Attention, Perception, & Psychophysics (2020) 82:877–890
https://doi.org/10.3758/s13414-019-01862-w

40 YEARS OF FEATURE INTEGRATION: SPECIAL ISSUE IN MEMORY OF ANNE TREISMAN

Task-driven and flexible mean judgment for heterogeneous luminance ensembles

Yusuke Takano[1] · Eiji Kimura[2]

Published online: 12 September 2019

Abstract

Spatial averaging of luminances over a variegated region has been assumed in visual processes such as light adaptation, texture segmentation, and lightness scaling. Despite the importance of these processes, how mean brightness can be computed remains largely unknown. We investigated how accurately and precisely mean brightness can be compared for two briefly presented heterogeneous luminance arrays composed of different numbers of disks. The results demonstrated that mean brightness judgments can be made in a task-dependent and flexible fashion. Mean brightness judgments measured via the point of subjective equality (PSE) exhibited a consistent bias, suggesting that observers relied strongly on a subset of the disks (e.g., the highest- or lowest-luminance disks) in making their judgments. Moreover, the direction of the bias flexibly changed with the task requirements, even when the stimuli were completely the same. When asked to choose the brighter array, observers relied more on the highest-luminance disks. However, when asked to choose the darker array, observers relied more on the lowest-luminance disks. In contrast, when the task was the same, observers' judgments were almost immune to substantial changes in apparent contrast caused by changing the background luminance. Despite the bias in PSE, the mean brightness judgments were precise. The just-noticeable differences measured for multiple disks were similar to or even smaller than those for single disks, which suggested a benefit of averaging. These findings implicated flexible weighted averaging; that is, mean brightness can be judged efficiently by flexibly relying more on a few items that are relevant to the task.

Keywords Ensemble perception · Summary statistics · Averaging · Brightness · Lightness

The visual system can summarize complex scenes by rapidly forming statistical summary descriptions of sets of similar items. For example, rapid and accurate averaging has been reported with many visual features, including motion (Watamaniuk & Duchon, 1992; Watamaniuk, Sekuler, & Williams, 1989), orientation (Dakin & Watt, 1997; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), and size (Ariely, 2001; Chong & Treisman, 2003, 2005; for reviews, see Alvarez, 2011; Bauer, 2015; Whitney, Haberman, & Sweeny, 2014). Such statistical processing, called *ensemble coding*, has also been found for higher-level properties of images, including the emotion and gender of faces (Haberman & Whitney, 2007, 2009), the headings of point-light walkers (Sweeny, Haroz, & Whitney, 2013), or the semantic meaning of symbols (Corbett, Oriet, & Rensink, 2006; Sakuma, Kimura, & Goryo, 2017; Van Opstal, de Lange, & Dehaene, 2011). Considering the statistical regularity and redundancy of the natural world, this ability can be critical to efficiently representing natural scenes.

Understanding how this ensemble coding is implemented has attracted great interest. Early studies argued that the extraction of ensemble means involves a parallel and holistic process (Ariely, 2001; Chong & Treisman, 2003, 2005). Supporting evidence included findings that the mean judgment was very efficient and effortless; it was not affected by variations in display size (Ariely, 2001), and was relatively immune to variations in stimulus duration and in item distributions (Chong & Treisman, 2003). Chong and Treisman (2003) argued for the importance of parallel processes to

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13414-019-01862-w) contains supplementary material, which is available to authorized users.

✉ Eiji Kimura
eiji.kimura@chiba-u.jp

[1] Graduate School of Science and Engineering, Chiba University, Chiba-shi, Chiba, Japan

[2] Department of Psychology, Graduate School of Humanities, Chiba University, Chiba-shi, Chiba, Japan

preattentively pool feature information in order to represent the statistical properties of visual scenes, and they discussed a link to parallel processing in feature search tasks (Treisman & Gormican, 1988). However, subsequent studies indicated that alternative strategies, such as subset sampling, could also account for the results (e.g., Myczek & Simons, 2008). This claim has aroused much controversy (Ariely, 2008; Chong, Joo, Emmanouil, & Treisman, 2008; Simons & Myczek, 2008), but there is now converging evidence for the notion that not all items are uniformly weighted in the computation of ensemble mean representations. Subsampling in ensemble coding has been studied for various perceptual attributes (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013; de Gardelle & Summerfield, 2011; Kanaya, Hayashi, & Whitney, 2018; Marchant, Simons, & de Fockert, 2013; Maule & Franklin, 2016; Solomon, Morgan, & Chubb, 2011).

Despite this recent surge in the study of ensemble coding, there has been little work regarding ensemble means of brightness. This may reflect the belief that for low-level visual attributes such as brightness, or color in general, the mean could be encoded by early visual mechanisms with large receptive fields; the mechanisms simply pool local features in order to represent the mean (see also Myczek & Simons, 2008). In fact, spatial averaging of luminances over a variegated region has been assumed in visual processes involved in light adaptation (e.g., Buchsbaum, 1980; Shapley & Enroth-Cugell, 1984), texture segmentation (e.g., Chubb, Econopouly, & Landy, 1994; Chubb, Landy, & Econopouly, 2004), and lightness scaling (e.g., Bressan, 2006). However, despite the importance of these visual processes, how ensemble mean brightness can be computed remains largely unknown.

One exception is Bauer (2009), who investigated mean brightness for a briefly presented stimulus set of achromatic disks. Observers judged whether a single probe disk presented following the stimulus set was brighter or darker than the mean brightness of the set. The results showed that the point of subjective equality (PSE) for mean brightness systematically deviated from the arithmetic mean. The PSE varied with the luminance level of the set and was found to be between 0.72 and 1.31 times the arithmetic mean of the set. Bauer (2009) interpreted this systematic deviation of the PSE as supporting the idea that ensemble mean brightness follows Stevens's power law, just as the brightness of single items does. However, in our opinion, if both the brightness of a single probe and the ensemble mean brightness follow the power law, the relation between the PSE (measured with the single probe) and the arithmetic mean luminance of the ensemble should have been described by a linear function in a log–log plot, as has been demonstrated in cross-modality matching (J. C. Stevens, Mack, & Stevens, 1960; S. S. Stevens, 1975). The slope of the function would correspond to the ratio of the two original exponents of the power function. If the exponents were similar, the slope would be close to unity. This prediction did not hold. Thus, the results suggested that some distinctive processing might underlie the perception of mean brightness.

In fact, differential weighting of some specific items in a stimulus set has been implicated in the context of lightness perception. Toscani, Valsecchi, and Gegenfurtner (2013a) found that, when asked to judge the lightness of objects, observers tended to direct their eyes to the brightest parts of the objects, and their lightness matches were biased toward the luminance of the fixated region. Toscani et al. further showed that this effect was mediated by attention as well as by eye fixations (see also de Fockert & Marchant, 2008, for the effects of attention on mean size). Other previous studies have argued that when judging the surface color or lightness of an object, the stimulus region of the highest contrast can be most informative (Anderson & Winawer, 2005, 2008; Wollschläger & Anderson, 2009). Moreover, previous studies on color averaging of multicolored textures also reported the bias favoring the region of the highest contrast (saturation) (Kimura, 2018; Kuriki, 2004; Sunaga & Yamashita, 2007). Recently, Kimura (2018) showed that the bias depended on stimulus variability around the mean. When the color variability in the texture was large, the mean color consistently deviated from the colorimetric mean toward the most-saturated color in the set. These findings accord with the idea that favoring the highest-luminance or highest-contrast items can be a good heuristic for the visual system to efficiently represent the surface properties of objects, particularly when sensory information is noisy.

An alternative type of differential weighting strategy—that is, trimmed or robust averaging—has also been demonstrated in ensemble coding (de Gardelle & Summerfield, 2011; Haberman & Whitney, 2010; Myczek & Simons, 2008). *Robust* averaging refers to down-weighting or excluding outlying items (those that fall far from the mean of the set, such as of the highest or lowest values) when making mean judgments. It can lead to reliable judgments by reducing the influences of less trustworthy evidence. De Gardelle and Summerfield (2011) showed that when asked to classify the mean hue (e.g., red vs. blue), observers discounted outlying items and based their decisions principally on the items close to the mean. Robust averaging may also be used in the computation of ensemble mean brightness, although the weighting of the extreme values seems to be in the opposite direction from the one discussed above. De Gardelle and Summerfield used trial-by-trial feedback, and thus perceptual learning might be an important factor for robust averaging of color (see also Fan, Turk-Browne, & Taylor, 2016). Taking these and previous findings together, investigating how heterogeneous luminance signals are summarized into an ensemble mean brightness could contribute to furthering our understanding of ensemble coding and brightness perception, as well as of the perception of illumination and surface color.

The objective of the present study was to investigate how accurately and precisely observers can judge the mean brightness of briefly presented heterogeneous achromatic arrays. Specifically, we asked the question of whether mean brightness judgments can be done in an efficient and parallel fashion or whether they involve some "smart" strategy, such as differential weighting. For this purpose, we introduced a novel ensemble brightness comparison paradigm in which two stimulus arrays were composed of different numbers of achromatic disks (e.g., 6 and 12 disks; see Fig. 1). The observer's task was to choose the brighter (or darker) array. This paradigm can allow us to indicate whether observers differentially weighted the highest (or lowest) luminances, rather than holistically averaging the brightness of the disks, as we describe below.

If two stimulus arrays are composed of the same number of disks, the brighter (or darker) array will tend to contain the highest- (or lowest-)luminance disk. Thus, strongly weighting the highest (or lowest) luminance in the computation of the ensemble mean brightness is a good strategy for comparing mean brightness. Moreover, even if observers cannot integrate luminance signals across the disks, simply finding the highest- (or lowest-)luminance disk in the entire stimulus display and choosing the array containing that disk can generally lead to the correct answer. (This can be considered an extreme case of differential weighting—that is, disk luminances other than the highest one are weighted at zero.) This shortcut is troublesome, because it is very difficult to tell whether the observer's choice relied on the simple shortcut or on actual averaging of the disk luminances, when the stimulus arrays are composed of the same number of disks. The present paradigm, with different numbers of disks, allowed us to reduce the covariance between the mean and the highest (or lowest) luminance. We could have pairs of arrays in which the darker array contained the highest-luminance disk, as will be fully explained in the later section and in Fig. 2. Thus, simply choosing the array containing the highest-luminance disk would result in a specific bias (Fig. 2), and this bias could be used to indicate whether the results were based on the shortcut or on holistic averaging. Similar control of the covariance can be achieved while keeping the number of disks constant, if we manipulate luminance variability around the mean. However, larger luminance steps might lead to texture segmentation, and the disks having different luminances might not be perceived as a group (Maule & Franklin, 2015; Utochkin & Tiurina, 2014).

Another possible strategy that observers could utilize for mean brightness judgments was choosing the array with the greater total brightness. When the stimulus arrays are composed of the same number of disks, the array of the greater total brightness always has the greater mean brightness. Therefore, it would be impossible to distinguish averaging from summing the luminance signals. However, the total-brightness strategy does not work well in the present paradigm with different numbers of disks. It would always lead to the wrong answer when the array with less disks had the greater mean brightness. Overall, the present paradigm can be effective to indicate the involvement of several strategies other than holistic rote averaging. Thus, the investigation of mean brightness judgments using the present paradigm can provide critical insights into whether and how an ensemble mean is computed for brightness.

Using the present paradigm, we showed that mean brightness judgments were precise, in the sense that the just-noticeable difference (JND) was small, but they were also mildly biased. On a dark background, the bias could be accounted for by observers heavily relying on either the highest-luminance or the highest-

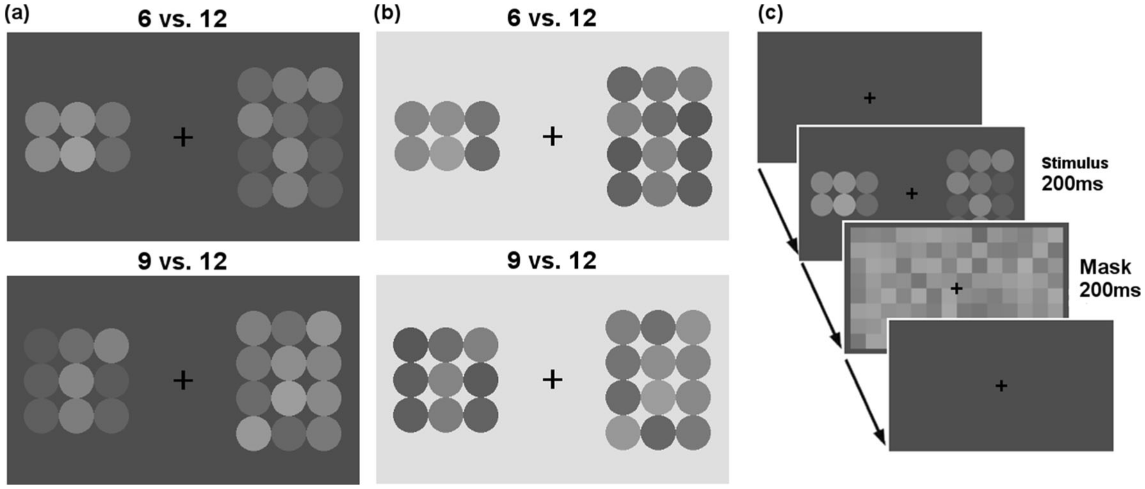

Fig. 1 Stimulus configurations (**a** and **b**) and the stimulus sequence (**c**) used in this study. A dark background was used in Experiment 1 (**a**), whereas a white background was used in Experiments 2 and 3 (**b**). The top row shows the 6-versus-12 disk configuration, and the bottom row the 9-versus-12 configuration. In the measurement, the left and right positions of the disk arrays were randomly switched from trial to trial. Note that, in these examples, the disk luminances on the dark (**a**) are the same as those on the white (**b**) background. Nonetheless, the apparent contrast of the disks greatly changes with the background luminance. The background of the stimulus examples was trimmed to reduce the size of the drawings, and the drawings are not to scale

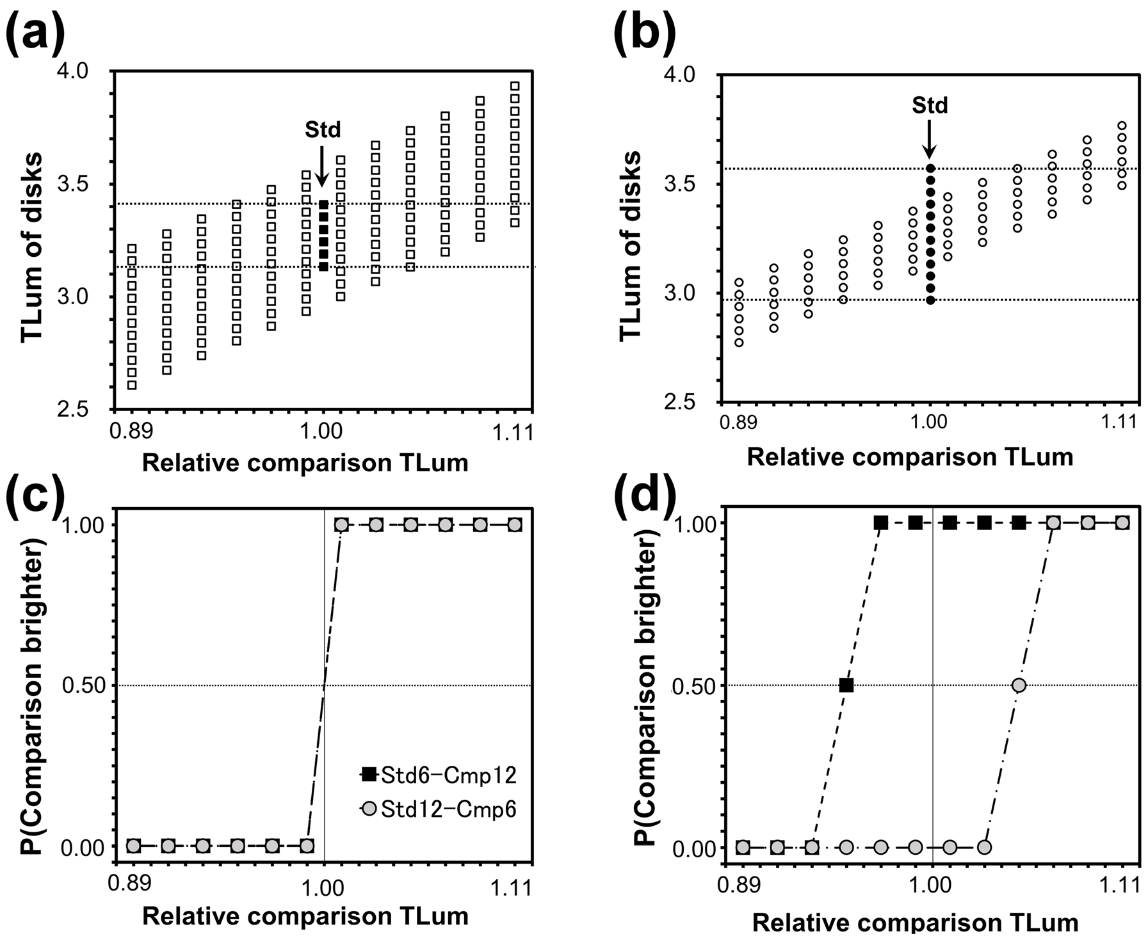

Fig. 2 (**a** and **b**) Transformed luminance (TLum) of the disks for each comparison stimulus in the Std6–Cmp12 (**a**) and Std12–Cmp6 (**b**) conditions. Here, TLum is defined by luminance transformed according to the power function with an exponent of 1/3—that is, TLum = luminance$^{1/3}$. The TLums of the disks in the standard stimulus are shown by solid symbols. Two horizontal dotted lines indicate the highest and lowest TLums of the standard stimulus. (**c** and **d**) Predictions of the results in Experiment 1: (**c**) Accurate brightness averaging. (**d**) Choosing the stimulus array containing the highest-luminance disk (highest-luminance shortcut). Proportions of responses indicating that the comparison stimulus was brighter are plotted as a function of the relative comparison TLum. In each panel, black squares represent the prediction in the Std6–Cmp12 condition, whereas gray circles show the prediction in the Std12–Cmp6 condition. In the Std6–Cmp12 condition, for the relative comparison TLum of 0.95, the highest luminances in the standard and comparison stimuli were very similar (Std = 3.409 vs. Cmp = 3.410; see panel **a**), and thus the proportion of responses was set to .50 (black squares in panel **d**). A similar situation occurred in the Std12–Cmp6 condition for the relative comparison TLum of 1.05 (Std = 3.574 vs. Cmp = 3.572)

contrast disk (Exp. 1). In Experiment 2 we used a white background, to differentiate the predictions based on the two strategies, and found that the bias resulted from favoring the highest-luminance disk. In Experiment 3 the task was changed from choosing the brighter to the darker array, and the results revealed that the bias was task-dependent. These findings implicate efficient and flexible weighted averaging of luminance signals for ensemble mean brightness judgments.

Experiment 1: Choosing the *brighter* array on a *dark* background

To investigate how accurately and precisely observers can judge the mean brightness of achromatic disks, we measured the PSE and the JND for mean brightness using two stimulus arrays with different numbers of disks. Observers' task was to choose the brighter array on a dark background. To evaluate the precision of the mean brightness judgments, the JND for mean brightness was compared with that measured using single disks. This comparison could indicate how precise mean brightness judgments are, relative to single-disk brightness judgments.

According to Stevens's law, brightness is a nonlinear function of luminance (but see also Nam & Chubb, 2000). Attempting to compensate for the nonlinearity and make the spacing of disk luminance perceptually linear, we transformed disk luminances according to Stevens's power law with an exponent of 1/3 (S. S. Stevens, 1975). The selection of perceptual metric would influence the interpretation of any possible bias. This issue will be discussed in General Discussion, after we present all the experimental results.

Method

Observers Eleven observers in total, who had normal or corrected-to-normal visual acuity, participated in Experiment 1. All the observers except for the first author (S_1) were naïve with regard to the purpose of the experiments. Six of them participated in the main experiment, seven of them took part in the random-mean experiment, and a different subset of seven performed in the single-disk experiment. The observers who participated in this and the following experiments gave informed consent after thorough explanation of the procedures before the experiment, and a specific number was assigned to each of them (e.g., S_1). The experiments were conducted in accordance with the Declaration of Helsinki and were approved by the university's Human Research Ethics Committee.

In this study, the number of observers was small. We attempted to analyze mean brightness judgments as they were manifested at the individual-observer level, as well as at the group-mean level. Because we collected a lot of data for each observer, as will be described in Procedure section, statistical power was sufficient within observers and did not much rely on averaging across observers.

Apparatus and stimuli The stimuli were generated using Matlab (The MathWorks Inc.) in conjunction with the Psychophysics Toolbox 3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). They were displayed on a 22-in. Mitsubishi color monitor (RDF223H), driven by an NVIDIA video card with a pixel resolution of 1,600 × 1,200 and a frame rate of 85 Hz. The intensity of each phosphor could be varied with 10-bit resolution. A Minolta CS-1000 spectroradiometer and an LS-100 luminance meter were used to measure the spectral radiance characteristics of the monitor's three phosphors and the gamma function for each phosphor. Then these calibration data were used to produce stimuli with desired tristimulus coordinates. The CIE *xy* chromaticity coordinates of all stimuli were set to those of CIE illuminant D65 ($x = 0.313$, $y = 0.329$), and only their luminance was varied. A chin and forehead rest was used to maintain a viewing distance of 57 cm.

Each stimulus display was composed of the standard and comparison stimuli, which were presented on the left and right sides of a fixation cross presented in the center of the screen (Fig. 1). The sides of the standard and comparison stimuli were randomly changed from trial to trial. Two different stimulus configurations were used for multiple disks: 6 versus 12 disks or 9 versus 12 disks (Fig. 1a). The 6-, 9-, and 12-disk arrays were composed of 3 × 2, 3 × 3, and 3 × 4 arrays of disks, respectively. The disks were immediately adjacent to each other in order to facilitate grouping of the heterogeneous disks. The stimulus condition will be specified in the form "Std6–Cmp12," meaning that the standard stimulus was a 6-disk array and the comparison stimulus was a 12-disk array. The four configuration conditions were Std6–Cmp12, Std12–Cmp6, Std9–Cmp12, and Std12–Cmp9. The stimulus conditions using the 6-versus-12 configuration (Std6–Cmp12 and Std12–Cmp6) might be sufficient to discourage observers from using other luminance clues (e.g., the highest or lowest luminance) rather than the mean luminance. Nonetheless, the 9-versus-12 configuration was included in the measurement mainly to discourage observers from focusing their attention on a small part of the stimulus. As is shown in Fig. 1a, the 6-disk array corresponds to the two central rows of the 12-disk array. Thus, limiting the configuration only to 6 versus 12 disks might have promoted observers' focusing their attention on the common two central rows. In addition, including the 9-versus-12 configuration increased the variety of the stimulus display, which could discourage observers from memorizing the stimulus. (See the Possible Effects of Heavily Weighting the Highest Disk Luminance section below for further discussion about the 9-versus-12 configuration.) Although a 6-versus-9 configuration was also possible, it was not used in the experiment.

Each of the disks in the stimulus array subtended 1.5°. The arrays were vertically centered on the screen, and the horizontal center of the arrays was shifted from the fixation point by 4.5°. The background subtended 35.3° (horizontally) by 26.9° (vertically) and had a luminance of 5 cd/m^2 (Fig. 1a). The standard and comparison stimuli were presented simultaneously for 200 ms (Fig. 1c). The stimuli were followed by a dynamic pattern mask for 200 ms, to disrupt visual processing of the stimulus after the offset and minimize possible effects of afterimages of the stimulus. The mask was a 14° × 8° random mosaic composed of 1° squares, and its mean luminance was set to 5 cd/m^2 (i.e., the background luminance). A blank screen was presented until the observer's response.

Stimulus luminance was transformed according to the power function, with an exponent of 1/3 (Bauer, 2009; S. S. Stevens, 1975). For the sake of simplicity, let us call the transformed luminance TLum—that is, TLum = luminance$^{1/3}$. The mean TLum of the standard stimulus was fixed at 3.271 (35 cd/m^2). The mean TLum of the comparison stimulus was defined relative to the mean standard TLum (relative comparison TLum, hereafter), and varied from 0.89 to 1.11 in steps of 0.02 (the mean luminance ranged from 25 to 48 cd/m^2). More precisely, let R_{CTL} be the relative comparison TLum, and $TLum_{MS}$ and $TLum_{MC}$ be the mean standard and comparison TLums, respectively. Then, $TLum_{MC}$ can be described as

$$TLum_{MC} = R_{CTL} \times TLum_{MS}.$$

For example, the lowest mean TLum of the comparison stimulus was 0.89 × 3.271 = 2.911, which is equal to 2.911^3 = 24.67 cd/m^2. The N TLums, TLum$_k$ (k = 1, 2, . . . , N), in a given comparison array with a mean TLum (TLum$_{\mathrm{MC}}$) were calculated as

$$\mathrm{TLum}_k = \mathrm{TLum}_{\mathrm{MC}} + \left(k - \frac{N+1}{2}\right) \times 0.055.$$

That is, the TLums of all disks in a given stimulus array were distinct (Fig. 2a and b) and equally spaced in 0.055 steps, centered on TLum$_{\mathrm{MC}}$. With this definition, the range of TLum was smaller for the array of the smaller number of disks. The spatial position of disks within the array was shuffled randomly between trials.

Because the mean standard TLum was constant (3.271) in the measurement, two additional luminance conditions were also included to discourage observers from memorizing the stimulus luminances and basing their judgments on the memory of the stimulus rather than sensory inputs. In one condition, the mean TLums of the two stimulus arrays were 3.500 and 3.631 (1.07 and 1.11 in relative units, respectively), and in the other condition, the mean TLums were 2.911 and 3.042 (0.89 and 0.93, respectively).

We also conducted an auxiliary random-mean experiment in which the mean standard TLum was varied randomly from trial to trial in a range from 3.042 to 3.500 (0.93 to 1.07 in relative units). The other aspects were kept the same as in the main experiment.

Procedure The method of constant stimuli was used in this and the following experiments. The experiment was run in a dark room. At the beginning of each daily session, observers dark-adapted for at least 5 min, and then preadapted to the background for 2 min. On each trial, the observer's key press initiated the stimulus sequence (Fig. 1c). Observers were asked to compare the mean brightness of each array of disks and to indicate which array, left or right, was brighter. All observers except for the first author did not know how the stimuli were constructed. Auditory feedback was given on every trial, indicating whether the observer's response was correct or wrong. The brighter array was defined as the one that had a higher mean TLum. Although TLum was used for comparison, we confirmed that calculation based on linear luminance (in candelas per square meter) would have provided the same answer. Observers were instructed that they could have as many practice trials as they wanted in order to familiarize themselves with the task. Each observer had ten to a few dozen practice trials before the main experiment.

A randomized block design was used, and all 56 stimulus conditions for multiple disks [4 configuration conditions (Std6–Cmp12, Std12–Cmp6, Std9–Cmp12, Std12–Cmp9) × 14 luminance conditions (12 comparison luminance steps and two additional luminance conditions)] were repeated 10 times in a daily session. For each observer, each session was repeated three times on different days to obtain reliable data. The total number of trials for each stimulus condition was 30. In the analysis of the results, the proportion of responses indicating that the comparison stimulus was brighter was plotted as a function of the relative comparison TLum in each configuration condition for each observer. The PSE (i.e., the relative comparison TLum producing 50% performance) and the JND (i.e., the difference between the relative comparison TLum producing 75% performance and that corresponding to the PSE) were determined by fitting a cumulative normal distribution function to the psychometric function using psignifit 4 (Schütt, Harmeling, Macke, & Wichmann, 2016). The results in the additional conditions using different mean TLums were excluded from the PSE and JND analysis but are described in the supplementary results.

To investigate the precision of mean brightness judgments, the JND for mean brightness was compared with that measured using single disks. In the measurement, both the standard and comparison stimuli were composed of a single disk. Each disk was centered vertically on the screen and was positioned horizontally 4.5° away from the fixation point, so that the center of the disk corresponded to the position of the multiple disks in the main experiment. The relative comparison TLum was varied from 1.01 to 1.11 (in 0.02 steps). Other aspects were the same as in the main experiments.

Possible effects of heavily weighting the highest disk luminance To better understand the results of the experiment, it would be helpful to preexamine possible effects of heavily weighting the highest disk luminance (highest-luminance strategy) in mean brightness judgments. As a basis for comparison, if the brightness of multiple disks were averaged accurately, regardless of the difference in the numbers of disks, the PSE would be expected to be 1.0 in relative TLums (i.e., the point of objective equality) in both the Std6–Cmp12 and Std12–Cmp6 conditions (Fig. 2c). Additionally, if the brightness of multiple disks were summarized with robust averaging, the PSE would also be expected to be 1.0, because, in the present stimulus, the distribution of TLum is symmetric to the mean, and thus, down-weighting the extreme values does not shift the PSE. Here, in Fig. 2c and d we assume that observers' responses are based on ideally precise luminance discrimination, but the actual response should be more stochastic.

In contrast to accurate averaging, heavily weighting the highest disk luminance in the computation of the mean would increase the mean brightness of the stimulus array containing the disk. Thus, that array would be chosen more as the brighter of the two, even when the mean TLums were physically the same between the standard and comparison stimuli. Although the present paradigm with different numbers of disks would discourage observers from using the highest-luminance strategy, through the trial-by-trial feedback, observers might still use it, particularly when the task was difficult. How biased the observers' choices would be would depend on how heavily the highest luminance was weighted. Here, for the sake of simplicity, we assume an extreme case of the highest-luminance strategy—that is, that observers would simply choose the array containing the highest-luminance disk as the brighter one (Fig. 2d). We call this strategy the highest-luminance *shortcut*. This shortcut predicts that the 12-disk array would be chosen as the brighter one even when it had a lower mean TLum, because it had the wider TLum range, and thus the highest-luminance disk would be included in this array (Fig. 2a and b). In the Std6–Cmp12 condition, this would happen when the mean comparison TLum was in the range of 0.95 to 0.99 (Fig. 2a). Choosing the 12-disk array (i.e., the comparison stimulus) as the brighter one would lead to a leftward shift of the psychometric function (black squares in Fig. 2d), and thus produce a biased lower PSE. A similar situation can be found when the mean comparison TLum was in the range of 1.01 to 1.05 in the Std12–Cmp6 condition (Fig. 2b). Choosing the 12-disk array (i.e., the standard stimulus) as the brighter one would produce a biased higher PSE (gray circles in Fig. 2d). An important prediction is that the highest-luminance shortcut, and the highest-luminance strategy in general, predicts that the biases would be in the opposite directions in the Std6–Cmp12 and Std12–Cmp6 conditions.

More specifically, if observers simply chose the array containing the highest-luminance disk (i.e., took the highest-luminance shortcut) and the observers' discrimination were ideally precise, we would expect to see a shift of PSE by ± 0.05 units of TLum, as is shown in Fig. 2d (about a – 14% or + 16% luminance difference, respectively, in candelas per square meter). The size of the shift would be smaller if observers relied and averaged more than one disk luminance and if internal and external noises were taken into consideration. Nonetheless, because the highest-luminance strategy would be expected to shift the PSE in opposite directions in the Std6–Cmp12 and Std12–Cmp6 conditions, a within-subjects analysis of the PSE difference between the two conditions could allow us to detect the small bias in mean brightness judgments.

The same pre-examination for the 9-versus-12 configuration indicated that the shift of PSE would be ± 0.02 units of TLum (about a – 5.9% or + 6.1% luminance difference, respectively, in candelas per square meter), and could be smaller in the actual measurement. Consequently, detecting the PSE shift in the 9-versus-12 configuration was expected to be very difficult. Nonetheless, the 9-versus-12 configuration was included in the measurement mainly to increase the variety of the stimulus displays, as is described in Apparatus and Stimuli section. Our main analysis focused on the results in the 6-versus-12 configuration, and the results in the 9-versus-12 configuration are described in the supplementary results.

Note that the present prediction based on the highest-luminance shortcut is the same as that based on the highest-*contrast* shortcut—that is, the strategy of simply choosing the array containing the highest-contrast disk. Because the background was dark in Experiment 1, the highest-luminance disk also had the highest contrast to the background. Differentiating the predictions based on highest contrast from those based on the highest-luminance shortcut would require using a brighter background (see Exp. 2).

For the sake of completeness, if observers could only utilize the total brightness of the arrays, instead of the mean brightness, their performance would be very poor, particularly when the relative comparison TLum was close to 1.0.

Results and discussion

The results of Experiment 1 (the 6-versus-12 disk configuration) revealed that observers could carry out mean brightness judgments fairly well, but their judgments exhibited consistent and moderate bias (Fig. 3). The size of the bias was different for different observers. The psychometric functions for some observers (e.g., S_4) exhibited a very small shift in the horizontal position (Fig. 3a), and their results might be considered accurate averaging. However, the functions for other observers (e.g., S_3) showed a mild bias (Fig. 3b). The results for PSE (Fig. 4a) revealed a small to medium bias for all observers. The shift of 0.02 in relative TLum corresponded to about a 6% difference in mean luminance expressed in candelas per square meter. The difference in PSE averaged across different observers was statistically significant between the Std6–Cmp12 and Std12–Cmp6 conditions [$t(5) = 3.77$, $p = .013$, Cohen's $d = 3.03$]. Although the size of the PSE shift was relatively small, the effect size (d) of the mean PSE difference was large. We also show the 95% confidence interval for each mean PSE (the rightmost data points in Fig. 4a). The intervals for both the Std6–Cmp12 and Std12–Cmp6 conditions did not include 1.0 (the point of objective equality). The biases were smaller than those predicted on the basis of the assumption that observers choose the array containing the highest-luminance disk (the second leftmost data points in Fig. 4a; see also Fig. 2d). These findings indicated that accurate averaging, robust averaging, or linear spatial pooling of local luminances cannot explain the results well. The results are more consistent with differential weighting of the highest luminance in averaging the disk luminances. Smaller biases are consistent with the accounts that observers relied on more than one disk luminance and/or that some noise contributed to the judgments. The results in the 9-versus-12 disk configuration exhibited smaller biases (see the supplementary materials for results from the 9-versus-12 disk configuration in this and subsequent experiments).

Similar results were found in the auxiliary experiment, in which the mean luminance of the standard stimulus was

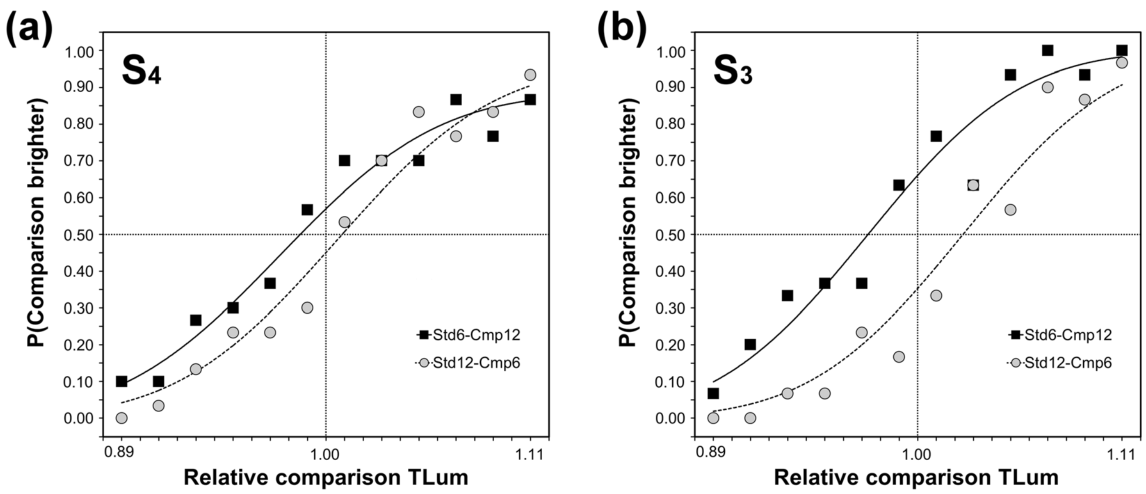

Fig. 3 Psychometric functions for representative observers in Experiment 1: (**a**) observer S_4 and (**b**) observer S_3. The proportion of comparison-was-brighter response was plotted as a function of the relative comparison TLum. In each panel, black squares represent the results in the Std6–Cmp12 condition, whereas gray circles show those in the Std12–Cmp6 condition. The solid and dashed lines illustrate cumulative normal distribution functions fitted to the data

varied randomly from trial to trial (Fig. 4b). The difference in PSEs was statistically significant between the Std6–Cmp12 and Std12–Cmp6 conditions [$t(6) = 3.61$, $p = .011$, $d = 2.53$]. Thus, the findings in the main experiment are not specific to situations in which the standard mean TLum was fixed on most of the trials. Moreover, the results were also similar in the additional luminance conditions of the main experiment, in which the mean TLums of the two arrays were different from the constant standard TLum (i.e., 3.271, which is 35 cd/m^2); the proportions of responses selecting the brighter array were similar to those found in the conditions with comparable TLum differences (see the supplementary material for the results from the additional luminance conditions in this and subsequent experiments).

Another important finding is that JNDs measured with multiple disks for mean brightness judgments appear to be smaller than JNDs measured with single disks (Fig. 5). The difference in JNDs between single disks and the main experiment was not statistically significant [$t(11) = 0.94$, $p = .37$, $d = 0.57$], but the difference between single disks and the random-mean experiment was statistically significant [$t(9) = 2.50$, $p = .034$, $d = 1.44$]. Smaller JNDs with multiple disks are consistent with a benefit of averaging—that is, averaging canceled out random noises in the individual signals and thus increased the precision of judgments. The size of the benefit might not have been very large when observers averaged (or highly weighted) only a few disk luminances.

Overall, the present results demonstrated that mean brightness judgments were efficient and precise. However, they were biased and not free from the effects of different array sizes. The bias was in the same direction as that predicted by the highest-luminance shortcut. However, as has already been discussed, similar biases can also be predicted by the highest-contrast shortcut. In Experiment 2, the predictions by the two shortcuts were differentiated by using a white background.

Experiment 2: Choosing the *brighter* array on a *white* background

In Experiment 2 we investigated the effects of changing the background luminance while keeping the disk luminance and the task the same as in Experiment 1. Making the background white greatly changed the luminance contrast of each disk with the background (Fig. 1b), so that the lowest-luminance disk became the highest-contrast disk. Therefore, if the biased judgments in Experiment 1 resulted from the observers relying more on the highest-contrast (or most salient) disk, the bias would be reversed in Experiment 2. However, if the highest-luminance disk by itself was important, the direction of the bias would be the same.

Method

In Experiment 2, the PSE measurement was conducted using a white background (75 cd/m^2) (Fig. 1b). The background had the highest luminance in the stimulus display throughout the measurement. The other aspects were kept the same as in Experiment 1. Eight observers, who had normal or corrected-to-normal visual acuity, participated in Experiment 2.

Results and discussion

The results for PSE (Fig. 6a) were similar to those in Experiment 1 (Fig. 4a), even though the apparent contrast of the disks was greatly changed against the white background

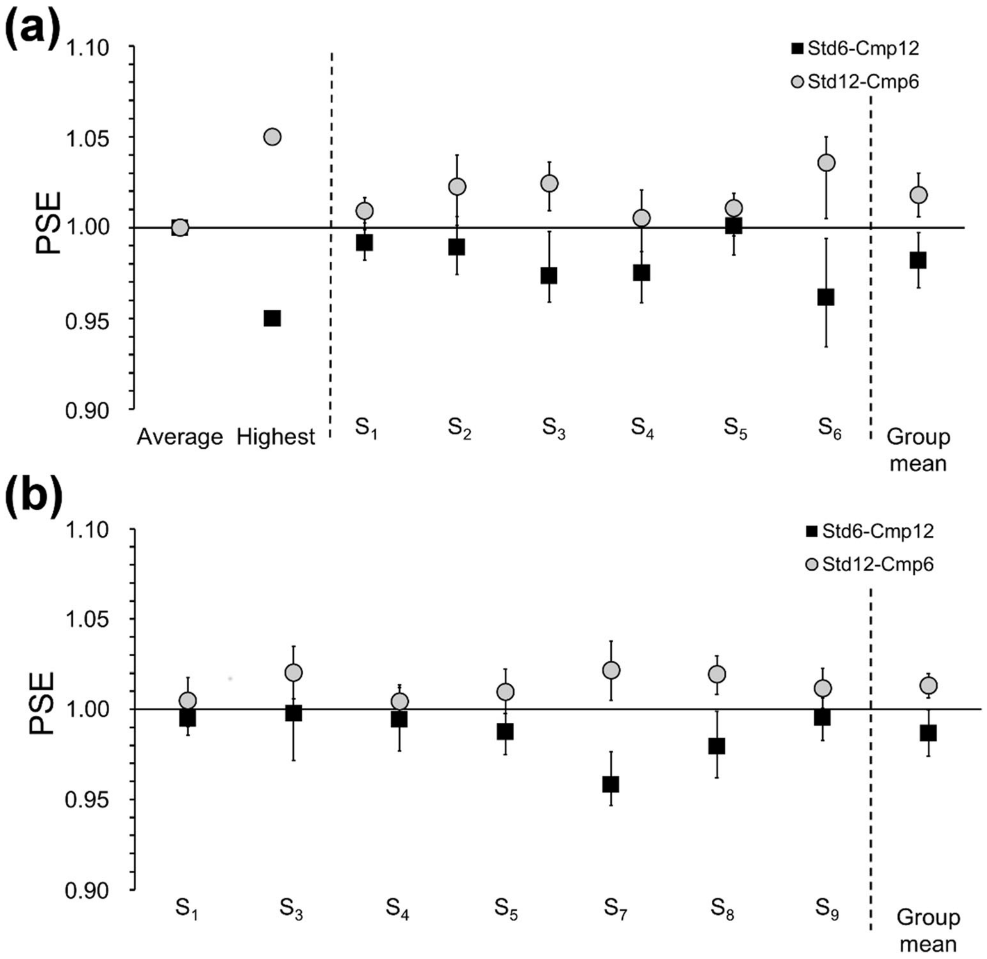

Fig. 4 Results of Experiment 1. (**a**) Predictions and results for the point of subjective equality (PSE). The data points shown on the left side are the predictions (see Fig. 2c and d). The symbols marked as "Average" represent the prediction for accurate brightness averaging; the PSEs are expected to be 1.0 (Fig. 2c). The symbols marked as "Highest" represent the prediction based on the highest-luminance shortcut. The PSEs are expected to be highly biased (Fig. 2d). The data points on the right side are the results for six observers. Error bars represent the 95% credible intervals (Schütt et al., 2016). The rightmost symbols show the averages across the different observers. Error bars designate the 95% confidence intervals. Note that this confidence interval can be used for inferring whether each mean PSE is different from 1.0 (the point of objective equality), but not for inferring the difference in mean PSEs between the Std6–Cmp12 and Std12–Cmp6 conditions. (**b**) Results for the PSE in the auxiliary random-mean experiment, in which the standard mean TLum was varied randomly from trial to trial

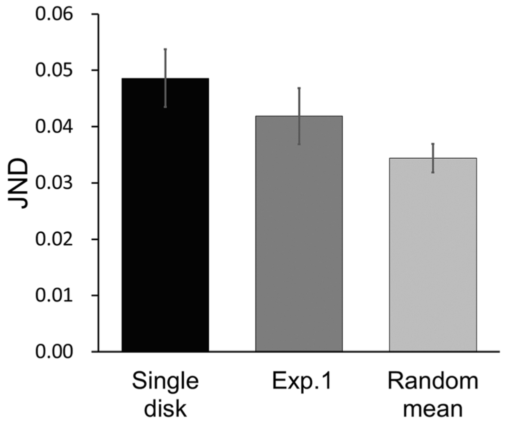

Fig. 5 Just-noticeable differences (JNDs) for mean brightness judgments measured with multiple disks in the main and random-mean experiments (dark- and light-gray bars, respectively), and JNDs measured with single disks (black bar). Error bars designate ± 1 *SEM* across observers

(cf. Fig. 1a and b). The PSE was mildly biased for most observers, and the direction of the bias was the same as in Experiment 1. The difference in PSEs was statistically significant between the Std6–Cmp12 and the Std12–Cmp6 conditions [$t(8) = 3.60$, $p = .009$, $d = 2.57$]. The effect size in Experiment 2 (2.57) was also very similar to that in Experiment 1 (2.51). This result could be explained if the observers who had participated in Experiment 1 adhered to the same strategy in Experiment 2, ignoring the change in the stimulus. However, this possibility is not likely, because at least one (S_{11}) of the three new observers (S_{10}–S_{12}) exhibited a similarly biased PSE. Thus, the bias does not seem to be associated with observers' persistent responses.

The result that the direction of the bias for mean brightness judgments remained the same with both dark and white backgrounds was consistent with the highest-luminance, but not

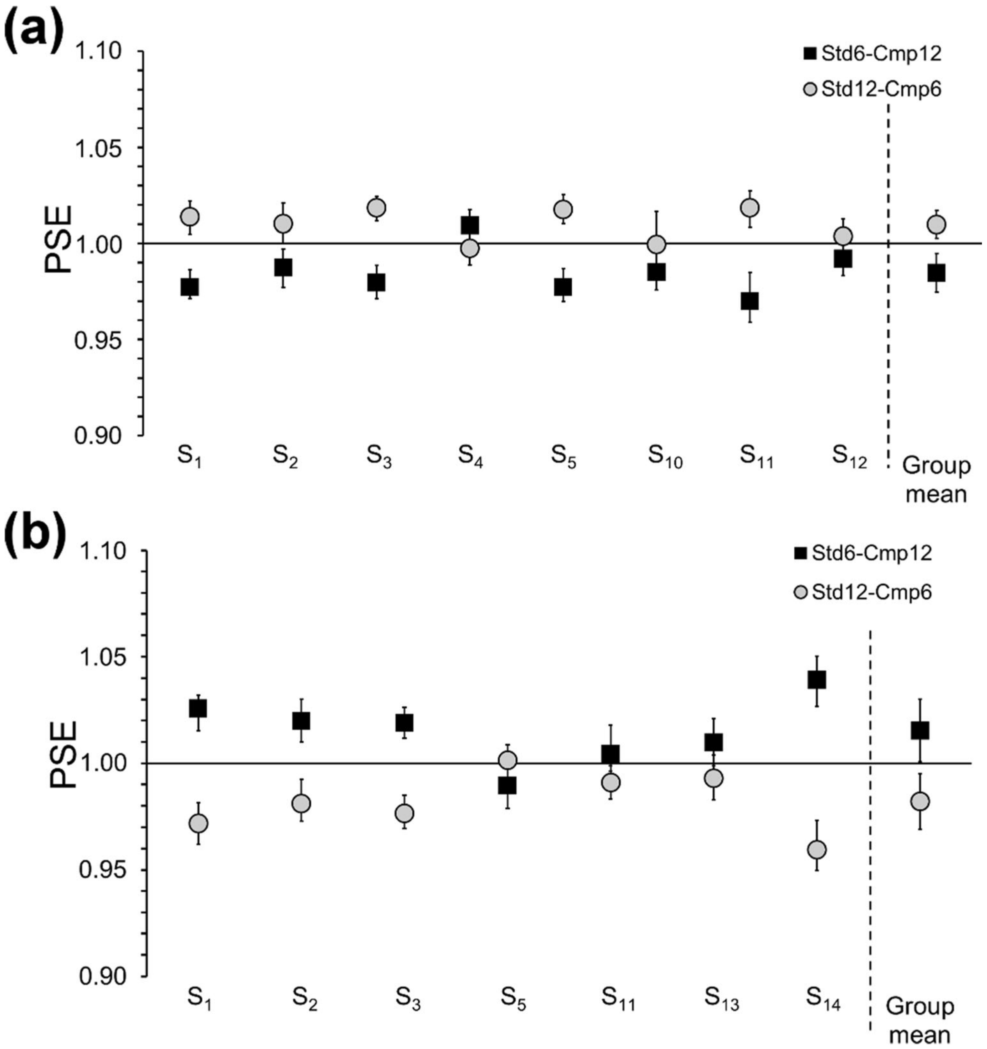

Fig. 6 Results for the point of subjective equality (PSE) in Experiments 2 (**a**) and 3 (**b**), respectively. Other aspects are the same as those in Fig. 4

with the highest-contrast, shortcut. This finding raises the question of why observers relied more on the highest-luminance disk even when it was not the most salient feature in the stimulus display. This might be because observers were asked to choose the *brighter* array of the stimulus. That is, observers might have relied on the information more relevant to the task—that is, on brighter rather than darker disks—particularly when the task was difficult. This account predicts that if we changed the task from choosing the *brighter* to choosing the *darker* array, the direction of the bias would be reversed. This prediction was tested in Experiment 3.

Experiment 3: Choosing the *darker* array on a *white* background

In Experiment 3, we investigated the effects of changing the task instructions, and observers were asked to choose the darker array. Because the stimuli remained unchanged, the accounts based on the stimulus properties predicted the same bias as in Experiment 2.

Method

In Experiment 3, only the instructions were changed from the settings in Experiment 2. Seven observers, who had normal or corrected-to-normal visual acuity, participated in Experiment 3.

Results and discussion

The results showed that, although the stimuli were completely the same as in Experiment 2, the opposite bias, favoring the lowest-luminance disk, was now found in most of the observers (Fig. 6b). The difference in PSEs was statistically significant between the Std6–Cmp12 and Std12–Cmp6 conditions [$t(7) = 2.94$, $p = .026$, $d = 2.38$]. The results in Experiment 3 strongly suggest that the bias was not stimulus-dependent, but task-dependent. The task instructions changed how the different disk luminances were weighted, and now the lowest disk luminance was weighted more strongly than the others.

Interestingly, the observers who had also participated in Experiment 2 reported that they did not intentionally change their strategy in response to the instruction change. Thus, the reversal of the bias might not be a conscious response.

General discussion

The present study demonstrated that observers can efficiently and flexibly select a set of disks that is particularly informative in view of the task requirement, and that they rely strongly on that subset of disks in making mean brightness judgments (Figs. 3, 4, and 6). These findings cannot be accounted for by assuming that the most salient disk attracted the observers' attention and that the allocation of attention biased the brightness judgments (see de Fockert & Marchant, 2008). The subset selection for weighted averaging seemed more deliberate. The mean judgments were accomplished efficiently even though observers had only 200 ms to view two heterogeneous arrays composed of 6 and 12 disks of different luminances. The judgments were also precise, in the sense that the JNDs measured with multiple disks were similar to, or even smaller than, the JND for simple brightness comparison with single disks (Fig. 5).

A bias similar to the one reported in this study could be observed even if the actual perceptual transformation of the luminance signals was greatly different from that assumed for setting up the stimulus luminances. Following Stevens's law (S. S. Stevens, 1975), we used a power function with an exponent of 1/3 to perceptually linearize luminance, but this might not have been effective enough to eliminate the bias. Additionally, the exponent of the function might not have been the same between the ensemble mean judgments with multiple disks and judgments with single disks (Bauer, 2009). Moreover, Nam and Chubb (2000) even reported that texture luminance judgments are not mediated by a compressive nonlinearity such as the ones both Stevens's law and Fechner's law assume for brightness perception. They showed that, when observers were asked to choose one of two texture patches that had greater total luminance, their judgments depended approximately linearly on the luminance of the texture elements. Thus, in retrospect, there is ambiguity concerning how luminance should be perceptually linearized. However, a possible incompleteness of perceptual linearization cannot account for the present findings. In Experiments 2 and 3, the stimuli were completely the same, but the direction of the bias was reversed, depending on the task. This is strong evidence for task-driven flexible brightness judgments.

Previously in the ensemble-coding literature, effects of task instructions have been reported. Im, Park, and Chong (2015) showed that mean size judgments were improved by matching task instructions. That is, when the task instructions favored larger mean sizes, mean size judgments were more accurate for the sets with larger mean sizes, whereas when the instructions favored smaller mean sizes, judgments were more accurate for the sets with smaller mean sizes. These findings indicated "ensemble-based" attention—that is, that ensembles can be flexibly selected by top-down (instruction-driven) attentional mechanisms (their other results also indicated the involvement of bottom-up attention). However, the present finding is distinct from the previous one, in that we showed that how observers computed mean brightness changed, depending on the task instructions.

Although we cannot rule out the possibility that observers only relied on the highest or lowest luminance of the disks [i.e., the highest- (or lowest-)luminance shortcut], several findings are more consistent with the notion that the observers actually integrated multiple luminance signals across different disks. One finding is that the observed biases were smaller in size than those predicted by the highest-luminance shortcut (Fig. 4a). If observers relied only on the highest-luminance or lowest-luminance disk, the bias would have been larger than the one we observed. However, the prediction was based on ideally precise luminance discrimination (Fig. 2d), and some judgment noise might need to be incorporated into simulations for more realistic predictions. Another finding is that the JND for mean brightness judgments can be smaller than the JND measured with single disks, which was obtained as a reference measure of the precision in judgments (Fig. 5). This finding presumably reflects the power of averaging (Alvarez, 2011), that averaging multiple noisy measures provides a more precise estimate of the mean than do the individual measures themselves, because random noise in one measure tends to cancel out the noise in another measure. The improvement in JND may not be very large (as is shown in Fig. 5), but that can be accounted for if observers average (or highly weight) a few disk luminances. Together, the biased but precise mean brightness judgments may be better understood in view of differentially weighted averaging.

It is remarkable that observers accomplished deliberate selection of a task-relevant subset of disks within a short duration of 200 ms, given that attentional dwell time is estimated to be in the range of 200 to 500 ms (Duncan, Ward, & Shapiro, 1994; Wolfe, 2003). Moreover, this selection was done even when different spatial configurations (i.e., 6-versus-12 and 9-versus-12 configurations; Fig. 1a and b) were randomly presented in the measurement. In this situation, exhaustive visual search for the task-relevant items over the stimulus display might not be possible, and thus observers might have primarily focused on the disks near fixation. Previous studies also showed that luminance signals at or near fixation significantly affect both lightness (Toscani et al., 2013a; Toscani, Valsecchi, & Gegenfurtner, 2013b) and brightness (Toscani, Gegenfurtner, & Valsecchi, 2017) judgments. In the ensemble-coding literature, a previous study showed that observers did not rely on the size of items presented near fixation to compute the average size (Chong, Joo, Emmmanouil, & Treisman, 2008), whereas another study on size variability discrimination showed that observers mostly sampled items close to fixation (Lau & Brady, 2018). Future studies will be needed to explore how observers select the task-relevant items during a short presentation time when making their mean brightness judgments.

The present study revealed evidence for strongly weighting extreme samples rather than inlying ones (i.e., those falling near the mean) in stimulus arrays. Thus, the results are contradictory to the robust or trimmed averaging that has been demonstrated in several types of ensemble mean judgments (de Gardelle & Summerfield, 2011; Fan et al., 2016). As the trial-by-trial feedback was given in the present experiments, the observers' biased responses were corrected during the measurement. Nonetheless, observers kept relying more on the extreme samples. One possible interpretation of this result is that learning robust averaging is considerably difficult for mean brightness judgments. But another, more intriguing interpretation is that the observers' weighting strategy itself was also task-dependent. When observers are asked to *compare* the means of two stimuli, as in the present study, strongly weighting extreme samples might be preferential, because they can be more informative about the *difference* in means. In contrast, when observers are asked to *classify* the means into different categories, as in de Gardelle and Summerfield (2011), strongly weighting inlying samples (robust averaging) might be preferential, because these samples can be more informative about the *location* of the mean.[1] Thus, observers may have sensibly selected the optimal strategy depending on task requirements. This interpretation may also help understand other apparently contradictory findings in previous studies on lightness and color. Some studies have shown that observers chose the strategy of heavily relying on extreme samples, such as the most saturated or highest-contrast items, when they needed to integrate variegated signals (Anderson & Winawer, 2005, 2008; Kimura, 2018; Kuriki, 2004; Sunaga & Yamashita, 2007; Wollschläger & Anderson, 2009). In contrast, Milojevic, Ennis, Toscani, and Gegenfurtner (2018) showed that, when asked to classify the color of natural stimuli (leaves), observers' judgments were predicted better by the mean chromaticity of the stimuli than by the most saturated color. These differences may be reconciled by considering flexible selection of the strategy or flexible weighting of some particular samples, depending on the nature of the task. In addition, previous studies have reported that stimulus variability around the mean can be a key factor in the effects of extreme samples (Kimura, 2018; Kuriki, 2004).

The present finding that observers displayed biased judgments even with trial-by-trial feedback made automatic extraction of the mean for ensemble brightness less likely. In fact, we used the feedback in the measurement, because we found in preliminary experiments that otherwise some observers exhibited a strong bias to choose a specific stimulus array (e.g., either the 6- or the 12-disk array) when mean brightness judgments were difficult. Previous studies also reported that several observers responded in a nonsystematic fashion in the mean brightness judgment task (Bauer, 2009). The present finding indicated that although spatial averaging of local luminances has been assumed in early visual processing (e.g., Bressan, 2006; Buchsbaum, 1980; Chubb et al., 1994; Chubb et al., 2004; Shapley & Enroth-Cugell, 1984), an explicit representation of mean brightness may not be immediately available to observers, even if it exists (see also Webster, Kay, & Webster, 2014, and Maule & Franklin, 2016, for similar discussions for the mean hue of multiple color patches).

Although the present results themselves did not identify what kind of visual mechanisms may underlie flexible mean brightness processing, some relevant findings have been reported. Investigating the discrimination of achromatic textures composed of small square elements, Silva and Chubb (2014) reported results supporting the existence of four channels differentially sensitive to grayscale textures. Two of the four channels, which are complementary to each other, are most relevant to this study: "up-ramped" and "down-ramped" channels. The sensitivity of the up-ramped (or down-ramped) channel increases (or decreases) linearly with increasing luminance and reaches its maximum (or minimum) near the high end. Silva and Chubb assumed that the outputs from the four channels are linearly combined in order to produce grayscale filters and that the weights of the channels are optimized for the task at hand. Their model describes the results for texture discrimination quite well. Although the spatial composition of the stimuli is different between the present and Silva and Chubb's studies, flexible mean brightness processing may be accounted for if we assume that observers selectively recruited the up-ramped and down-ramped channels in a task-dependent fashion.

Overall, the present findings suggest that the mean brightness of heterogeneous achromatic disks can be processed efficiently and precisely. This processing is presumably neither automatic nor based on rote averaging. Rather, it can be characterized as being flexible and task-dependent; some items that are important for the task at hand are processed preferentially. This kind of flexibility can be general and might be

[1] One of the reviewers suggested that another relevant difference, rather than difference versus location, can be the nature of the perceptual dimension, in that the dimension—that is, hue—tested by de Gardelle and Summerfield (2011) is a *qualitative* dimension, whereas brightness, which we tested in this study, is a *quantitative* dimension.

observed in ensemble processing of other perceptual attributes.

Acknowledgments This work was partly supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science to E.K. (Grant Nos. 26285162 and 25285197). We are grateful to Sang Chul Chong, Charlie Chubb, and an anonymous reviewer for their helpful comments and suggestions on earlier versions of the manuscript.

Open Practices Statement The data and materials for all experiments are available upon request from the authors.

References

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research*, *83*, 25–39. https://doi.org/10.1016/j.visres.2013.02.018

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*, 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Anderson, B. L., & Winawer, J. (2005). Image segmentation and lightness perception. *Nature*, *434*, 79–83.

Anderson, B. L., & Winawer, J. (2008). Layered image representations and the computation of surface lightness. *Journal of Vision, 8*(7), 18: 1–22. http://journalofvision.org/8/7/18/, https://doi.org/10.1167/8.7.18

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*, 157–162. https://doi.org/10.1111/1467-9280.00327

Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & Psychophysics*, *70*, 1325–1326. https://doi.org/10.3758/PP.70.7.1325

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *Psychological Record*, *59*, 171–186.

Bauer, B. (2015). A selective summary of visual averaging research and issues up to 2000. *Journal of Vision, 15*(4), 14:1–15, https://doi.org/10.1167/15.4.14

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

Bressan, P. (2006). The place of white in a world of grays: A double-anchoring theory of lightness perception. *Psychological Review*, *113*, 526–553. https://doi.org/10.1037/0033-295X.113.3.526

Buchsbaum, G. (1980). A spatial processor model for object colour perception. *Journal of the Franklin Institute*, *310*, 1–26. https://doi.org/10.1016/0016-0032(80)90058-7

Chong, S., Joo, S., Emmmanouil, T.-A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics*, *70*, 1327–1334. https://doi.org/10.3758/PP.70.7.1327

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*, 393–404. https://doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, *45*, 891–900. https://doi.org/10.1016/j.visres.2004.10.004

Chubb, C., Econopouly, J., & Landy, M. S. (1994). Histogram contrast analysis and the visual segregation of IID textures. *Journal of the Optical Society of America A*, *11*, 2350–2374. https://doi.org/10.1364/JOSAA.11.002350

Chubb, C., Landy, M. S., & Econopouly, J. (2004). A visual mechanism tuned to black. *Vision Research*, *44*, 3223–3232. https://doi.org/10.1016/j.visres.2004.07.019

Corbett, J. E., Oriet, C., & Rensink, R. A. (2006). The rapid extraction of numeric meaning. *Vision Research*, *46*, 1559–1573. https://doi.org/10.1016/j.visres.2005.11.015

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*, 3181–3192. https://doi.org/10.1016/S0042-6989(97)00133-8

de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, *70*, 789–794. https://doi.org/10.3758/PP.70.5.789

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences*, *108*, 13341–13346. https://doi.org/10.1073/pnas.1104517108

Duncan, J., Ward, R., & Shapiro, K. (1994). Direct measurement of attentional dwell time in human vision. *Nature*, *369*, 313–315. https://doi.org/10.1038/369313a0

Fan, J. E., Turk-Browne, N. B., & Taylor, J. A. (2016). Error-driven learning in statistical summary perception. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 266–280. https://doi.org/10.1037/xhp0000132

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*, R751–R753. https://doi.org/10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 718–734. https://doi.org/10.1037/a0013899

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*, 1825–1838. https://doi.org/10.3758/APP.72.7.1825

Im, H. Y., Park, W. J., & Chong, S. C. (2015). Ensemble statistics as units of selection. *Journal of Cognitive Psychology*, *27*, 114–127. https://doi.org/10.1080/20445911.2014.985301

Kanaya, S., Hayashi, M. J., & Whitney, D. (2018). Exaggerated groups: Amplification in ensemble coding of temporal and spatial features. *Proceedings of the Royal Society B*, *285*, 20172770:1–9, https://doi.org/10.1098/rspb.2017.2770

Kimura, E. (2018). Averaging colors of multicolor mosaics. *Journal of the Optical Society of America A*, *35*, B43–B54. https://doi.org/10.1364/JOSAA.35.000B43

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*(ECVP Abstract Suppl.), 14.

Kuriki, I. (2004). Testing the possibility of average-color perception from multi-colored patterns. *Optical Review*, *11*, 249–257. https://doi.org/10.1007/s10043-004-0249-2

Lau, J. S.-H., & Brady, T. F. (2018). Ensemble statistics accessed through proxies: Range heuristic and dependence on low-level properties in variability discrimination. *Journal of Vision*, *18*(9), 3:1–18, https://doi.org/10.1167/18.9.3

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica*, *142*, 245–250. https://doi.org/10.1016/j.actpsy.2012.11.002

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision*, *15*(4), 6:1–18. https://doi.org/10.1167/15.4.6

Maule, J., & Franklin, A. (2016). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. *Journal of the Optical Society of America A*, *33*, A22–A29. https://doi.org/10.1364/JOSAA.33.000A22

Milojevic, Z., Ennis, R., Toscani, M., & Gegenfurtner, K. R. (2018). Categorizing natural color distributions. *Vision Research*, *151*, 18–30. https://doi.org/10.1016/j.visres.2018.01.008

Myczek, K., & Simons, D. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, *70*, 772–788. https://doi.org/10.3758/pp.70.5.772

Nam, J.-H., & Chubb, C. (2000). Texture luminance judgments are approximately veridical. *Vision Research*, *40*, 1695–1709. https://doi.org/10.1016/S0042-6989(00)00006-7

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*, 739–744. https://doi.org/10.1038/89532

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442. https://doi.org/10.1163/156856897X00366

Sakuma, N., Kimura, E., & Goryo, K. (2017). Rapid proportion comparison with spatial arrays of frequently used meaningful visual symbols. *Quarterly Journal of Experimental Psychology*, *70*, 2371–2385. https://doi.org/10.1080/17470218.2016.1239747

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123. https://doi.org/10.1016/j.visres.2016.02.002

Shapley, R., & Enroth-Cugell, C. (1984). Visual adaptation and retinal gain controls. In N. Osborne & G. Chader (Eds.), *Progress in retinal research* (Vol. 3, pp. 263–346). Oxford, UK: Pergamon Press.

Silva, A. E., & Chubb, C. (2014). The 3-dimensional, 4-channel model of human visual sensitivity to grayscale scrambles. *Vision Research*, *101*, 94–107. https://doi.org/10.1016/j.visres.2014.06.001

Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics*, *70*, 1335–1336. https://doi.org/10.3758/PP.70.7.1335

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision, 11*(12), 13:1–11, https://www.ncbi.nlm.nih.gov/pubmed/22011381 , https://doi.org/10.1167/11.12.13

Stevens, J. C., Mack, J. D., & Stevens, S. S. (1960). Growth of sensation on seven continua as measured by force of handgrip. *Journal of Experimental Psychology*, *59*, 60–67. https://doi.org/10.1037/h0040746

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

Sunaga, S., & Yamashita, Y. (2007). Global color impressions of multicolored textured patterns with equal unique hue elements. *Color Research and Application*, *32*, 267–277. https://doi.org/10.1002/col.20330

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 329–337. https://doi.org/10.1037/a0028712

Toscani, M., Gegenfurtner, K. R., & Valsecchi, M. (2017). Foveal to peripheral extrapolation of brightness within objects. *Journal of Vision*, *17*(9), 14:1–14. https://doi.org/10.1167/17.9.14

Toscani, M., Valsecchi, M., & Gegenfurtner, K. R. (2013a). Optimal sampling of visual information for lightness judgments. *Proceedings of the National Academy of Sciences*, *110*, 11163–11168. https://doi.org/10.1073/pnas.1216954110

Toscani, M., Valsecchi, M., & Gegenfurtner, K. R. (2013b). Selection of visual information for lightness judgements by eye movements. *Philosophical Transactions of the Royal Society B*, *368*, 20130056. https://doi.org/10.1098/rstb.2013.0056

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*, 15–48. https://doi.org/10.1037/0033-295X.95.1.15

Utochkin, I. S., & Tiurina, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta Psychologica*, *146*, 7–18. https://doi.org/10.1016/j.actpsy.2013.11.012

Van Opstal, F., de Lange, F. P., & Dehaene, S. (2011). Rapid parallel semantic processing of numbers without awareness. *Cognition*, *120*, 136–147. https://doi.org/10.1016/j.cognition.2011.03.005

Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research*, *32*, 931–941. https://doi.org/10.1016/0042-6989(92)90036-I

Watamaniuk, S. N. J., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: The integration of direction information. *Vision Research*, *29*, 47–59. https://doi.org/10.1016/0042-6989(89)90173-9

Webster, J., Kay, P., & Webster, M. A. (2014). Perceiving the average hue of color arrays. *Journal of the Optical Society of America A*, *31*, A283–A292. https://doi.org/10.1364/JOSAA.31.00A283

Whitney, D., Haberman, J., & Sweeny, T. D. (2014). From textures to crowds: Multiple levels of summary statistical perception. In J. S. Werner & L. M. Chalupa (Eds.), *The new visual neurosciences* (pp. 695–709). Cambridge, MA: MIT Press.

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Sciences*, *7*, 70–76. https://doi.org/10.1016/S1364-6613(02)00024-4

Wollschläger, D., & Anderson, B. L. (2009). The role of layered scene representations in color appearance. *Current Biology*, *19*, 430–435. https://doi.org/10.1016/j.cub.2009.01.053

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.