Manuscript title:
A comparison of partition scaling and magnitude estimation for brightness scaling.

Authors and affiliations:

Laurens Van de Perre [1]	laurens.vandeperre@kuleuven.be
Kevin A. G. Smet [1]	kevin.smet@kuleuven.be
Marc Dujardin [2]	marc.dujardin@kuleuven.be
Peter Hanselaer [1]	peter.hanselaer@kuleuven.be
Wouter Ryckaert [1]	wouter.ryckaert@kuleuven.be

[1] ESAT/Light&Lighting Laboratory, KU Leuven, Ghent, Belgium;
[2] Faculty of Architecture, KU Leuven, Ghent, Belgium;

A comparison of partition scaling and magnitude estimation for brightness scaling.

ABSTRACT

Several psychophysical procedures exist to capture the relationship between the magnitude of an optical stimulus (e.g. luminance) and the corresponding perceived brightness. A commonly used psychophysical method is magnitude estimation (ME). However, some drawbacks of this method have been identified in literature: the method could be prone to some biases such as order effects, range effects and centring bias. On the other hand, partition scaling (PS) is one of the oldest, yet rarely used, psychophysical method to derive a similar luminance-brightness relationship. In the present study, the two methods are compared in terms of robustness and susceptibility to possible biases. Psychophysical experiments with simple achromatic discs were set up to obtain a brightness scale as a function of luminance using ME. Regarding PS, very similar experiments have been conducted in a previous study. The experiments were conducted for four luminance ranges: three subranges (low-, mid- and high-range of luminance) equally divided from 5 to 175 cd/m^2 and one overlapping full-range. Results indicate that observers had difficulties in accurately estimating the mid- and high-ranges for both methods, because the brightness between discs was too similar. Perceptual brightness scales could be obtained for the full- and low-range, although strong evidence was found for a range bias using ME. When pooling the full- and low-range data, both ME and PS results converge. Results show that PS is a valid alternative compared to ME for deriving brightness scales, with the advantage that it is more reliable and resilient to range bias.

Keywords:
Magnitude estimation, Partition scaling, Psychophysical methods, Brightness perception, Visual Perception

1 INTRODUCTION

Determining the relationship between the luminance of a stimulus and its perceived brightness is crucial for several fields such as colour appearance modelling, lighting design, estimating visual comfort criteria and lighting research in general. There are several psychophysical procedures to capture the relationship between the magnitude of a (psycho)physical stimulus and the magnitude of its corresponding percept (henceforth referred to as 'perceived magnitude'). One of the most frequently used methods establishing such a relation is magnitude estimation (ME), while one of the oldest methods is partition scaling (PS). Both methods are considered appearance based scaling procedures based on the classification scheme presented in (Kingdom and Prins 2016). Although PS is the oldest method, it is rarely used nowadays. On the other hand, ME is commonly used, but literature indicates that ME can suffer from several of the disadvantages discussed further below.

1.1 Magnitude Estimation

ME was initially used by Richardson and Ross (1930) and then further developed by S. S. Stevens in the 1950's (Moskowitz 1977; Stevens 1953). It is categorised as a non-forced-choice scaling procedure in which observers have to assign a numerical estimate for the perceived magnitude (e.g. brightness) of a test stimulus compared to the perceived magnitude of a reference stimulus. Both stimuli are usually displayed one after the other (sequentially) or at the same time (simultaneously). The reference stimulus is assigned a fixed numeric value or is chosen by the observers themselves. The observer numeric response results in a ratio of the perceived magnitude of the test and reference stimuli. As observers have to assign numeric estimates which represent ratio's, it is advised to test the math ability of observers prior to the experiment.

ME has been used to investigate the effects of adaptation on perceived brightness (Stevens and Stevens 1963) and is frequently used to develop colour appearance models (CAM) such as CIECAM02 (Luo and Rhodes 1999; Moroney *et al.* 2002), CAM15u (CIE 2004; Withouck *et al.* 2015b) and CAM18sl (Hermans *et al.* 2018a, 2019). However, ME can suffer from several disadvantages. Observers can find it difficult to provide a numerical estimate of perceived brightness, since it requires them to translate a perceptual sensation into a numeric representation. In addition, in an attempt by the observer to be self-consistent (Stevens 1955), the ME-method is also sensitive to sequential effects (order effects), wherein the current estimated value depends not only on the current stimulus but also on the recent history of shown stimuli (Ward 1979). Sequential effects can be partially counterbalanced by changing the presentation sequence of the stimuli for each observer (Cross 1973). Furthermore, ME can also be prone to centring bias as the observer tends to centre its range of responses on the range of presented stimuli (Poulton 1979). The result is an over- (under-) estimation of small (large) magnitudes in the tested stimuli dimension range. Centring bias is also known as 'regression effect', 'central tendency effect' or 'regression to the mean' (Thurley 2016). Another type of bias is range effect, whereby the outcome of the perceptual sensation depends, to some extent, on the selected stimulus range (Stevens 1975) and even on the selected value for the reference stimulus (Engen and Ross 1966). An overview of possible biases, together with a Bayesian framework for understanding biases in ME, is presented in (Petzschner *et al.* 2015).

Several analysis techniques exist for magnitude estimates (ASTM 2012; Han *et al.* 1999; Moskowitz 1977). Generally, the magnitude estimates are averaged across all observers for each stimulus using the geometric mean. A psychophysical scale is established by fitting the average observer ratios to the physical stimulus dimension (e.g. luminance).

1.2 Partition scaling

A partition scaling (PS) procedure constructs an interval scale of a psychological attribute (e.g. brightness) directly from the judgments of an observer without any translation from perceptual to numerical representation. A well-known experiment that used PS was conducted by Whittle (1992). Observers were instructed to adjust the brightness values of several discs in such a way that an equal-interval scale was created.

A PS procedure commonly consists of one or more bisection tasks, wherein an observer is shown two stimuli and is asked to adjust a third stimulus such that two equal-appearing perceptual intervals are produced. A progressive solution method is used and the two obtained equal-perceptual intervals are further progressively bisected in smaller intervals until the required number of equal-appearing perceptual intervals are obtained. The output of this progressive solution method is an equisection scale.

Like most psychophysical procedures, PS also suffers from some biases. For example, when using the progressive solution method, cumulative errors can occur, as each subsequent interval is dependent on the previously set interval. Order effects were found in a brightness experiment where the adjusted stimulus value was dependent on the order of presentation (Stevens 1961; Stevens and Stevens 1960). More details regarding possible PS biases and types of PS techniques can be found in (Van de Perre *et al.* 2019).

The CIE report on spatial brightness methodology (CIE 2014) provides an excellent overview of several biases found in various psychophysical procedures. While the report did not explicitly include ME or PS, the overview of possible biases, best practices and recommendations are also applicable to ME and PS.

In this paper, the ability of ME to derive a perceptual scale for brightness as a function of luminance is investigated and is compared to PS in terms of robustness and susceptibility to possible biases. Magnitude estimation and partition scaling experiments were conducted to determine brightness scales. These experiments were conducted as a part of a series of experiments in which different psychophysical procedures were investigated to produce brightness scales. The results of the ME experiment will be presented and compared to the results of the PS experiments. The latter is published in (Van de Perre *et al.* 2019).

2 METHODOLOGY

2.1 Experiment setup

A calibrated monitor (ColorEdge CG246; 60 Hz, 1920 x 1200 pixels and 10 bits per channel) was used to present neutral circular stimuli varying only in luminance. The background on the monitor surrounding the discs was black (< 0.5 cd/m²) and the room was completely darkened during the experiments (unrelated stimuli), with the exception of the observer monitor and experiment supervisor monitor (fully dimmed). Figure 1 shows an image of the experimental room on the left and a schematic layout of the room on the right. The monitor was calibrated up to approximately 180 cd/m² (CIE 1931 2° observer) using a colorimetric imaging camera (TechnoTeam LMK5-5 Color). Luminance accuracy was checked for each disc location and was within 1% of the requested values. Observers were sitting at a distance of roughly 60 cm from the monitor, resulting in a field of view for the monitor of approximately 40° and of about 10° for the central disc. The experiment was programmed in MATLAB R2018a and stimuli were generated using the Psychophysics toolbox (version 3.0.14) (Brainard 1997).

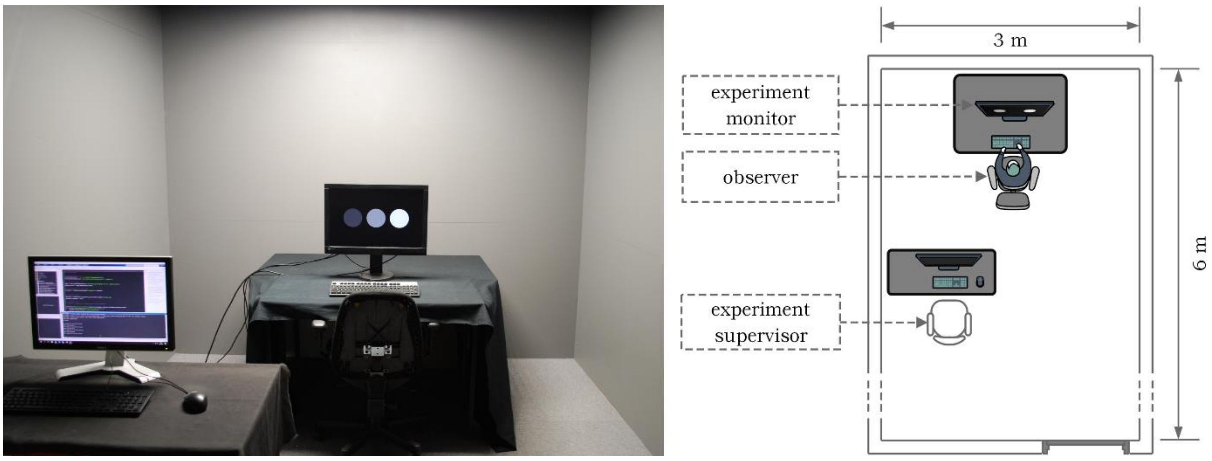

Figure 1 – Left: Image of experiment room. Note that during the experiment the stimuli and the supervisor monitor were the only light source in the room. In the image, the ambient lighting in the room was on for visual clarity. Right: A schematic layout of the experiment room.

2.2 Experiments

Brightness scales were determined for both experiment types: a magnitude estimation experiment and a partition scaling experiment. The experiments were divided over two sessions. The first session consisted of the ME and a PS experiment and observers started randomly with either. In this session, the lowest luminance anchor point in the PS experiment was always presented on the left on the monitor, data for the right position was obtained in session two.

Four luminance ranges were investigated for both experiment types. A *full-range*, from 5 cd/m² to 175 cd/m² and three subranges: a *low-range*, from 5 cd/m² to 82.3 cd/m²; a *mid-range*, from 51.4 cd/m² to 128.6 cd/m² and a *high-range* from 97.7 cd/m² to 175 cd/m². Each subrange had a 40% luminance overlap with its neighbouring range(s). The selected luminance values for each range per experiment type is shown in Figure 2.

Subdividing the full luminance range using an exponential or logarithmic function would result in more equally divided perceptual subranges. However, a linear subdivision was chosen to avoid any prior knowledge of the brightness scale which could potentially influence the results.

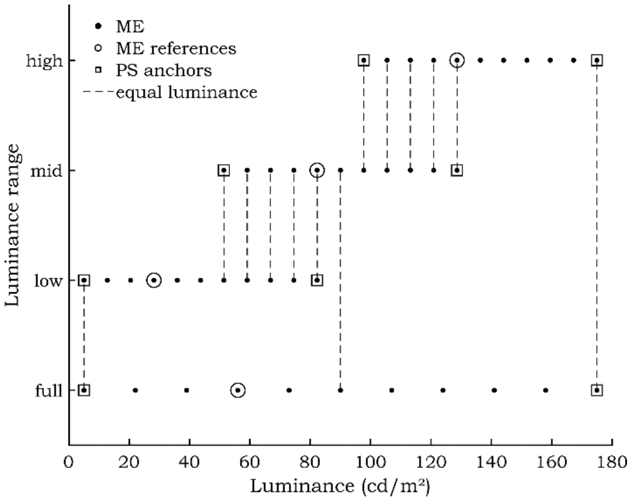

Figure 2 – The four luminance ranges (denoted on the vertical axis) are shown for both experiment types. A dot represents the stimuli in the ME experiment, while for each range, a circle indicates the ME reference stimuli and the two squares represent the PS anchors. Stimuli that have equal luminance are denoted with a dotted vertical line.

Fifteen observers – six females and nine males – with ages ranging from 24 to 30 years (average 26.3 years) participated in both sessions. Prior to the experiment, participants gave a written informed consent. The study was conducted in agreement with the social and societal ethics committee (SMEC) of KU Leuven. External participants were compensated at a rate of 10 €/hour.

2.2.1 Magnitude Estimation (ME)

During the ME experiment observers were asked to estimate the brightness of a neutral disc compared to that of a neutral reference disc with an assigned value of 100. If the test disc was twice as bright observers were asked to give a value of 200, if half as bright, a value of 50 and so on. The test and reference disc were presented simultaneously. The luminance of the reference disc for each range was chosen to be close to the ranges' geometric mean luminance (ASTM 2012). These were 56.0, 28.2, 82.3 and 128.6 cd/m² for the full-, low-, mid- and high-range, respectively (shown in Figure 2 as circles). For each range, 11 test (shown in Figure 2 as dots) and 3 repeated luminance values were selected and randomly presented to the observer twice. The minimum presentation time for both test and reference discs was 3 seconds. Between estimations a dark screen (< 0.5 cd/m²) was presented for 1 second. The position on the display (left/right) of the test and reference disc was randomly switched for each observer and was fully counterbalanced for each range. All observers started with the full-range followed by the subranges presented in a random order. Observers were given a break after completing all estimations of a range, but could pause at their request at any time. Each observer completed 28 (11 test and 3 repeated discs, presented left and right) brightness ratio estimations per range or 112 for all 4 ranges (not including the training phases).

At the beginning of the ME experiment observers were given brief verbal instructions (see appendix). Before each range, observers were asked to estimate the brightness of 4 training discs to familiarize them with the procedure and the luminance range. Two discs have a luminance equal to the range's 2 endpoints and two have a luminance value between that of the reference disc and the range's minimum and maximum luminance. The reference disc position was randomized per observer and per range and then kept fixed for all 4 training discs. After training, brightness evaluations were made for all 14 discs presented in random order keeping the reference position fixed. This was repeated, but with test and reference disc positions switched.

2.2.2 Partition scaling (PS)

In the PS experiment, an observer is shown two discs, one the left and one at the right of the display, each with a specific “anchor” luminance, and a third central disc. The observer task consists of bisecting the brightness difference between two anchor discs into two perceptually equal intervals by adjusting the luminance of the central disc. Two new sets of anchor points (luminance values) are thereby obtained, i.e. left-centre and centre-right, which can each be further progressively bisected in smaller luminance intervals as the level of subdivision increases, as shown in Figure 3. The PS experiment consisted of three subdivision levels, which results in eight equal perceived brightness intervals (partition scale) and nine stimulus luminance values (two fixed endpoints and seven observer adjusted discs show in Figure 3 as “F” and “A”, respectively).

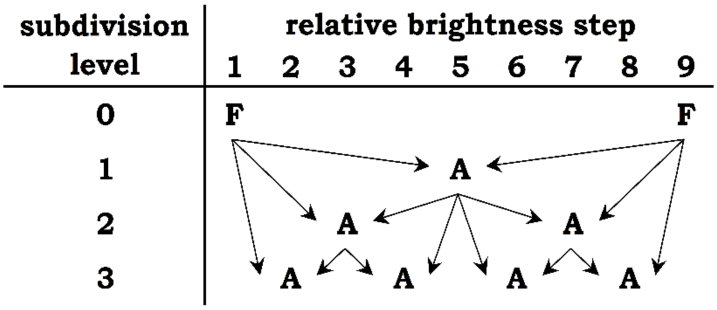

Figure 3 – Schematic overview of a partition scaling experiment with eight equal perceptual brightness intervals and three sublevels. The starting anchor discs (fixed anchor disc) are denoted as “F”. An adjustable disc is denoted with “A” and the corresponding anchor discs for that bisection task are denoted with two arrows going from the previous sublevel(s) to the current task.

The starting luminance of the central disc was either dark or bright and the position of the anchor points were swapped, which results in four ways of running the PS for each luminance range. In total, each observer completed 112 bisection tasks, i.e. 7 adjustable stimuli per complete PS times 4 luminance ranges times 4 possible ways running the PS. More details regarding the methodology of the PS experiment can be found in (Van de Perre *et al.* 2019).

Note that the number of collected data points obtained for each observer for both methods was equally balanced, 112 bisection tasks and 112 brightness ratio estimations.

2.3 Brightness scale models

Steven’s power law is used to model the luminance-brightness relationship (Stevens 1975). It states that the brightness (Q) changes as a function of luminance (L) as follows:

$$Q = aL^b \qquad (1)$$

Where a is a scaling factor and b the exponent of the power function. Commonly reported values for b in literature are between 1/3 and ½ (Stevens 1957, 1975; Withouck *et al.* 2015a). For ME, the brightness (Q) can be seen as the observer estimate and L is the luminance of the test disc. Note that the luminance-brightness relationship is specific to each luminance range due to the use of different reference luminance. A more appropriate, more general approach is to also consider the luminance and brightness values of the reference disc:

$$\begin{aligned} Q = \frac{ME_{test}}{ME_{ref}} &= \frac{aL_{test}{}^{b}}{aL_{ref}{}^{b}} \\ &= \left(\frac{L_{test}}{L_{ref}}\right)^{b} \end{aligned} \qquad (2)$$

The brightness (Q) is now considered a brightness ratio estimate between the brightness estimate of the test disc (ME_{test}) and the fixed brightness value of the reference disc (ME_{ref}). The luminance of the test and reference disc is L_{test} and L_{ref}, respectively. Equation (2) is referred to as the general model.

3 RESULTS & ANALYSIS

The results of the ME experiment are discussed first, followed by a comparison between the results of the ME and PS methods.

3.1 Magnitude estimation

First, the intra- and inter-observer variability is discussed, followed by the results of the separate and pooled luminance ranges for the average observer. Afterwards, the general brightness model is fitted to the individual observer data. Finally, a linear mixed-effects model is conducted, to investigate possible biases.

3.1.1 Intra- and inter-observer variability

The intra- and inter-observer variability were assessed using the standardized residual sum of squares (STRESS) (García *et al.* 2007; Melgosa *et al.* 2011). It is a measure of the goodness of fit between two sets of data, where 1 and 0 denote no and perfect agreement, respectively.

The intra-observer variability indicates how much an observer's response varies between observations of identical stimuli. The consistency of the observer's response (ME values) for the repeated stimuli per luminance range and per reference stimulus position was evaluated. STRESS values were calculated per observer and averaged over all observers for repeated stimuli per luminance range and reference position (Table 1). In addition, the consistency of an observer between all responses of left versus right reference position per luminance range has been calculated (Table 1). All arithmetic mean intra-observer STRESS values are below 0.15 and are in line or are lower than those obtained in similar ME brightness experiments (Hermans *et al.* 2018b; Withouck *et al.* 2015b). The small differences in STRESS values indicate that the difference between the left and right position of the reference is negligible, which indicates the absence of reference position bias. Results also show that the STRESS value decreases with higher luminance range, indicating observers being more consistent. However, observer feedback indicated that the higher subranges (mid- & high-range) were very difficult to estimate as the luminance difference (and hence perceived brightness) was almost indistinguishable. The decrease in STRESS value could therefore also be due to observers responding closer to the ME reference stimulus value.

Table 1. Mean intra-and inter-observer variability STRESS values. First row: average intra-observer consistency for repeated stimuli per luminance range and per reference position. Second row: average intra-observer consistency between all responses of left versus right reference position. Last row shows the average inter-observer variability per luminance range.

Type	Reference position:	Full-range		Low-range		Mid-range		High-range	
		Left	Right	Left	Right	Left	Right	Left	right
Intra	repeated stimuli	0.125	0.128	0.077	0.062	0.051	0.044	0.048	0.041
Intra	all stimuli	0.145		0.128		0.067		0.059	
Inter	all stimuli	0.098		0.089		0.056		0.044	

Inter-observer variability is a measure of how consistent an observer is with the 'average' observer's response. Per observer, one ratio estimate is calculated per unique stimulus for each range, averaging (repeated) stimuli for both left and right reference position, using the geometrical mean. The average observer was calculated by taking these values for all observers and calculating the geometric mean. The average inter-observer variability per luminance range was calculated using the arithmetic mean over each individual observer inter-observer variability (Table 1). The same effect is found as in intra-observer variability, STRESS values decrease with higher luminance range.

The duration for each observer task (estimating ratio) per observer was also recorded during the experiment. On average, observers estimated a brightness ratio in 8 seconds.

3.1.2 Brightness perception: individual luminance ranges

The observer estimates for the left and right reference position were combined for each luminance range, as there was no indications of reference position bias. The brightness perception of the average observer (shown in Figure 4) was estimated by calculating the geometric mean of the magnitude estimation ratios over all observers for each luminance range. The dashed lines indicate the reference stimulus luminance of each range. The error bars are standard errors of the geometric mean (Alf and Grossberg 1979). As can be expected, the standard errors increase with luminance difference between reference and test stimulus.

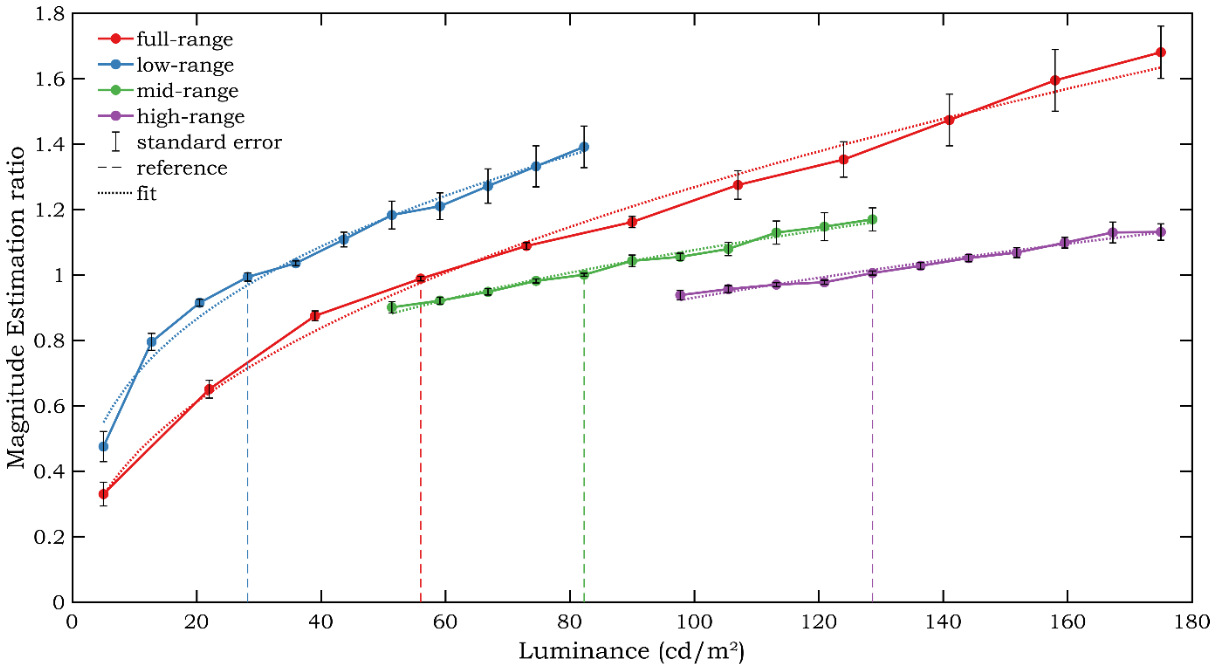

Figure 4 – Average observer brightness perception for each luminance range is shown. The luminance ranges: full-, low-, mid- and high-range are highlighted in red, blue, green and purple lines, respectively. The error bars are standard errors of the geometric mean. Each luminance range is fitted with a power law equation denoted as a coloured dotted line. The reference stimulus is indicated by a dashed line for each range in their respective colour.

For each luminance range a power-law brightness model has been fitted using (1) where Q is the average observer magnitude estimation ratio. The fits are shown in Figure 4 as the dotted lines. The parameters values and 95% confidence intervals (CI) of a and b, together with the coefficient of determination (R^2) of the fits, are shown in Table 2. The parameter a is a scaling factor and depends mostly on the scale units (determined by the choice of the brightness value assigned to the reference stimulus). Exponent b characterizes the shape of function and is therefore expected to be similar across all luminance ranges. However, this appears not to be the case. The b values for the subranges are similar, with a mean of 0.325, but are significantly different from the b value (0.452) of the full-range. The confidence interval of the full-range doesn't overlap with any of the subranges confidence intervals, indicating that a range effect could be present. All fits have very high R^2. However, note that the R^2 values are inflated, as the model has been fitted to the average observer data, whereby inter-observer variance was lost due to averaging.

Table 2. Fitting results per range with stimulus luminance as input and average observer brightness ratio as output. The 95% confidence intervals (CI) and values for parameters *a* and *b* are shown. The coefficient of determination (*R²*) of the model is shown in the last column.

	parameter a		**parameter b**		**R²**
	value	***CI***	***value***	***CI***	
full-range	0.158	[0.130, 0.186]	0.452	[0.415, 0.490]	0.993
low-range	0.324	[0.276, 0.372]	0.328	[0.290, 0.366]	0.983
mid-range	0.272	[0.236, 0.307]	0.299	[0.270, 0.328]	0.984
high-range	0.189	[0.150, 0.227]	0.347	[0.305, 0.388]	0.976

3.1.3 Brightness perception: pooled luminance ranges

In a classic brightness ME setup, the whole luminance range of interest would be conducted in one experiment. If the investigated luminance range is too large for observers to accurately estimate brightness, it is possible, although rather uncommon, to split the range into smaller subranges and conduct several ME experiments. The previous section showed possible evidence for a range effect between the full-range and the subranges in the present study. Pooling the subrange data and comparing them with the full-range gives more insight into the possible presence of a range effect. To pool the ME data of the subranges, they were transformed to a common scale using multiplicative scaling (linear transformation with zero offset) to ensure the ratio properties of a ME scale are kept intact. Although several approaches are possible, the following method was adopted. First, for each luminance range a brightness model is fitted with (1) using the luminance of the stimuli as input and the average observer brightness perception as output (show in Figure 4 as dotted lines). Second, a multiplicative scaling function is fitted to the predicted brightness values – using the fitted brightness models for each range – for the shared luminance interval between each subrange and the full-range. These scaling functions (factors) are then used to transform the brightness data of the subranges to that of the full-range (Figure 5).

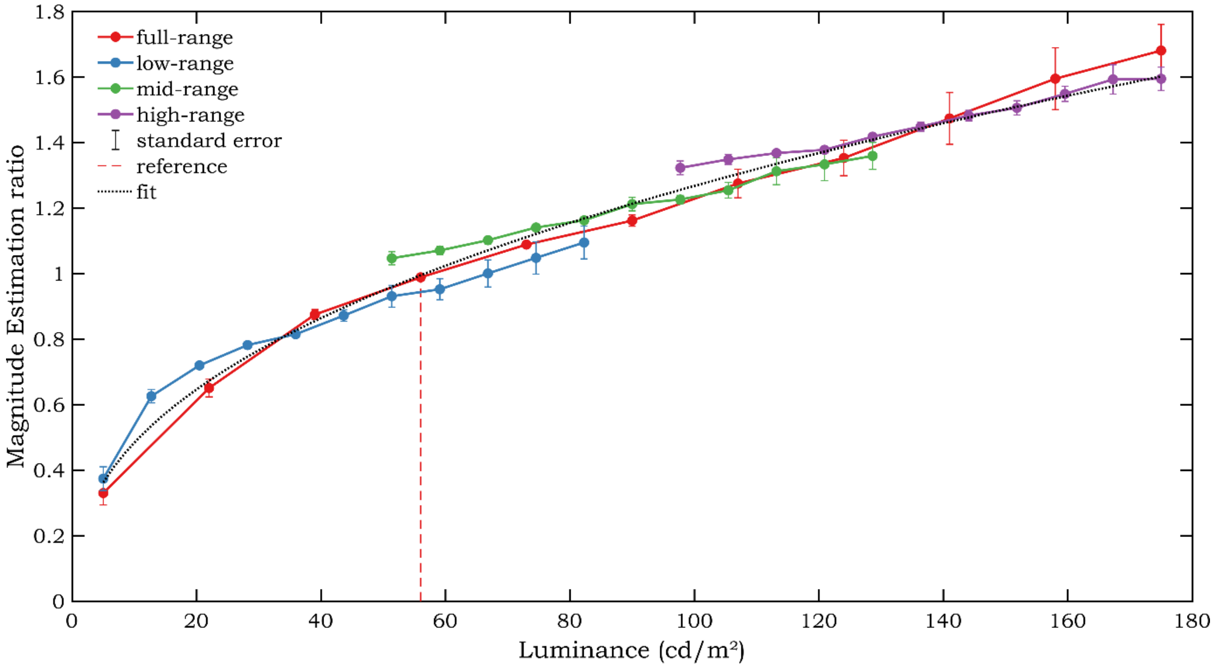

Figure 5 – Average observer brightness estimates rescaled to the full-range scale. The luminance ranges: full-, low-, mid- and high-range are highlighted in red, blue, green and purple, respectively. The reference stimulus luminance is indicated with a red dashed line. The error bars are rescaled standard errors of the geometric mean. The brightness model fitted to the pooled data is shown as a black dotted line.

Figure 5 indicates the presence of a centring bias: the high-end stimuli of the low-range are underestimated and the low-end stimuli of the mid-range are overestimated. A range effect is also visible, the estimations of the overlapping range for the low- & mid-range do not overlap or cross and have a substantial offset towards each other. Both effects are also present for the mid- & high-range. In (Petzschner *et al.* 2015; Thurley 2016) the main behavioural characteristics of ME are discussed and presented with figures showing several biases, such as range and regression effects (centring bias) for ME, similar to the effects seen in Figure 5.

The brightness model fitted to the pooled, rescaled brightness estimates of the average observer from all four ranges is shown in Figure 5 as a black dotted line. The fit had a very high R^2 of 0.982, fitted parameters were $a = 0.185$ (CI: [0.166, 0.204]) and $b = 0.418$ (CI: [0.396, 0.440]). The exponent b is similar to the b value (0.452) of the separate fit for the full-range, but is substantially higher than the fits to the individual subranges, again providing evidence for a range effect.

3.1.4 Brightness perception: a general model approach

The previous rescaling and pooling method is only applicable to ME data and is dependent on the presence of overlapping luminance ranges. The use of equation (2) does not require several overlapping luminance ranges and avoids the need for prior rescaling of the individual ranges. It also allows the comparison of b values over several luminance ranges and experimental methods.

The ME ratio is predicted using the luminance of test and reference stimuli and a brightness exponent b. Figure 6 shows each observer response as a dot. Different luminance ranges are indicated with different colours. Brightness models fitted to the data using (2) are shown as dotted coloured lines. The surface fitted to the brightness data of all four luminance ranges is plotted as the black grid in Figure 6.

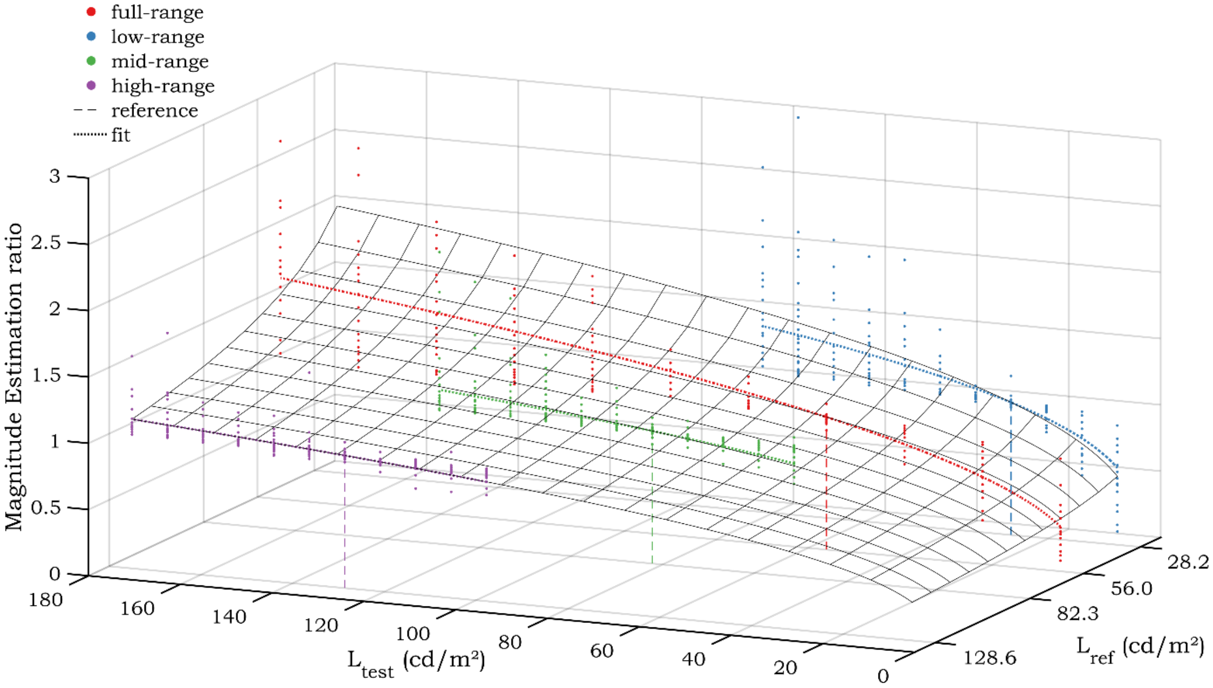

Figure 6 – The individual observer brightness ratio values for each luminance range are shown as dots. Data for the full-, low-, mid- and high-range are highlighted in red, blue, green and purple, respectively. The reference stimulus luminance for each range is indicated with a coloured dashed line. Each luminance range is fitted with the general model using (2) denoted as a coloured dotted line. The brightness model fitted to the pooled data is plotted as a black grid.

Values for the exponent b, its 95% confidence intervals and the coefficients of determination (R^2) for fits to the data are shown in Table 3 for the individual ranges and several pooled ranges. The general model fits (using (2)) for the individual ranges result in substantially lower

R^2 values compared to the corresponding fits using (1) (see Table 2). Reasons are two-fold. First, the results in Table 2 were obtained by fitting to average observer data, while those in Table 3 by fitting to the individual observer data. By averaging the data before model fitting, the spread due to inter-observer variability does not contribute to the residuals, resulting in a higher R^2. Second, (2) has only one free parameter, while (1) has two. This extra degree of freedom can increase the R^2.

From Table 3, it can be seen that the R^2 value for the full-range was highest, whereas the R^2 values for the subranges show a systematic decrease from the low- to the high-range. This indicates observers become more and more inconsistent with increasing luminance range. This is also consistent with observer feedback stating that the higher subranges (mid- & high-range) were harder to estimate and that they could sometimes barely make out any brightness difference between stimuli.

For each luminance range, the fitted b values using (1) are similar to the fitted b values using (2). It is expected that the exponent b should be similar across all luminance ranges. However, as before, this appears not to be the case. With a value of 0.319, the exponent b obtained for the pooled subranges (low- to high-range) value is significantly different from that of the full-range (0.450), as the CIs do not overlap. This again indicates that a range effect could be present. Based on the substantially lower R^2 values of the mid- & high-range subranges (0.369 and 0.357), including any of these ranges when pooling the data could bias the outcome (since observers had difficulties with consistently estimating these subranges). The R^2 values of the full- & low-range are substantially higher. However, their b values (resp. 0.450 and 0.315) differ substantially from each other, as their confidence intervals do not overlap. This significant difference was also found using the average observer data and again indicates a range effect between the full- and low-range.

Pooling the full- & low-range and comparing with a pooling of all ranges results in nearly identical b values (0.385 and 0.382), with similar R^2 values, but slightly higher for fits to the pooled full- & low-range data. Despite the low R^2 values of the separate fits to the mid- & high-range data, the impact on the fitting results of including these ranges (when pooling the data) is very small. This indicates that luminance ranges that were (too) difficult to estimate can still be pooled without substantially influencing the results, although a range effect is still present when comparing the full-range with either the pooled full- & low-range or all ranges pooled.

Table 3. Exponent b, 95% CI and R^2 for fits of the general model to the individual observer data for each of the luminance ranges and several pooling combinations. The first four rows show fits for separate ranges and the last three rows show fits for the pooled ranges.

(pooled) ranges	parameter *b*		R^2
	value	*CI*	
Full-range	0.450	[0.427, 0.472]	0.720
Low-range	0.315	[0.294, 0.337]	0.587
Mid-range	0.336	[0.299, 0.372]	0.369
High-range	0.385	[0.343, 0.428]	0.357
Full- & low-range	0.385	[0.369, 0.401]	0.649
All ranges	0.382	[0.370, 0.394]	0.633
Low- to high-range	0.319	[0.305, 0.333]	0.546

3.1.5 Range bias b exponent values

Figure 5 and the discrepancy in b values between the full-range and the subranges in Table 2 and Table 3 suggest the presence of a range bias. This was confirmed by performing a linear mixed-effects model (LMM) analysis using the *R* language (R Core Team 2019) and package *lme4* (Bates *et al.* 2015). Restricted maximum likelihood was used as estimation method for the model. The fixed effect luminance range was considered as categorical and each observer had an intercept as the only random effect. Visual inspection of residual plots and Q-Q plots did not reveal any deviations from homoscedasticity or normality. The residuals were tested on normality using Shapiro-Wilk's method and normality is assumed ($p = 0.591$). A Wald Chi-square test indicated that the fixed effect luminance range was significant ($\chi^2(3) = 11.957, p = 0.008$). Contrast tests using Bonferroni adjustment for multiple comparisons revealed that the b values between the full- and low-range were significantly different ($p = 0.016$). All other pair of ranges were found to have no significant difference in b values. This confirms the range bias between the full- & low-range using ME. A similar analysis in (Van de Perre *et al.* 2019) using the results of the PS experiment showed that there was a range bias between all pairs, except the full- & low-range. However, the mid- & high-range were not reliable as observers had difficulties estimating these ranges because the brightness of stimuli were too similar for these ranges.

3.2 Magnitude Estimation versus Partition scaling

Observers could better estimate the full- & low-range for both ME and PS (Van de Perre *et al.* 2019). These results are compared using the average and individual observer data.

3.2.1 Pooled luminance ranges

Figure 7 shows the pooled average observer brightness estimates of the full- & low-range, rescaled to the full-range for ME and PS, on the left and right, respectively.

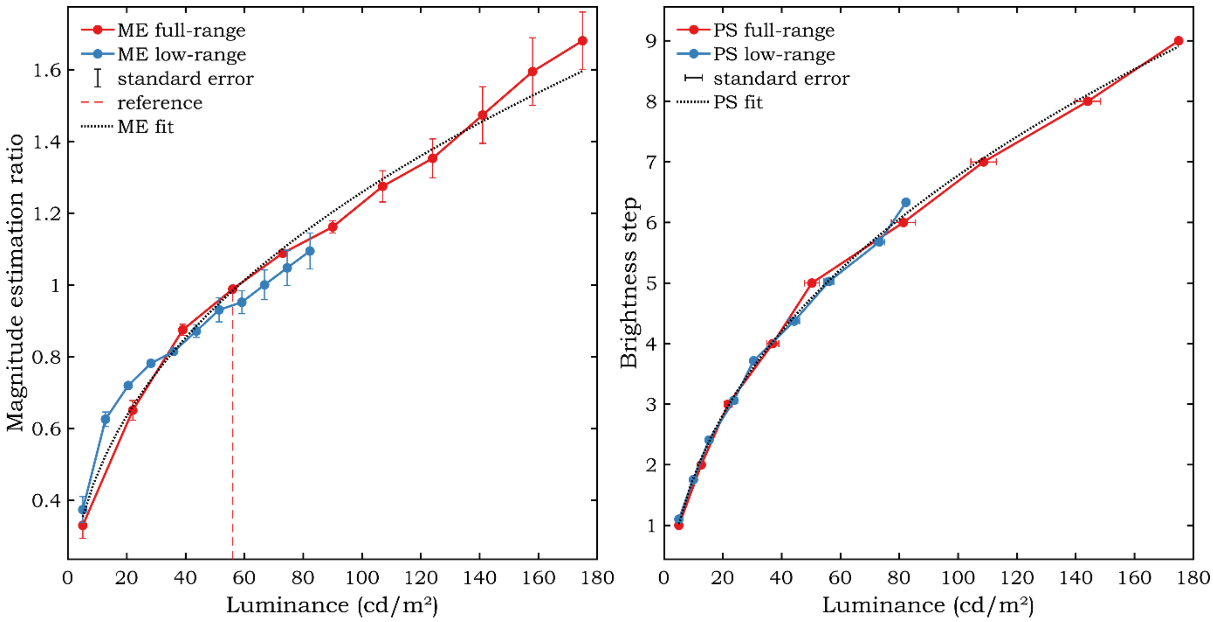

Figure 7 – Average observer brightness estimates of the full- & low-range rescaled to the full-range, for ME and PS on the left and right, respectively. The luminance ranges full- and low-range are highlighted in red and blue, respectively. The ME reference stimulus luminance is indicated with a red dashed line. The errors bars for ME are rescaled standard errors on the geometric mean, whereas for PS these are standard errors on the arithmetic mean per observer. A brightness model is fitted to the pooled data as a black dotted line.

The results show that both experiments have similar curvature (related to exponent b). However, the full- & low-range results for PS align better compared to those for ME. In the ME plot, an offset is visible due to range bias. Figure 7 also shows a brightness model fitted to the pooled data of ME and PS using (1). As PS generates an interval scale it requires an additional constant (c). The fitting results (coefficient values with 95% CI and R^2) are shown in Table 4. The fitted b values show remarkably similar values with overlapping CI, indicating that ME or PS brightness experiments could lead to similar results when pooling luminance ranges, despite both methods generate a ratio and interval scale, respectively. However, the rescaled average observer data for ME (Figure 5) shows a clear effect of range bias, while for PS this effect was not present. The PS results for the average observer data reported in (Van de Perre *et al.* 2019) also have higher R^2 values compared to the ME average observer results reported in Table 2.

Table 4. Fitting results for average observer data of the pooled full- & low-range for both types of experiment.

type	parameter *a*		parameter *b*		parameter *c*		R^2
	value	*conf. int.*	*value*	*conf. int.*	*value*	*conf. int.*	
ME	0.178	[0.153, 0.204]	0.424	[0.391, 0.457]			0.980
PS	1.168	[0.814, 1.522]	0.419	[0.370, 0.468]	-1.273	[-1.899, -0.646]	0.998

3.2.2 Brightness exponents for individual observers

The general model approach allows to assess any difference in the brightness function derived from data obtained with both types of experiment. Table 5 shows the exponent b, 95% CI and R^2 for fits, using the general brightness model, to the individual observer data obtained using both type of experiments for the full-, low- and pooled range. It is clear that the obtained R^2 values for fits to the individual observer data are substantially lower compared to those of fits to the average observer data. However, the lowest R^2 of any individual range in PS was higher than the highest *R²* in ME, possibly indicating that PS leads to more robust results. The CI of exponent b between the full- and low-range do not overlap for ME, whereas there was a slight overlap for PS. Both methods lead to similar exponent b values for the pooled full- & low-range. However, for PS the *R²* (0.948) were substantially higher than for ME (0.649), indicating that PS results were more robust. Comparing the pooled full- & low-range results of the brightness model using average observer data and individual observer data (general brightness model) shows that PS lead to almost equal brightness exponents, while there was a slight drop in the value of exponent b for ME.

Table 5. Fitting results for the individual observer data using the general brightness models for both types of experiment. The full, low and pooled luminance ranges is shown.

Method	(pooled) ranges	parameter *b*		*R²*
		value	conf. int.	
ME	**Full-range**	0.450	[0.427, 0.472]	0.720
ME	**Low-range**	0.315	[0.294, 0.337]	0.587
ME	**Full- & low-range**	0.385	[0.369, 0.401]	0.649
PS	**Full-range**	0.403	[0.363, 0.444]	0.942
PS	**Low-range**	0.490	[0.436, 0.543]	0.936
PS	**Full- & low-range**	0.414	[0.384, 0.444]	0.948

Figure 8 shows a boxplot of the b-values, for each experiment (ME and PS) and luminance range (separate and pooled full- & low-range), fitted to each individual observer's data using the general model. A substantial inter-individual variation in brightness exponents b is visible within each boxplot, although slightly larger for PS. Literature shows that the power exponent varies across individuals. In (Tsubomi *et al.* 2012), the 'inner psychophysics' of brightness perception within individuals was investigated using ME and fMRI (functional magnetic resonance imaging). The power function exponents for subjective brightness ratings varied from 0.14 to 0.46 with a mean of 0.32 for nine observers. A disk subtending 10° in diameter and with luminance ranging from 1 to 100 cd/m² was presented to the observers which were instructed to rate brightness using a ME procedure. These conditions and results are somewhat similar with the low-range ME results where a value of 0.315 for the exponent b value was found using the general model.

A LMM with the individual observer b values as input has been used to analyse potential differences between both methods and between full-, low- and the pooled full- & low-range. Restricted maximum likelihood was used as estimation method for the model. The fixed effects luminance range and method type was considered as categorical and each observer had an intercept as only random effect. Visual inspection of residual plots and Q-Q plots did not reveal any deviations from homoscedasticity or normality. The residuals were also tested on normality using Shapiro-Wilk's method and normality is assumed ($p = 0.498$). A Wald Chi-square test indicated that the fixed effect luminance range was significant ($\chi^2(2) = 10.822, p = 0.004$), the fixed effect method was not significant ($\chi^2(1) = 0.930, p = 0.335$), but the interaction effect was ($\chi^2(2) = 15.247, p \leq 0.001$).

Comparing the full-range with the low-range shows a substantial drop in exponent b values for ME, whereas for PS only a moderate increase occurred. Contrast tests using Bonferroni adjustment for multiple comparisons revealed that there was only one pair of luminance ranges with a significant difference namely between the full- and low-range for ME ($p = 0.005$) whereas for PS this pair had no significant difference ($p = 0.109$). This confirms previous results that there was a range bias between the full- & low-range for ME but not for PS. When comparing the pooled full- & low-range no significant difference ($p = 0.401$) was found between both methods. This is consistent with previous results (Figure 7), indicating that ME and PS converge to the same results when pooling the full- and low-range.

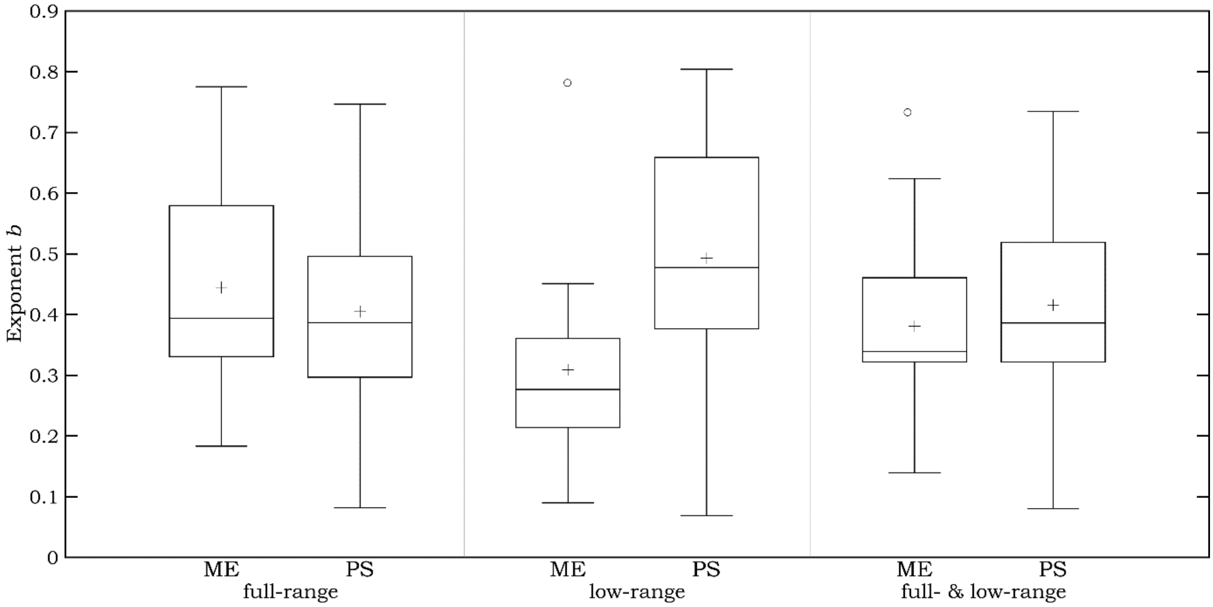

Figure 8 – Boxplots of the exponent *b* values fitted to the individual observers' data using the general model. The boxplots are structured by range (separate and pooled full- & low-range) and experiment method (ME and PS). Per boxplot the mean is shown as a '+' and the median is a horizontal solid black line. Outliers are shown as an 'o'.

4 CONCLUSION

Psychophysical data was obtained to derive a perceptual brightness/luminance scale using two appearance-based methods, namely ME and PS. Both psychophysical experiments were conducted with simple neutral circular stimuli (disks) presented on a calibrated monitor with a dark background in a dark environment. Four luminance ranges were investigated for each method. A full-range, from 5 cd/m² to 175 cd/m² and three subranges (low-, mid- and high-range) were carefully chosen with luminance intervals equally spaced over the full luminance range.

For ME, the intra- and inter-observer variability (assessed using STRESS) decreased when evaluating luminance ranges at high luminance. However, observers reported having more difficulties making estimates at higher luminance ranges (mid- & high-range), as the perceived brightness ratio between test and reference stimulus was sometimes too indistinguishable. A similar effect was found using PS.

The observer's inability to accurately estimate the mid- and high-range was not reflected in the very high R^2 values of fits based on average observer data. Whereas this was clearly visible with the R^2 values of fits using the individual observer data, as these were substantially lower compared to the R^2 of the full- & low-range. This indicates that R^2 on average observer data is not necessarily a good estimate to indicate whether or not observers could actually complete the task correctly.

The inter-individual spread of the brightness exponent value b is large for both types of experiments, in line with literature. Reporting the spread of the observer data and the confidence intervals of the fitted parameters should be considered as best practice.

Pooling all ranges and rescaling using the full-range as a 'common scale' shows no indication for a range bias with PS, whereas there is a clear range bias using ME. For the latter, substantial offsets between neighbouring subranges and the full-range were found. However, ME and PS converge to approximately similar fitted brightness exponents when combining the full- & low-range, ranges that observers could accurately estimate. This indicates that a range bias with ME might be counterbalanced by estimating and combining several luminance ranges.

When comparing R^2 values of fits using the general model on individual observer data, PS had overall higher R^2 values compared to ME, even so that the lowest R^2 of the individual ranges for PS was higher than the highest R^2 for ME, indicating that PS is a more robust method. However, PS is not flawless; it can be prone to cumulative errors, as each subsequent bisection interval is dependent on the previously set interval. Although, this can be counterbalanced by conducting several repeats of the same luminance range or use an improved PS method, as proposed in (Van de Perre *et al.* 2019). While PS is more reliable than ME, it is not always a viable option. This is the case when only one stimulus can be shown at the same time, as it can be difficult to use a sequential presentation, which needs switching between the two anchor stimuli and the adjustable stimuli. Another example is when investigating multiple stimulus dimensions. Then PS can be difficult to implement because a stimulus is adjustable in more than one dimension (e.g. coloured stimuli: hue and saturation). In both cases, ME could be a suitable alternative. A disadvantage of ME is that stimulus values need to be selected prior to the experiment whereas for PS only a selection of the investigated stimulus range is required. In any case, both methods are viable alternatives given different circumstances. However, when using ME one should consider the potential range bias and counterbalancing is mandatory.

ACKNOWLEDGMENT

The authors want to express their gratitude towards the observers and the anonymous reviewers for their constructive feedback.

FUNDING

This research was supported by Impulse Fund KU Leuven (IMP/16/027 – "*Correlaties tussen objectieve en subjective descriptoren van binnenverlichting*"), granted to W. R. Ryckaert and Marc Dujardin. Kevin A. G. Smet was supported by Internal KU Leuven Funds (STG/17/013).

DISCLOSURE STATEMENT

The authors have no financial interests to declare.

REFERENCES

Alf, E.F. and Grossberg, J.M., 1979. The geometric mean: Confidence limits and significance tests. *Attention, Perception, & Psychophysics*, 26 (5), 419–421.

ASTM, 2012. *Standard Test Method for Unipolar Magnitude Estimation of Sensory Attributes*. West Conshohocken, No. E1697-05:e1.

Bates, D., Mächler, M., Bolker, B., and Walker, S., 2015. Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, 67 (1), 1–48. doi: 10.18637/jss.v067.i01.

Brainard, D.H., 1997. The Psychophysics Toolbox. *Spatial Vision*, 10 (4), 433–436. doi: 10.1163/156856897X00357.

CIE, 2004. *A colour appearance model for colour management systems: CIECAM02*. Vienna: Commission Internationale de L'Eclairage, Technical Report No. CIE 159:2004.

CIE, 2014. *Guidance towards Best Practice in Psychophysical Procedures Used when Measuring Relative Spatial Brightness*. Vienna: Commission Internationale de L'Eclairage, Technical Report No. CIE 212:2014.

Cross, D. V., 1973. Sequential dependencies and regression in psychophysical judgments. *Perception & Psychophysics*, 14 (3), 547–552. doi: 10.3758/BF03211196.

Engen, T. and Ross, B.M., 1966. Effect of reference number on magnitude estimation. *Perception & Psychophysics*, 1 (1), 74–76. doi: 10.3758/BF03207825.

García, P. a, Huertas, R., Melgosa, M., and Cui, G., 2007. Measurement of the relationship between perceived and computed color differences. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 24 (7), 1823–1829. doi: 10.1364/JOSAA.24.001823.

Han, S.H., Song, M., and Kwahk, J., 1999. A systematic method for analyzing magnitude estimation data. *International Journal of Industrial Ergonomics*, 23 (5–6), 513–524. doi: 10.1016/S0169-8141(98)00017-1.

Hermans, S., Smet, K.A.G., and Hanselaer, P., 2018a. Color appearance model for self-luminous stimuli. *Journal of the Optical Society of America A*, 35 (12), 2000. doi: 10.1364/JOSAA.35.002000.

Hermans, S., Smet, K.A.G., and Hanselaer, P., 2018b. Brightness Model for Neutral Self-Luminous Stimuli and Backgrounds. *LEUKOS - Journal of Illuminating Engineering Society of North America*, 1–14. doi: 10.1080/15502724.2018.1448280.

Hermans, S., Smet, K.A.G., and Hanselaer, P., 2019. Exploring the applicability of the CAM18sl brightness prediction. *Optics Express, Vol. 27, Issue 10, pp. 14423-14436*, 27 (10), 14423–14436. doi: 10.1364/OE.27.014423.

Kingdom, F.A.A. and Prins, N., 2016. *Psychophysics. A Practical Introduction*. 2nd ed. Academic Press.

Luo, M.R. and Rhodes, P.A., 1999. Corresponding-colour datasets. *Color Research and Application*, 24 (4), 295–296. doi: 10.1002/(SICI)1520-6378(199908)24:4<295::AID-COL10>3.0.CO;2-K.

Melgosa, M., García, P.A., Gómez-Robledo, L., Shamey, R., Hinks, D., Cui, G., and Luo, M.R., 2011. Notes on the application of the standardized residual sum of squares index for the assessment of intra- and inter-observer variability in color-difference experiments. *Journal of the Optical Society of America A*, 28 (5), 949–953. doi: 10.1364/JOSAA.28.000949.

Moroney, N., Fairchild, M.D., Hunt, R.W.G., Li, C., Luo, M.R., and Newman, T., 2002. The CIECAM02 Color Appearance Model. *In*: *Color and Imaging Conference*. Society for Imaging Science and Technology, 23–27 doi: 10.1.1.77.8398.

Moskowitz, H.R., 1977. MAGNITUDE ESTIMATION: NOTES ON WHAT, HOW, WHEN, AND WHY TO USE IT. *Journal of Food Quality*, 1 (3), 195–227. doi: 10.1111/j.1745-4557.1977.tb00942.x.

Van de Perre, L., Ryckaert, W.R., Dujardin, M., Hanselaer, P., and Smet, K.A.G., 2019. Derivation of Brightness Scales Using Partition Scaling. *LEUKOS*, 1–15. doi: 10.1080/15502724.2019.1635890.

Petzschner, F.H., Glasauer, S., and Stephan, K.E., 2015. A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, 19 (5), 285–293. doi: 10.1016/j.tics.2015.03.002.

Poulton, E.C., 1979. Models for biases in judging sensory magnitude. *Psychological Bulletin*, 86 (4), 777–803. doi: 10.1037/0033-2909.86.4.777.

R Core Team, 2019. R: A Language and Environment for Statistical Computing.

Richardson, L.F. and Ross, J.S., 1930. Loudness and telephone current. *Journal of General Psychology*, 3 (2), 288–306. doi: 10.1080/00221309.1930.9918206.

Stevens, J.C. and Stevens, S.S., 1963. Brightness Function : Effects of Adaptation. *Journal of the Optical Society of America*, 53 (3), 375. doi: 10.1364/JOSA.53.000375.

Stevens, S.S., 1953. On the brightness of lights and the loudness of sounds. *Science*, 118, 576.

Stevens, S.S., 1955. The measurement of loudness. *The journal of the acoustical society of america*, 27 (5), 815–829.

Stevens, S.S., 1957. On the psychophysical law. *Psychological Review*, 64 (3), 153–181. doi: 10.1037/h0046162.

Stevens, S.S., 1961. To Honor Fechner and Repeal His Law. *Science*, 133 (3446), 80–86. doi: 10.1126/science.133.3446.80.

Stevens, S.S., 1975. *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York Wiley.

Stevens, S.S. and Stevens, J.C., 1960. *The dynamics of visual brightness*. Psychophysical Laboratory Monograph: Harvard University.

Thurley, K., 2016. Magnitude Estimation with Noisy Integrators Linked by an Adaptive Reference. *Frontiers in Integrative Neuroscience*, 10, 6. doi: 10.3389/fnint.2016.00006.

Tsubomi, H., Ikeda, T., and Osaka, N., 2012. Primary visual cortex scales individual's perceived brightness with power function: Inner psychophysics with fMRI. *Journal of Experimental Psychology: Human Perception and Performance*, 38 (6), 1341–1347. doi: 10.1037/a0030025.

Ward, L.M., 1979. Stimulus information and sequential dependencies in magnitude estimation and cross-modality matching. *Journal of Experimental Psychology: Human Perception and Performance*, 5 (3), 444–459. doi: 10.1037/0096-1523.5.3.444.

Whittle, P., 1992. Brightness, discriminability and the 'Crispening Effect'. *Vision Research*, 32 (8), 1493–1507. doi: 10.1016/0042-6989(92)90205-W.

Withouck, M., Smet, K.A.G., and Hanselaer, P., 2015a. Brightness prediction of different sized unrelated self-luminous stimuli. *Optics Express*, 23 (10), 13455–13466. doi: 10.1364/OE.23.013455.

Withouck, M., Smet, K.A.G., Ryckaert, W.R., and Hanselaer, P., 2015b. Experimental driven modelling of the color appearance of unrelated self-luminous stimuli: CAM15u. *Opt Express*, 23 (9), 12045–12064. doi: 10.1364/OE.23.012045.

APPENDIX: OBSERVER INSTRUCTIONS

The instructions for the ME experiment were given verbally in Dutch or English, depending on the preference of the observer. Text highlighted in italic were the verbally given instructions to the observer.

English instructions:
"You will see two circles one left and one right, the task is to give a brightness estimate for one circle compared to the other, called test and reference circle, respectively. The reference circle is assigned a fixed number of 100. If you perceive the test circle compared to the reference circle twice as bright you give a brightness value of 200, if half as bright, a value of 50 and so on. There is no upper limit to the value of brightness. If the test circle is three times as bright compared to the reference circle what brightness value would you give? (this was asked to make sure the observer understood the brightness estimates using numeric representation). *You are not limited to integer numbers or multiples of 10 or any other kind, for example you can also say 25.5 or 123 as a brightness value. So you can use any kind of numbers you like, fractions, decimals or whole numbers. You also don't need to worry about consistency, try to estimate the current brightness regardless of previously given estimates. The keyboard is not needed, whenever you give a brightness estimate I will register your response. There is no time limit imposed when estimating a brightness value. You can take a break whenever you need by asking me* (the experimenter in question). *In total there will be four sub parts, each subpart will start with a training phase followed by several brightness estimates then the reference and test circle will switch positions, again followed by several brightness estimates. I will inform you whenever the reference circle changes position. It is*

also possible to take a break after each subpart." If the observer had no more questions the ME experiment started with the training phase of the full luminance range.

Dutch instructions:
"Je zal twee cirkels zien, de ene links en de andere rechts, de taak is om een schatting te geven van de helderheid van de ene cirkel ten opzichte van de andere, respectievelijk de test- en referentiecirkel genoemd. De referentiecirkel krijgt een vast getal van 100 toegewezen. Als je de testcirkel ten opzichte van de referentiecirkel twee keer zo helder waarneemt dan geef je een helderheidswaarde van 200, als deze de helft zo helder is, een waarde van 50 enzoverder. Er is geen bovengrens aan de waarde van de helderheid. Als de testcirkel drie keer zo helder is als de referentiecirkel, welke helderheidswaarde zou je dan geven? Je bent niet beperkt tot gehele getallen of veelvouden van 10 of een ander soort, je kunt bijvoorbeeld ook 25.5 of 123 als helderheidswaarde geven. Je kan dus alle soorten getallen gebruiken die je wilt, zoals breuken, decimalen of gehele getallen. Je hoeft je ook geen zorgen te maken over de consistentie van je helderheidswaarde, probeer de huidige helderheid te schatten, ongeacht van eerder gegeven schattingen. Het toetsenbord is niet nodig, wanneer je een helderheidsschatting geeft zal ik je schatting noteren. Er is geen tijdslimiet opgelegd bij het schatten van een helderheidswaarde. Je kunt een pauze nemen wanneer je maar wilt door mij te informeren. In totaal zullen er vier subdelen zijn, elk subdeel zal beginnen met een trainingsperiode gevolgd door verschillende helderheidsschattingen, dan zal de referentie- en testcirkel van positie veranderen, gevolgd door weer verschillende helderheidsschattingen. Ik zal je informeren wanneer de referentiecirkel van positie verandert. Het is ook mogelijk om na elk subdeel een pauze in te lassen".