficient within an experiment. An alternative procedure, frequently used in lighting research, is to use a well mixed order of stimuli by randomising or counterbalancing the order of presentation [Cliff 1973]. Whilst a mixed stimulus order does not eliminate the sequence effect it removes the effect of bias from the experimental results by distributing the error randomly or in a counterbalanced way.

2.2 Response Range Bias

The number of categories in a response scale must reflect the ability of subjects to use categories and also the accuracy by which the recorded data represents the subjects intended response.

Heise used a seven-point range to demonstrate typical semantic differential scaling, with the ends labelled Good (3) and Bad (-3) and the central point being labelled neutral (0) [Heise 1969]. Previous lighting studies have tended to use seven-point rating scales, for example a scale ranging from 1 = dim to 7 = bright, and this is commonly used in definitions of the semantic rating task;

> *The semantic differential consists of a set of bipolar,* ***seven****-category rating scales* [Tiller & Rea 1992].

> *Semantic differential rating scales – a* ***seven*** *category range between the extremes* [Houser & Tiller 2003].

There is some evidence that subjects are able to reliably distinguish between approximately seven categories of a uni-dimensional stimulus, and this is apparent for a broad range of sensory judgements, but with more than seven categories confusions become more frequent [Miller 1956].

The accuracy by which a response scale represents a subject's intended response is the need to force responses into a limited set of categories whose mid-points are assumed to be uniformly spaced, integer values. This is tantamount to subjecting the original response to a non-linear transformation. For example, if the subject's real assessment of brightness were really 1.2, 1.9 and 3.3, the subject would be forced to treat these values as 1, 2 and 3. While the order of the scale is maintained, the ratios of their scale separations are not [Green & Rao 1970]. Green and Rao created a synthetic data set to examine data recovery using 2-, 3-, 6- and 18-point response categories. Their data suggests diminishing returns of correlation beyond 6 response categories, but that a fewer number (2 or 3) leads to poor recovery [Green & Rao 1970].

There is a tendency for respondents to avoid using the ends of a scale, to underestimate large sizes and overestimate small sizes, and ratings will thus converge toward the centre of the response range. This response contraction is enhanced if the response range has an obvious middle value and can reduce the distinction between stimuli [Poulton 1989]. Such an outcome can be observed in the findings of previous lighting research:

- Wake et al [Wake et al 1977] used 7-point scales, and for their brightness rating they concluded "*the differences among lamps are extremely small*".
- Akashi and Boyce [Akashi & Boyce 2006] used 5-point scales (-2 to +2) with a middle neutral point marked "0" and found "*The mean ratings … do not indicate any strong opinions, i.e. all mean responses are around neutral*"

Because of potential response contraction bias it is not clear whether there really is no difference of brightness between the lamps used in these studies, under the particular conditions used, or if the test failed to reveal a difference. Further lighting studies have also highlighted the middle value of a category scale as being neutral [DeLaney et al 1978, Ishida et al 2007, McNelis et al 1985, Oi & Takahashi 2007].

In an example of good experimental practise Akashi and Boyce used two data gathering tools, the category rating task and questionnaires seeking a yes/no

response to statements regarding the visual environment – this latter might be considered a two-category response scale [Akashi & Boyce, 2006]. The occupants of four offices judged their lighting over a period of several months. Initially (*stage 1*), each office had similar lighting, a correlated colour temperature (CCT) of 3500K and mean desk illuminances of 544 to 586 lux. After nine months the first intervention (*stage 2*) was to reduce the illuminance in two offices by approximately one third (by removal of one of the three lamps in each luminaire) and to increase the correlated colour temperature to 6500K in two offices by replacement of the lamps: these changes were balanced so that one office was unchanged, one office had a reduction in illuminance and an increase in CCT, and in the remaining two offices just one change was made. The category rating task employed a five point response scale and these results did not reveal any significant effect. The two-category task did however reveal some significant effects. In the office where illuminance was reduced, this lead to a significant increase in judgements that the office appeared gloomy; in the office where the reduction in illuminance was accompanied by an increase in CCT, there was a significant reduction in judgements that the lighting was too dim.

To counter a potential contraction bias, rating scales should avoid an obvious centre to the rating scale, i.e. use an even number of response categories. [Poulton 1989] Following from Miller [Miller 1956] and Green and Rao [Green & Rao 1970] this might be a response range of 6 or 8 categories.

2.3 Anchoring the Response Range

Consider a subject instructed to rate a visual environment using a defined response scale. In the absence of visual demonstration of the meaning of the upper and lower ends of the response scale each subject must develop their own internal criteria. These decisions are difficult to make, particularly early in an experiment when the subject has not yet seen the range of possibilities. It has therefore been recommended that pre-experimental standards are used to define to observers the meaning of the upper and lower limits of a rating scale, anchoring the response range to the stimulus range [Tiller & Rea 1992, Houser & Tiller 2003]. For example, in acoustics research experimenters present subjects with sample tones to define a response scale of loudness differences [Schneider et al 1978]. Response range anchoring has the potential to reduce the variance in the data, or at least to increase the internal consistency of each subject, and to reduce response contraction bias [Poulton 1989].

One study has done this [Fotios & Cheal 2007] although of their eight rating items they demonstrated only the brightness dimension. This was done by showing the test stimuli predicted by pilot studies to be the most and least bright. Whilst it is relatively straight forward to visually illustrate brightness, and perhaps colourfulness, it is more difficult (if at all practical) to define other items used in evaluation of lighting such as pleasantness, acceptability, spaciousness and likeness. Vrabel, Bernecker and Mistrick circumvented this by giving written definitions rather than visual demonstration [Vrabel, Bernecker & Mistrick 1998].

Anchors can help to produce an even distribution of ratings and to produce a linear function. Without anchors, functions are likely to be S-shaped – floor and ceiling effects. Consider the smallest stimulus in the range; this can be confused only with larger stimuli, and thus all of its confusions increase its average rating, making it close to the rating of the next stimulus; the next stimulus is less affected because it can be confused with stimuli that are both smaller and larger; this is a floor effect, the function becoming shallow at the lower end of the stimulus range, and a similar ceiling effect is expected at the higher end [Poulton 1989]. Anchors at both ends of a stimulus range help to counteract floor and ceiling effects. However, anchors may over-correct for the floor/ceiling effects, leading to a function that is too steep at both ends, and this can be compensated for by having anchors located just beyond both ends of the stimulus range [Poulton 1989].

2.4 Grouping Bias

The rating task is affected by the number of response categories and the number of stimulus magnitudes [Poulton 1989]. If there are fewer categories than stimulus magnitudes then observers attempt to discriminate one stimulus from another, putting together in the same category stimuli that are most easily confused with each other. If there are only a few categories and an equal number of stimulus magnitudes, the task changes from one of grouping to one of placing each stimulus in its own specific category. If there are 20 or more categories and relatively few different stimulus magnitudes, some categories hence being unused, the task begins to resemble a numerical magnitude judgement.

Thus, if the response scale has fewer categories than there are stimuli, several stimuli will need to be grouped within each category, and this may hide the difference between two stimuli when this difference is small but nonetheless real. Consider the

study by Boyce and Cuttle (their *Experiment 1*) which used 22 stimulus conditions, including four types of lamp, and a five-point response range [Boyce & Cuttle 1990]. Their participants would thus need to group several stimuli within each response category. Only one of the 19 rating items (dim) was found to be significantly affected by lamp type, and this at $p<0.05$ may be a Type I error (i.e. erroneous rejection of the null hypothesis).

3 The Effect of Bias on Test Results

The influence of response contraction bias in the category rating task is demonstrated through analysis of repeated measures data. In tests employing repeated measures, subjects see a range of stimulus magnitudes; for the issue under discussion these usually comprise controlled variations in illuminance and SPD (lamp type). Variations in illuminance tend to be large: e.g. test illuminances of 300 and 600 lux [Boyce 1977], 30, 90, 225 and 600 lux [Boyce & Cuttle 1990], and 269, 592 and 1345 lux [Davis & Ginthner 1990]. Such differences are easily noticeable – in these three studies the effect of illuminance on ratings of brightness was significant and this shows that the category rating task is strong enough to reveal differences.

Table 2 allocates repeated measures studies into one of four categories according to the variables examined and the reported effect of lamp type on brightness. Consider studies which presented variations in both lamp type and illuminance [Boyce 1977, Boyce & Cuttle 1990, Davis & Ginthner 1990]; whilst these studies found significant effects of illuminance they did not find a significant effect of lamp type on brightness. Next consider three studies in which only lamp type was varied [Boyce & Cuttle 1990, Flynn & Spencer 1977, Vrabel, Bernecker & Mistrick 1998]; these studies did find a significant effect of lamp type on brightness. The difference between these two groups could be explained as a range equalizing bias [Poulton 1989]. When only lamp type is varied, sensitivity to these differences in brightness increases, and they expand to fill the response range. Brightness differences due to SPD are more difficult to evaluate than differences due to illuminance, so when both illuminance and lamp type are varied, illuminance, the easy-to-evaluate attribute, is the primary determinant of the evaluation [Hsee et al 1999]; illuminance differences will expand to fit the response range and differences between lamp type will register as smaller intervals on the response range.

[Table 2]

The studies in both of these groups are suspected to contain response contraction bias as described above (other than order effect, as stimuli were presented in random order). The difference between the groups is whether illuminance was included as a test variable – if it was, the effect of lamp type was found to be negligible. There is however one study in which both lamp type and illuminance were varied and which did find a significant effect of lamp type on brightness [Fotios & Cheal 2007] and what this study did differently was to take precautionary steps to counter possible response contraction bias. These steps were:

- A response scale with an even number of categories (8) and hence no obvious middle value.
- The brightness response range was anchored to the stimulus range by visual demonstration at the start of each test session.
- Stimuli were presented in randomised order.
- The number of stimulus magnitudes (10) and number of response categories (8) were similar.

This analysis suggests that response contraction bias is sufficient to hide the difference in brightness between lamps of different SPD when these differences are relatively small compared to simultaneous variation in illuminance, but that the difference in brightness between lamps can be revealed when the experimental procedure includes precaution against response contraction bias.

The results are not simply explained by chromatic adaptation. Table 3 identifies the time interval between exposure to the stimulus and the point at which the judgement was recorded: in some tests this data is reported, and in others an estimate has been made from description of the experimental procedure. Chromatic adaptation will be incomplete in exposures up to one minute and complete for exposures of two minutes or more [CIE 2004, Fairchild & Reniff 1995, Shevell 2001]. Chromatic adaptation diminishes the activity in the chromatic (parvocellular) pathways between the retina and brain. This might suggest that studies with incomplete chromatic adaptation would demonstrate a strong effect of lamp type on brightness, and studies permitting complete chromatic adaptation would reveal no effect, or a weak effect, of lamp type on brightness. Table 2 shows that this is not always the case. Davis and Ginthner used incomplete chromatic adaptation (one minute) and found no effect of SPD on brightness [Davis & Ginthner 1990]; the procedure used by Boyce & Cuttle would allow complete chromatic adaptation (approximately 15 minutes) yet in their

second experiment they found a significant effect of lamp type on brightness [Boyce & Cuttle 1990]. Even complete chromatic adaptation does not fully eliminate the effects of lamp spectral power on the perception of brightness.

[Table 3]

Chromatic adaptation needs further consideration for its impact on the time course of brightness judgements in practical situations. The important question is *what is important to the occupant?* If it is the first impression made when entering a space, or when abruptly changing the illumination, then the results of test using mixed or incomplete chromatic adaptation are relevant. Consider a person who walks into a room and thinks "*this lighting is bright*"; if it is assumed that a few minutes later they won't suddenly change their mind and think "*I was wrong, it's dark in here*" then first impressions count. The alternative is that people establish their opinion of brightness after long term exposure, perhaps several minutes or more, in which case the results of tests enabling complete chromatic adaptation are relevant. Further research is needed to determine whether it is the first impression or long term impression that has the greater impact on judgments of lighting.

4 Studies using independent samples

Three category-rating studies have used independent samples, where subjects participated in only one test condition [Baron, Rea & Daniels 1992, Knez 1995, 2001]. Independent samples are advantageous because they avoid order effects and the response scale grouping bias, and the aim of the experimental research is less likely to become obvious to the participant. Provided that the assignment of subjects to experimental groups is random then it is possible to infer that any difference between the groups is attributable to the experimental treatment. However, independent samples are less powerful than repeated measures and for a given number of subjects and a given effect size, a repeated measures trial is more likely to detect a significant change than is an independent samples trial.

Knez examined only the effect of lamp type and found no significant differences between lamps [Knez 2001]. Two studies [Baron, Rea & Daniels 1992, Knez 1995] examined both lamp type and illuminance, and in both studies it was found that illuminance had a significant effect on ratings of brightness. Baron et al found that lamp type did not affect ratings of brightness, although it did affect some other rating scales [Baron, Rea & Daniels 1992]. Knez found that there was a significant

difference in ratings of brightness for two lamps of low CRI ($p<0.01$) but not for two lamps of high CRI [Knez 1995]. The four lamps were used in separate trials (low CRI and high CRI) for which the overall results are not compared by Knez and there are insufficient data to do so; it is possible that this one significant effect of lamp type on brightness is a chance effect.

There are two reasons why any effect of lamp type on brightness would be diminished in these tests. Firstly, they used long periods of adaptation, these being 20 minutes [Baron, Rea & Daniels 1992], 105 minutes [Knez 2001] and 115 minutes [Knez 1995]. These are longer periods than used in the repeated measures tests (Table 3) and thus any activity in the chromatic visual pathways contributing to brightness would be smaller. Secondly, there are two potential causes of contraction bias; all three studies employed a five-category response scale, hence having an obvious middle value, and did not anchor the response range with visual stimuli. In the *separate* mode of evaluation [Hsee et al 1999] it is difficult to judge the unfamiliar magnitude of brightness, and this is more-so for the relatively small effect of lamp type on brightness rather than the relatively large effect of illuminance which did cause significant effect in the two studies in which it was varied. Therefore it is unlikely that the independent samples category rating task, which employs long adaptation periods and fails to counter potential contraction bias, would reveal differences in brightness from lamps of different SPD. This suggestion could be tested by carrying out an independent samples rating task using a shorter adaptation period (up to 15 minutes according to significant effects noted in Table 3) and taking steps to counter bias.

5 Conclusion

Category rating is prone to bias, suffering from six of the seven biases listed by Poulton where the matching task avoids them all [Poulton 1989, Table 3.4, p61]. This paper has discussed causes of response contraction bias that can hide the difference in brightness between lamps of different SPD: order effect, response range bias, failure to anchor the response range, and response range grouping bias.

It has been demonstrated that this bias is sufficient to hide differences in brightness between lamps when these variations are accompanied by variations in illuminance, but that if precautions against contraction bias are taken then the effect of lamp type can be revealed. Contraction biases are strong enough to hide relatively weak

effects, such as that of two SPDs that are not substantively different, but will not hide strong effects such as from large differences in illuminance.

In contrast with bias in the brightness matching task which tends to exaggerate differences between lamps [Fotios, Houser & Cheal 2008] bias in the category rating task tends to reduce differences between lamps. Robust conclusions demand the same stimuli are compared using a variety of psychophysical methods. A few studies have employed two or more methods to compare the same stimuli [Akashi & Boyce 2006, Boyce 1977, Fotios & Cheal 2007, Hu, Houser & Tiller 2006, Vrabel, Bernecker & Mistrick 1998] but most do not. Analysis of SPD effects across a range of studies relies on the assumption of converging evidence, but there are many differences between studies (e.g. evaluation mode, visual objectives, adaptation time and visual environment in addition to differences in the SPD of lamps used) and this confounds analysis of lamp SPD effects.

Our recommendations for assuaging the bias associated with the category rating task when judging brightness can be summarized as follows:

- To address bias associated with presentation order use one or more of these methods: 1) record judgements of a stimulus immediately following judgements of an identical stimulus, 2) ask subjects for only a single judgement, hence using independent samples to judge different stimuli, 3) present a fixed standard before each test stimulus, or 4) use a well mixed order of stimuli by randomising or counterbalancing the order of presentation.
- To counter a potential response range bias, use an even number of response categories. A response range of 6 or 8 categories is typically appropriate for brightness judgments.
- The response range should be anchored to the stimulus range using a pre-experimental visual demonstration. The anchors should be located just beyond both ends of the stimulus range.
- To avoid a grouping bias the number of stimuli and the number of response categories should be similar.

This list is a reinforcement of the need for good experimental design. Other features of careful experimental work include keeping the background and surrounding conditions constant so that judgements of different stimuli are not differently affected.

Together with clear directions, these procedures should also have the effect of reducing variance [Cliff 1973] and providing more defensible data.

In addition to good experimental design, a study of the SPD effect on spatial brightness is more useful if it considers the physiology of the eye and practical application of the findings. A sensitive experiment may detect a difference in brightness judgements with lighting of different SPD, but this difference may not be large enough to be of practical importance. Practical application will place the findings within the context of the built environment.

Fotios & Houser have announced their intention to collate data with which to test models of spatial brightness at photopic levels [Fotios & Houser 2007]. The current work suggests guidelines for identifying reliable evidence of lamp SPD effects from those studies using the category rating task: retain those studies where precaution was taken against bias and those repeated measures studies which did not take precautions against bias but in which lamp type is the sole variable; discard those studies which did not take precautions against bias and varied both illuminance and SPD. This would mean retaining data from Boyce and Cuttle (their e*xperiment 2*), Flynn and Spencer, and Vrabel, Bernecker & Mistrick [Boyce & Cuttle 1990, Flynn & Spencer 1977, Vrabel, Bernecker & Mistrick 1998] but ignoring data from the remaining studies. Given the limited number of past research that stands up to rigorous evaluation, there is a need for new studies that are intentionally and carefully designed to avoid the many potential experimental pitfalls that lead to questionable data.

References

Akashi Y, Boyce PR. 2006. A field study of illuminance reduction. Energy & Buildings 38(6): 588-599

Baron RA, Rea MS, Daniels SG. 1992. Effects of indoor lighting (illuminance and spectral distribution) on the performance of cognitive tasks and interpersonal behaviours: the potential mediating role of positive affect. Motivation & Emotion 16(1): 1-33

Bartholomew R. 1975. Lighting in the classroom. Building Research & Practice 3(1): 32-39

Berman SM, Jewett DL, Fein G, Saika G & Ashford F, Photopic luminance does not always predict perceived room brightness, Lighting Research & Technology 22(1) 37-41 (1990)

Berman SM, Fein G, Jewett DL, Saika G, & Ashford F, 1992. Spectral determinants of steady-state pupil size with full field of view, Journal of the Illuminating Engineering Society 21(2): 3-13.

Boray PF, Gifford R, Rosenblood L. 1989. Effects of warm white, cool white and full-spectrum fluorescent lighting on simple cognitive performance, mood and ratings of others. Journal of Environmental Psychology 9: 297-308

Boyce PR. 1977. Investigations of the subjective balance between illuminance and lamp colour properties. Lighting Research & Technology 9: 11-24

Boyce PR, Cuttle C. 1990. Effect of correlated colour temperature on the perception of interiors and colour discrimination. Lighting Research & Technology 22(1): 19-36

CIE. 2004. Chromatic Adaptation under Mixed Illumination Condition when Comparing Softcopy and Hardcopy Images. Commission Internationale De L'Éclairage, 162:2004

Cliff N. 1973. Scaling. Annual Review Psychology 24: 473-506

Cockram AH, Collins JB, Langdon FJ. 1970. A study of user preferences for fluorescent lamp colours for daytime and night-time lighting. Lighting Research & Technology 2(4): 249-256

Davis RG, Ginthner DN. 1990. Correlated color temperature, illuminance level and the Kruithof curve. Journal of the Illuminating Engineering Society, Winter, 27-38

DeLaney WB, Hughes PC, McNelis JF, Sarver JF, Soules TF. 1978. An examination of visual clarity with high colour rendering fluorescent light sources. Journal of the Illuminating Engineering Society 7(2): 74-84

Fairchild MD, Reniff L. 1995. Time course of chromatic adaptation for color-appearance judgements. Journal of the Optical Society of America, series A, 12(5): 824-833

Fleischer S, Krueger H, Schierz C. 2001. Effect of brightness distribution and light colours on office staff. Lux Europa, Rejkjavik, 18-20 June 2001, 76-80

Flynn JE, Spencer TJ. 1977. The effects of light source colour on user impression and satisfaction. Journal of the Illuminating Engineering Society 6: 167-179

Fotios SA. 2001. Lamp colour properties and apparent brightness: a review. Lighting Research and Technology 33(3): 163-181

Fotios SA, Cheal C. 2007. Lighting in subsidiary streets: investigation of SPD effects. Part 2 – Brightness. Lighting Research & Technology 39(3): 233-252

Fotios S, Houser K. 2007. Research of Lamp SPD Effects on the Perception of Interior Spaces: The Current State of Knowledge. 26th Session of the CIE, Beijing, 4-11 July 2007, vol. 1, pp. D1-111 to D1-114

Fotios SA, Houser KW, Cheal C. 2008. Counterbalancing needed to avoid bias in side-by-side brightness matching tasks. Leukos 4(4): 207-223

Gescheider GA. 1988. Psychophysical scaling. Annual Review Psychology 39: 169-200

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

Guth SL, Lodge HR. 1973. Heterochromatic additivity, foveal spectral sensitivity, and a new colour model. Journal of the Optical Society of America 63(4): 450-462

Heise DR. 1969. Some methodological differences in semantic differential research. Psychological Bulletin 72(6): 406-422

Hesselgren S. 1967. The language of architecture. Studentlitteratur, Lund, Sweden.

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

Houser KW, Tiller DK, Hu X. 2004. Tuning the fluorescent spectrum for the trichromatic visual response: A pilot study Leukos 1(1): 7-22.

Hu X, Houser KW, Tiller DK. 2006. Higher colour temperature lamps may not appear brighter. Leukos 3(1): 69-81. (*For more complete details of this work see Houser KW, Tiller DK, Hu X. 2003. Prototype demonstration of vision-tuned fluorescent lamps, Energy Innovations Small Grant (EISG) Program, Final Report for the California Energy Commission*)

Hunt RWG. 1995. Measuring Colour. Second edition. Ellis Horwood, London

Hsee CK, Loewenstein GF, Blount S, Bazerman MH. 1999. Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. Psychological Bulletin 125(5): 576-590

Ishida T, Ikeyama K, Toda N. 2007. Psychological evaluation of lighting with a wide range of colour temperatures and illuminances. 26th Session of the CIE, Beijing, 4-11 July 2007, pp.D1-178 to D1-181

Knez I. 1995. Effects of indoor lighting on mood and cognition. Journal of Environmental Psychology 15: 39-51

Knez I. 2001. Effects of colour of light on non-visual psychological processes. Journal of Environmental Psychology 21: 201-208

Lennie P, Pokorny J, Smith VC. 1993. Luminance. Journal of the Optical Society of America (A) 10(6): 1283-1293

McNelis JF, Howley JG, Dore GE, DeLaney WB. 1985. Subjective appraisal of colored scenes under various fluorescent lamp colors. Lighting Design & Application 15(6): 25-29

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

Nakamura H, Oki M. 2002. Effect of color temperature and illuminance on preference of atmosphere, and Kruithof curve. CIE/ARUP Symposium on Visual Environment, 24th & 25th April 2002, the Royal Society, London, CIE publication x024:2002, pp95-100

Oi N, Takahashi H. 2007. Preferred combinations between illuminance and color temperature in several settings for daily living activities. 26th Session of the CIE, Beijing, 4-11 July 2007, (Abstract pp D3-178; full paper not included, downloaded from authors website).

Parducci A, Perrett LF. 1971. Category rating scales: Effects of relative spacing and frequency of stimulus values. Journal of Experimental Psychology Monograph 89(2): 427-452

Poulton EC. 1989. Bias in quantifying judgements, Lawrence Erlbaum Associates Ltd; Hove & London

Pracejus WG. 1967. Preliminary report on a new approach to color acceptance studies. Illuminating Engineering 62(12): 663-673

Rubenstein G, Kirschbaum CF. 2003. Colour temperature and illuminance levels in offices. 25th Session of the CIE, vol 2, D3-110 to D3-113

Schneider B, Parker S, Valenti M, Farrel G, Kanow G. 1978. Response bias in category and magnitude estimation of difference and similarity for loudness and pitch. Journal of Experimental Psychology 4(3): 483-496

Shevell SK. 2001. The time course of chromatic adaptation. Color Research & Application, Supplement 26: S170-S173

Staddon JER, King M, Lockhead GR. 1980. On sequential effects in absolute judgement experiments. Journal of Experimental Psychology 6(2): 290-301

Tiller DK, Rea MS. 1992. Semantic differential scaling: Prospects in lighting research. Lighting Research & Technology 24(1): 43-52

Vrabel PL, Bernecker CA, Mistrick RG. 1998. Visual performance and visual clarity under electric light sources: Part II - Visual Clarity. Journal of the Illuminating Engineering Society 27(1): 29-41

Wagner G, Boynton RM. 1972. Comparison of four methods of heterochromatic photometry. Journal of the Optical Society of America 62(12): 1508-1515

Wake T, Kikuchi T, Takeichi K, Kasama M, Kamisasa H. 1977. The effects of illuminance, color temperature and colour rendering index of light sources upon comfortable visual environments in the case of the office. Journal of Light & Visual Environment 1(2): 31-39

Ward LM, Lockhead GR. 1970. Sequential effects and memory in category judgements. Journal of Experimental Psychology 84(1): 27-34

Yaguchi H, Ikeda M. 1983. Contribution of opponent-colour channels to brightness. in Mollon JD, Sharpe LT (Eds), Colour Vision: Physiology & Psychophysics. Academic Press Inc. (London) Ltd; London. 353-360

Missing information or suspected error	Study affected
No statistical analysis presented and post-hoc analysis prevented by missing information, e.g. absence of variance data and sample size	Bartholomew, 1975; Cockram, Collins & Langdon, 1970; Delaney et al, 1978; Fleischer et al, 2001; Ishida et al, 2007; Nakamura & Oki, 2002; Oi & Takahashi, 2007; Pracejus, 1967; Rubenstein & Kirschbaum, 2003; Wake et al, 1977
Results presented graphically only and incomplete annotation hinders interpretation.	Ishida et al, 2007; Oi & Takahashi, 2007; Rubenstein & Kirschbaum, 2003
Multiple attributes of room appearance were rated by subjects, but these results are grouped into one, or a few, categories with no statistical validation of the grouping: effect on individual items is not available.	Fleischer et al, 2001; McNelis et al, 1985
Subjects are instructed to apply ratings to assumed contexts rather than the actual visual environment.	Cockram, Collins & Langdon, 1970; Nakamura & Oki, 2002
The test apparatus and procedure are incompletely described.	Delaney et al, 1978; McNelis et al, 1985; Pracejus, 1967
Lamp type variable confounded by variations in illuminance.	Bartholomew, 1975; Cockram, Collins & Langdon, 1970

Table 1

Summary of data missing from previous studies using category rating to compare perception of the visual environment under different types of lamp. These omissions mean the studies give an unreliable account of the effect of lamp type on perceptual attributes.

	Lamp type and illuminance are varied	**Lamp type is the sole variable**
Effect of lamp type on *brightness* is significant	Fotios & Cheal, 2007	Boyce & Cuttle, 1990 (*Experiment 2*) Flynn & Spencer, 1977 Vrabel, Bernecker & Mistrick, 1998
Effect of lamp type on *brightness* is negligible	Boyce, 1977 Boyce & Cuttle, 1990 (*Experiment 1*) Davies & Ginthner, 1990	(none)

Table 2
Comparison of category rating studies using repeated measures. In tests which modulated only lamp type, this was found to have a significant effect on ratings of brightness; in tests which modulated both lamp type and illuminance, then illuminance had significant effect on brightness but lamp type did not. In only one study [Fotios & Cheal 2007] examining both lamp type and illuminance was lamp type found to have significant effect on brightness and this study took precautions against experimental bias.

Adaptation group	Study	Test variables		Adaptation time (minutes)	Significant effect of SPD on brightness?
		Illuminance	SPD		
Incomplete chromatic adaptation	Davis & Ginthner, 1990	✓	✓	1	No
	Flynn & Spencer, 1977	X	✓	1	Yes
	Vrabel et al, 1998	X	✓	1*	Yes
Complete chromatic adaptation	Boyce, 1977	✓	✓	5*	No
	Fotios & Cheal, 2007 (*mesopic*)	✓	✓	5	Yes
	Boyce & Cuttle, 1990; experiment 1	✓	✓	15*	No
	Boyce & Cuttle, 1990; experiment 2	X	✓	15*	Yes

Table 3 Comparison of SPD effect on brightness and adaptation time in studies using the repeated measures category rating task.

*The adaptation times in these studies were not reported and hence estimated from description of the test procedure.

- Boyce, 1977: the category rating task followed a previous set of 34 rating scales.
- Boyce & Cuttle, 1990; the category rating task followed two preliminary colour discrimination tasks.
- Vrabel et al, 1998; they allowed a few minutes at the start of the test, but ratings may have been applied on immediate exposure to different lamps. There were eight rating items, hence on average perhaps 1 minute before considering brightness.