Effects of ensemble complexity and perceptual similarity on rapid averaging of hue

John Maule

The Sussex Colour Group, School of Psychology, University of Sussex, Brighton, UK

Anna Franklin

The Sussex Colour Group, School of Psychology, University of Sussex, Brighton, UK

The ability to extract the mean of features from a rapidly viewed, heterogeneous array of objects has been demonstrated for a number of different visual properties. Few studies have previously investigated the rapid averaging of color; those that did had insufficient stimulus control or inappropriate methods. This study reports three experiments that directly test observers' ability to extract the mean hue from a rapidly presented, multielement color ensemble. In Experiment 1, ensembles varied in number of elements and number of colors. It was found that averaging was harder for ensembles with more colors but that changing the number of elements had no effect on accuracy, supportive of a distributed-attention account of rapid color averaging. Experiment 2a manipulated the hue range present in any single ensemble (varying the perceptual difference between ensemble elements) while still varying the number of colors. Range had a strong effect on ability to pick the mean hue. Experiment 2b found no effect of color categories on the accuracy or speed of mean selection. The results indicate that perceptual difference of elements is the dominant factor affecting ability to average rapidly seen color ensembles. Findings are discussed both in the context of perception and memory of multiple colors and ensemble perception generally.

Introduction

It has been claimed that humans have the ability to extract summary statistics from a briefly viewed visual scene for a number of visual properties, including orientation (e.g., Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), motion speed (e.g., Watamaniuk & Duchon, 1992), motion direction (e.g., Watamaniuk, Sekuler, & Williams, 1989), brightness (e.g., Bauer, 2009b), size (e.g., Ariely, 2001), emotional expression of faces (e.g., Haberman & Whitney, 2009), facial identity (e.g., De Fockert & Wolfenstein, 2009; Leib et al., 2014), and direction of biological motion (e.g., Sweeny, Haroz, & Whitney, 2012). Much of the literature is focused on perceptual averaging—whether ensembles are encoded by their mean properties.

To date, relatively little published research has investigated ensemble perception of color (Demeyere, Rzeskiewicz, Humphreys, & Humphreys, 2008; Kuriki, 2004; Maule, Witzel, & Franklin, 2014; Webster, Kay, & Webster, 2014). This is despite color being a good candidate for investigation from the point of view of better understanding how color is perceived and encoded and also for understanding the perceptual averaging mechanism. For color scientists, perceptual averaging experiments provide a paradigm that could help elucidate questions about the shape and organization of perceptual color space (Webster et al., 2014). For those interested in ensemble coding mechanisms and functions more generally, color is an ideal substrate for investigation. It is well described and characterized in terms of human perception, is continuous yet also subject to categorization (e.g., Bird, Berens, Horner, & Franklin, 2014; He, Witzel, Forder, Clifford, & Franklin, 2014; Roberson, Pak, & Hanley, 2008), and can help answer ecologically valid questions about the appearance of surface colors (Giesel & Gegenfurtner, 2010; Sunaga & Yamashita, 2007). Furthermore, hue (the subjective experience of which is qualitative), along with the other dimensions of color, saturation, and lightness (which are matters of magnitude), provide an opportunity to understand how ensemble coding deals with the integration of multiple perceptual dimensions.

Ensemble coding of color

Various papers have addressed the question of color averaging, and although in most cases their findings

Citation: 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, doi:10.1167/15.4.6.

doi: 10.1167/15.4.6 Received December 22, 2014; published June 25, 2015 ISSN 1534-7362

hint at a meaningful color averaging mechanism, most have not employed the methods used commonly in investigations of ensemble perception or investigated averaging under rapid viewing conditions. It has been shown that when colorful ensembles are presented for a short time (500 ms), observers tend to find the unseen mean hue as familiar as the hues that were present in the original ensemble (Maule et al., 2014), a pattern reminiscent of Ariely's (2001) finding for ensembles of different sizes. This same pattern appears for ensembles of color in a patient with simultagnosia, a condition rendering the patient unable to reliably count more than one or two items (Demeyere et al., 2008), suggesting that the color averaging, as with face averaging in cases of prosopagnosia (Leib et al., 2012), can survive in spite of cognitive deficits limiting the encoding of individual items.

In addition to the indications of mean encoding, there is also evidence that other summary statistics might also play a role in the extraction of gist from a rapidly viewed colorful ensemble. For example, high-variance colorful ensembles tend to elicit slower reaction times (RTs) and less accurate responses when judging whether a mean of another ensemble is "blue" or "red" (de Gardelle & Summerfield, 2011). Variance is further implicated as an important determinant of the accuracy of summary statistics by evidence that priming with a colorful ensemble can speed a judgment about another ensemble of the same variance even when the mean changes from prime to target (Michael, de Gardelle, & Summerfield, 2014). Likewise, the tendency to find the mean hue familiar in a membership identification task disappears when the perceptual difference between ensemble hues is increased (Maule et al., 2014), suggesting that there may be a functional limit to the amount of variance that can be rapidly encoded by summary statistics. These findings support suggestions that summary statistics serve to help tune the visual system to the environment and support the stability of the visual world (Corbett & Melcher, 2014a, 2014b; Corbett, Wurnitsch, Schwartz, & Whitney, 2012; Lanzoni, Melcher, Miceli, & Corbett, 2014).

Although promising, results from membership identification tasks (Maule et al., 2014) provide only an indirect measure of the encoding of the mean following brief ensemble presentations. A more direct test of color averaging is required to establish whether observers demonstrate ensemble coding for color by the mean when presented with an ensemble very briefly. If a mean hue can be accurately extracted from a rapidly presented ensemble, this will add weight to claims that ensemble coding is a pervasive feature of the visual system, possibly driven by a mechanism common to many different visual domains (e.g., Alvarez, 2011; Haberman & Whitney, 2012). Nevertheless, the evidence of priming by the mean and variance of colorful ensembles (Michael et al., 2014) is a strong indication that summary statistics of color may be encoded by the visual system.

Previous research has also found that observers can approximate the colorimetric mean when adjusting a homogenous patch to represent a continuously presented multicolor mosaic but that these estimates are biased toward the salience or saturation of the mosaic elements (Kuriki, 2004) and the position of the unique hues (Sunaga & Yamashita, 2007)—i.e., the red, green, blue, or yellow that appears pure, unmixed with any other hue (see Kuehni, 2014). Such adjustments are also more variable when the perceptual distance between element colors is greater (Webster et al., 2014). These studies allowed observers to view ensembles for an unlimited amount of time while making their settings. Therefore, it remains unknown whether these averaging mechanisms are the same when observers have limited time to view the stimuli—conditions in which ensemble coding would be most beneficial (Alvarez, 2011).

In summary, although some research on mean color perception has been carried out, there is no equivalent of the studies investigating ensemble perception of other features. Studies that have probed directly for representations of the mean color (e.g., Kuriki, 2004; Sunaga & Yamashita, 2007; Webster et al., 2014) have allowed observers unlimited time to view the ensemble stimuli rather than using rapid presentation to encourage the deployment of distributed attention (e.g., Alvarez & Oliva, 2008; Baijal, Nakatani, van Leeuwen, & Srinivasan, 2013; Treisman, 2006). When rapid presentation has been used (e.g., de Gardelle & Summerfield, 2011; Maule et al., 2014; Michael et al., 2014), tasks have not probed directly for representations of the mean color. Some studies have also failed to adequately control, using color spaces, the perceptual differences between the colors used in the experiments (e.g., de Gardelle & Summerfield, 2011; Demeyere et al., 2008; Michael et al., 2014). This study seeks to address this gap in the literature as well as further explore the nature of rapid mean color encoding (if it exists).

Theoretical questions

This study also addresses four important theoretical questions relevant to both ensemble perception generally and color cognition. First, any claim of robust, rapid color ensemble coding also requires attention to be paid to the mechanism and limits of this process. As highlighted in the literature on size averaging (e.g., Ariely, 2001, 2008; Marchant, Simons & De Fockert, 2013; Simons & Myczek, 2008), a crucial distinction regarding the mechanism of ensemble perception is

between holistic processing with distributed attention (exhaustive ensemble processing) and subsampling of relatively few elements (limited capacity ensemble processing). Exhaustive ensemble processing implies the parallel integration of all the objects or items in an ensemble to provide a summary representation, often with inaccurate or no representation of the individual items. This has been a controversial suggestion as it postulates a mechanism that exceeds the limit of visual short-term memory (Alvarez, 2011) with some studies appearing to demonstrate exhaustive processing (e.g., Ariely, 2001, 2008; Chong & Treisman, 2003, 2005a, 2005b) while others propose models that use subsampling can account for the accuracy of mean judgments without the need for a holistic mechanism (e.g., Marchant et al., 2013; Myczek & Simons, 2008; Simons & Myczek, 2008; Whiting & Oriet, 2011). Varying the number of elements in ensembles will provide an indication as to whether averaging is affected by an increase in the number of objects to average. If the process requires focused attention to a subsample of elements, then the averaging process should be subject to more error when there are more elements. In contrast, if the process occurs across the whole ensemble, the rate of error will be unaffected.

Second, given the evidence for the importance of variance and/or range of colors in extracting a mean (de Gardelle & Summerfield, 2011; Maule et al., 2014; Michael et al., 2014; Webster et al., 2014), it is pertinent to investigate the limit of averaging for ensembles varying more or less widely in color. Some functional limit for perceptual averaging of color is likely to be present (larger variances in size of ensemble elements is also detrimental to averaging performance; Utochkin & Tiurina, 2014), but whether this is a limit dependent on the number of different colors present in an ensemble or the perceptual difference between the elements of the ensemble is an open question.

Third, the perception and cognition of hue has features that are not shared by other visual domains investigated in the ensemble perception literature. Variation in hue tends to be represented (in color spaces) as forming a circular perceptual continuum. It is possible that the circularity of hue perception will interfere with the encoding of the mean color as any pair of hues has two angular differences (clockwise and counterclockwise) that could describe their perceptual relationship (these two angular differences adding up to 360°). As these angular differences approach parity (180°), these competing interpretations of the mean color become equally likely and may make the extraction of the mean color more difficult. For this reason, it might be expected that hue may be unsuited to ensemble coding, in which case averaging color would be very difficult or effortful.

Fourth, color is also subject to verbal labels (e.g., "green," "yellow"), which (at least in English) divide the hue circle into more-or-less discrete categories. The position of linguistic boundaries has been claimed to have effects on color memory (e.g., Roberson & Davidoff, 2000), visual search (Daoutis, Pilling, & Davies, 2006), neural representation (e.g., Clifford et al., 2012; Clifford, Holmes, Davies, & Franklin, 2010), and color discrimination (Drivonikou, Clifford, Franklin, Ozgen, & Davies, 2011; but see Witzel & Gegenfurtner, 2013). Notably, however, the effects of categories appear to be postperceptual in origin (Bird et al., 2014; He et al., 2014; Roberson et al., 2008). Given these effects, we might predict that ensembles containing multiple categories would prove harder to average. This would have implications for ensemble perception in other domains with categorical labels, such as facial expression/identity. If the categorical content of an ensemble does not make any difference, this will provide some support for the early and automatic computation of mean color from rapidly presented ensembles.

The present study

The present study attempts to address these questions, building on and rectifying the methodological constraints identified in the past literature. Using a two-alternative forced-choice (2AFC) task, observers were tested on the correct identification of the mean hue following the rapid presentation of a multihue ensemble. Following Maule et al. (2014), the stimuli are controlled to ensure equality of difference between adjacent colors in terms of just-noticeable differences (JNDs; Witzel & Gegenfurtner, 2013). This ensures that ensembles of different colors can be used with reasonable assurance of their equivalence in perceptual terms and allows us to use a broad stimulus set, which reduces the probability of trial-by-trial learning of the stimuli (Bauer, 2009a). This stimulus control also supports the validity of any findings regarding ensemble perception as errors due to failures of discrimination will be minimized.

Experiment 1 aimed to establish the basic conditions in which observers can reliably extract an average hue from a multihue ensemble and how hue averaging is affected by the number of elements and number of colors in the ensemble. The design was similar to that of the Marchant et al. (2013) study on the effect of set size and heterogeneity on estimations of mean size with manipulations of the number of elements in ensembles and number of different colors in ensembles. That study used limited combinations of these parameters for their analysis; we have included independent

manipulations of levels of both number of colors and number of elements.

Experiment 2a built on the results of Experiment 1, investigating the effect of varying the range of colors in ensembles on the accuracy of mean extraction. The overall design and aim was similar to that of Utochkin and Tiurina's (2014) replication and extension of the work of Marchant et al. (2013). Utochkin and Tiurina attempted to parse the effect of number of elements, the effect of number of different sizes, and the effect of difference between ensemble elements on estimations of mean size. They found evidence for the variance of sizes being a strong determinant of accuracy when estimating mean size.

Experiment 2b reanalyzed the data from Experiments 1 and 2a on the basis of color naming data in order to observe the influence of categories on the extraction of mean hue and further investigate the mechanism of hue averaging.

Overall, the experiments aim to further characterize the conditions under which rapid averaging of color occurs, explore whether rapid averaging appears to be a result of focused or distributed attention, and demonstrate the effect of hue range and categorization on mean estimation.

Experiment 1

Methods

Participants

Eighteen observers (14 female, mean age = 22.4 years, $SD = 2.2$ years) naive to the purpose of the experiment took part. All reported normal or corrected-to-normal visual acuity and were assessed as having normal color vision using Ishihara plates (Ishihara, 1973) and the Lanthony test (Lanthony, 1998). All were undergraduate students at the University of Sussex and were paid £4.50 for their time. The research protocol was approved by the university ethics committee.

Stimuli

A stimulus range consisting of 24 hues was specified from a circle on an equiluminant plane in Derrington-Krauskopf-Lennie (DKL) space (Derrington, Krauskopf, & Lennie, 1984; Krauskopf, Williams, & Heeley, 1982) (see Figure 1). These hues were spaced such that each differed from its neighboring hues by two JNDs as measured by Witzel and Gegenfurtner (2013). Because the stimuli are recreating those from Witzel and Gegenfurtner's measurements, their position in color space describes a circle in DKL space; however, the experiment uses the JND-scaled version of this circle, so the discriminability of neighboring hues is consistent throughout the circle and is not warped by inhomogeneities in the hue specification of that, or any other, color space. A gray background (xyY [1931]: 0.310, 0.337, 30.039) was used throughout the experiment.

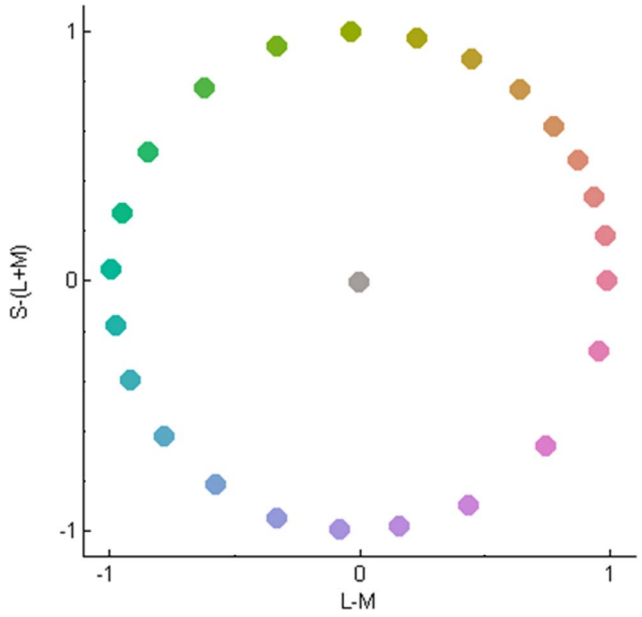

Figure 1. Arrangement of stimulus colors on an isoluminant plane in DKL space. Axes correspond to activation of the cone-opponent processes and are scaled arbitrarily according to the maximal monitor output described by Witzel and Gegenfurtner (2013). Adjacent hues are spaced by two JNDs (Witzel & Gegenfurtner, 2013), resulting in some inhomogeneity in the angular distance; discriminability is higher in some areas of this color space than in others. The gray point in the center represents the background color. In this (and all figures), colors are intended to approximate those used in the experiment but should not be taken to represent those colors precisely due to differences in printing and display equipment. This arrangement of hues was never shown to the observers in any experiment.

Apparatus

Stimuli were displayed on a 22-in. Mitsubishi DiamondPlus 2070SB Diamondtron CRT monitor with a resolution of 1600 × 1200 pixels, 24-bit color resolution, and a refresh rate of 100 Hz. Monitor primary values (red-green-blues) for all of the colors used in the experiment were selected manually using systematic adjustment of monitor primaries to output the correct xyY values as measured by a ColorCal colorimeter (Cambridge Research Systems).

The experiment took place in a blacked-out room with the monitor the only source of light. A cardboard viewing tunnel lined with black felt was used to obscure peripheral objects and colors that could otherwise be illuminated by the light from the monitor from the

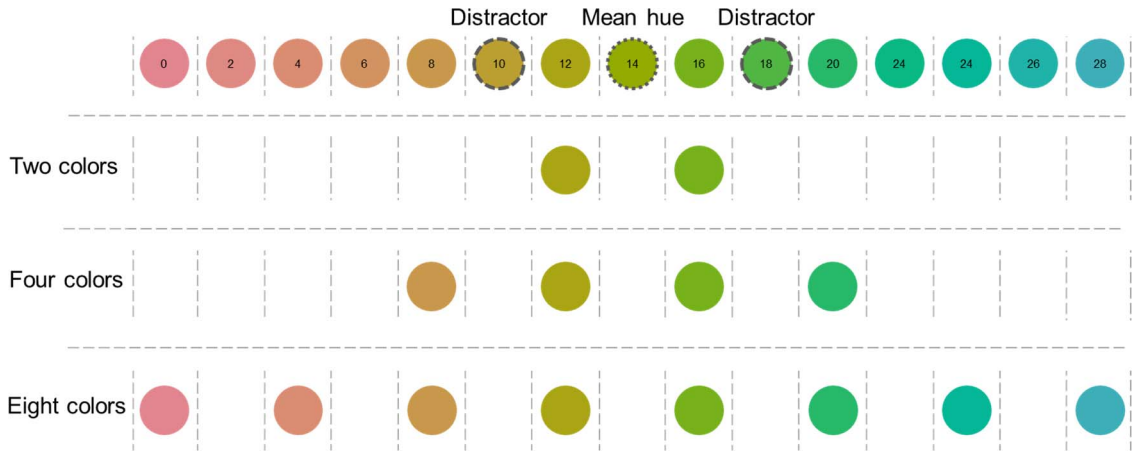

Figure 2. An example of stimulus selection for ensembles in Experiment 1. Numbers at the top refer to number of JNDs from the seed hue (leftmost, numbered 0). In this example, each distribution has a mean color corresponding to number 14, which would be paired with a distractor chosen from four JNDs either side for the 2AFC task. It should be noted that this is an example section of the hue continuum used, but in each trial, the seed color varied freely around the whole hue circle (Figure 1). Ensembles also varied in number of elements, but each hue was always present in equal number.

participant's field of vision. A chin rest was used to constrain viewing distance at 57 cm, and responses were given using the keyboard.

Design

Ensembles consisted of four, eight, or 16 circular, uniformly colored patches ("elements"). Elements were allocated at random to a cell in an invisible 4×4 grid subtending approximately 8° of visual angle. Each element patch subtended approximately 1.2° with a spatial location jittered randomly by up to 0.6° from the center of the allocated cell in the vertical and horizontal directions to remove the appearance of a regular arrangement of elements.

At the beginning of each trial, one hue was randomly selected from the 24-hue stimulus array. This "seed" hue, along with the required number of hues for the trial, was used to calculate the ensemble hues. As such, the particular segment of the hue circle represented by the elements in the ensemble varied randomly in each trial, reducing the possibility of trial-by-trial averaging affecting responses as can be caused by using a limited set of stimuli (Bauer, 2009a).

Each ensemble contained two, four, or eight different hues (see Figure 2), always represented in equal number across the elements. These two parameters (number of hues and number of elements) were varied independently such that each level was combined with the others when possible. This resulted in eight within-participant conditions (see Figure 3).

Following research into ensemble perception in other perceptual features, we tested recognition of average hue using a 2AFC task. In each case, the target patch was the "mean" hue of the preceding ensemble. Mean is defined as the centroid of the distribution of hues in the ensemble. The distractor hue was spaced four JNDs from the mean in either the clockwise or counterclockwise direction in DKL space, counterbalanced across trials. The 2AFC patches were the same size as the elements in the ensemble, arranged along the horizontal midline of the monitor to the left and right of the vertical midline and spaced by 3.5° of visual angle. The location of the target on the left or right was counterbalanced across trials.

Each trial began with a black fixation point, which appeared in the center of the display for 1000 ms; this was followed by the onset of the ensemble, present for 500 ms. After a blank screen interstimulus interval of 1000 ms, the 2AFC patches appeared. The 2AFC patches remained onscreen until the participant indicated his or her choice by pressing a key ("Z" for the left patch, "M" for the right patch). A 1000-ms intertrial interval followed. This is summarized in Figure 4.

Each participant completed four blocks of 96 trials. The eight element-color conditions were interleaved pseudorandomly within each block with the constraint that each block would present each condition 12 times. Thus, each participant provided responses to a total of 384 trials, 48 for each condition.

Procedure

Participants were briefed on the basic task of the study. On-screen instructions emphasized that participants should "try to get the gist of the set of colors" when shown the ensemble and then, when given the 2AFC patches, to "decide which you think is the mean color of the set of dots." If participants asked for

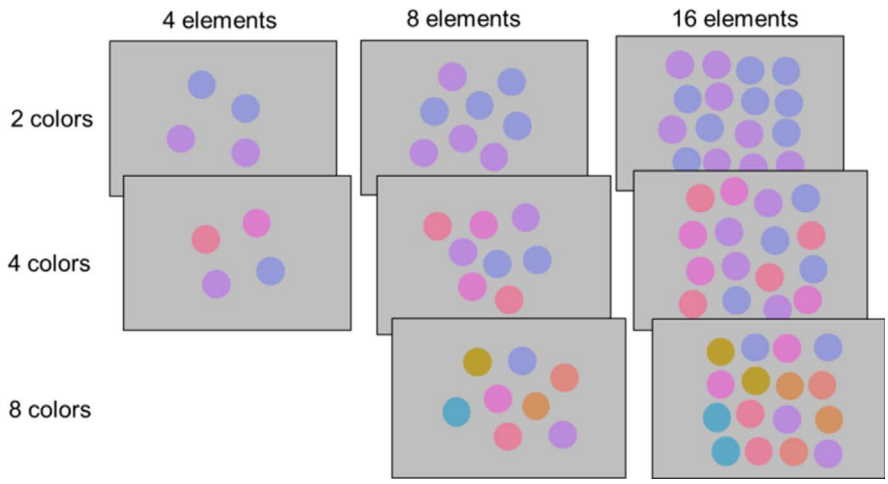

Figure 3. Variations in ensembles presented in Experiment 1. All of these ensembles have the same mean hue but vary in the number of elements and number of different colors.

clarification of how to interpret "mean color," the experimenter prompted them to "choose the color which you think best represents the whole set." Participants completed a short set of eight practice trials before beginning the experimental task. The time spent on reading and practice trials ensured adaptation to the white point. No feedback on performance was given at any point during practice or experimental trials (Bauer, 2009a).

Results

Data were screened for RT outliers prior to analysis. Trials with an RT less than 200 ms were removed along with trials in which the RT was more than three standard deviations above the participant's mean RT. This resulted in the removal of 143 trials (out of 6,912) or approximately 2% of trials. This rate of trial removal is comparable to the same RT screening process applied in a previous study on ensemble perception of hue (Maule et al., 2014). The interpretation of the statistical analysis that follows is the same whether these trials are removed or not.

A trial was coded correct if the participant selected the mean hue (i.e., the hue falling in the center of the ensemble distribution in terms of our two-JND hue circle) rather than the distractor in the 2AFC. The main dependent variables for analysis are the proportion of correct trials and RTs in correct trials with number of

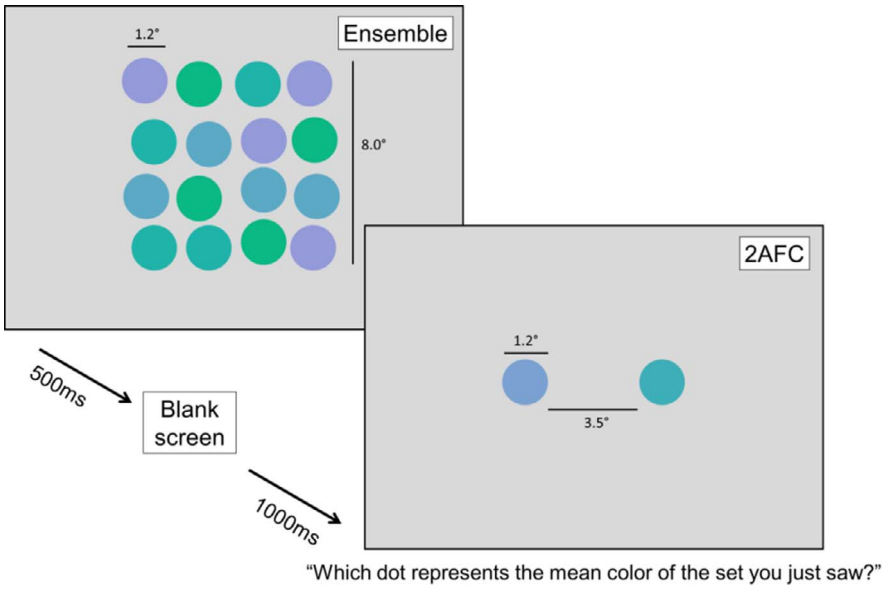

Figure 4. Trial structure, timing, and stimulus size and arrangement for Experiments 1 and 2. The 2AFC patches remained present until the participant responded.

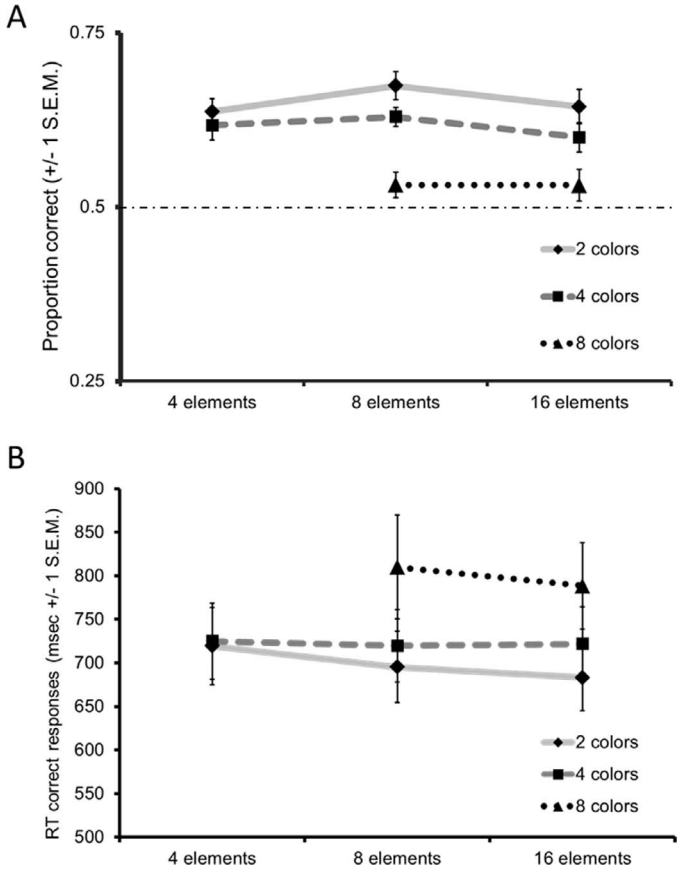

Figure 5. Results of Experiment 1. (A) Accuracy: mean proportion of trials in which the mean color was correctly chosen. The dotted line represents chance (0.5). (B) RTs: mean response latency for correct trials. All error bars represent 1 *SEM*.

elements and number of colors as factors. Figure 5 presents the across-participant mean proportion and RT of correct trials for each of the eight ensemble types. Inspection suggests that selecting the mean for an ensemble containing fewer colors is easier (more correct responses and faster RTs) than when there are more colors (the lines are "stacked" vertically), but there is no or little effect of number of elements.

Due to the "missing" cell in the elements × colors matrix (the impossible four-element, eight-color; see Figure 3), the analysis was divided into two repeated-measures ANOVAs. Reported effect sizes are generalized eta-squared (η^2_G) as recommended for use with repeated-measures designs by Bakeman (2005).

A 3 (number of colors: two, four, eight) × 2 (number of elements: eight, 16) repeated-measures ANOVA revealed a significant main effect of number of colors on accuracy, $F(2, 34) = 19.48$, $p < 0.001$, $\eta^2_G = .30$; no main effect of number of elements, $F(1, 17) = 1.68$, $p = 0.21$, $\eta^2_G = .01$; and no interaction between the factors, $F(2, 34) = 0.43$, $p = 0.66$, $\eta^2_G = .01$. These results were mirrored in a parallel analysis; a 2 (number of colors: two, four) × 3 (number of elements: four, eight, 16) repeated-measures ANOVA found a marginally non-significant main effect of number of colors, $F(1, 17) = 4.36$, $p = 0.05$, $\eta^2_G = .04$; no main effect of number of elements, $F(2, 34) = 1.86$, $p = 0.17$, $\eta^2_G = .02$; and no interaction, $F(2, 34) = 0.31$, $p = 0.74$, $\eta^2_G < .01$. Analysis of RTs in correct trials found a similar pattern. A 3 (number of colors: two, four, eight) × 2 (number of elements: eight, 16) repeated-measures ANOVA found a main effect of number of colors, $F(2, 34) = 15.75$, $p < 0.001$, $\eta^2_G = .06$, with no effect of number of elements, $F(1, 17) = 1.28$, $p = 0.27$, $\eta^2_G < .01$, and no interaction, $F(2, 34) = 0.73$, $p = 0.49$, $\eta^2_G < .01$. Figure 5 suggests that this is due to longer RTs being associated with more colors, particularly the eight-color ensembles.

After collapsing individual mean accuracy across elements, one-sample *t* tests (all two-tailed) revealed that performance was significantly above chance (50%) on two-color ensembles, $M = 0.65$, $SD = 0.07$, $t(17) = 9.19$, $p < 0.001$, and four-color ensembles, $M = 0.62$, $SD = 0.05$, $t(17) = 9.41$, $p < 0.001$, but was not above chance (albeit very marginally) for the eight-color ensembles, $M = 0.53$, $SD = 0.06$, $t(17) = 2.06$, $p = 0.06$.

Interim discussion

The results from Experiment 1 suggest that the number of elements in an ensemble has no effect on participants' ability to pick the mean hue of a briefly presented color ensemble whereas increasing the number of colors does have a deleterious effect on performance on this task. Likewise, RTs were longer for ensembles with more colors but insensitive to changes in number of elements, supporting the implication that more variegated ensembles are harder to average. The main effect of number of colors could be interpreted in a couple of ways. First, it could suggest that, although perceptual averaging of a stimulus attribute is robust when there are a manageable number of unique stimulus values in an ensemble, the averaging process becomes less accurate when this number increases (≥8). This would have implications for the suggestion that ensemble perception circumvents the limits of focused attention (e.g., Alvarez, 2011). An alternative explanation is that it is not the number of colors per se but the difference between them that makes averaging difficult. In this sense, it is the range in hue in the eight-color ensembles that depletes the ability to accurately choose the mean. It can be clearly seen in Figure 2 that range varies with number of colors in the ensembles. It is not possible to resolve the difference between these two explanations of the data from Experiment 1. Therefore, in Experiment 2a, we used fixed ranges for ensembles with

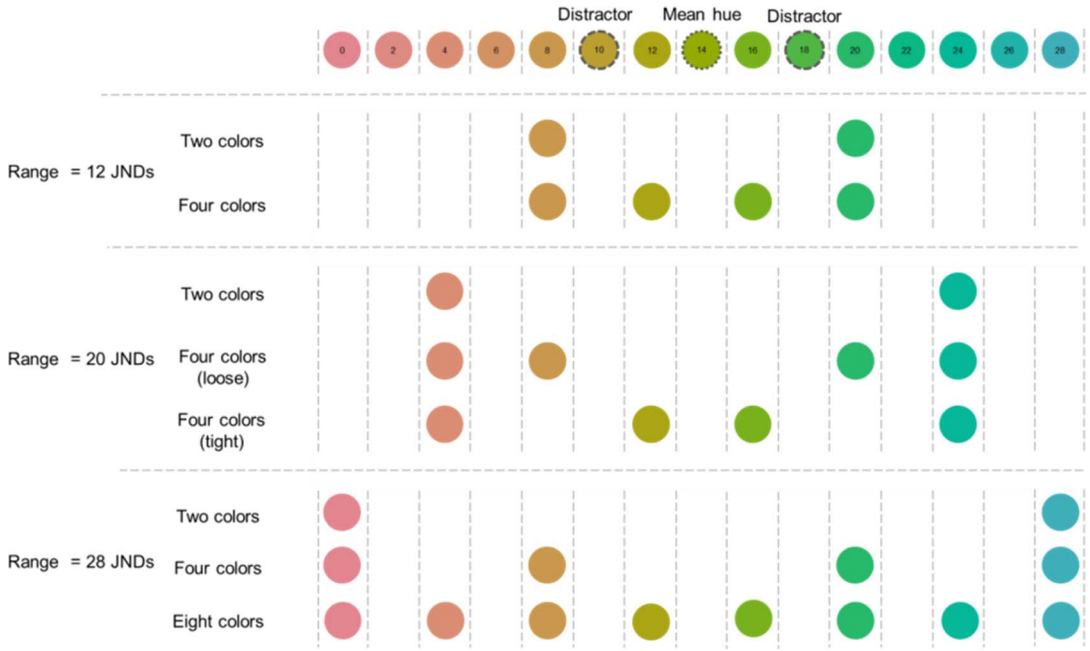

Figure 6. An example of stimulus selection for ensembles in Experiment 2. In contrast to the design in Experiment 1 (shown in Figure 2), the outermost hues are a fixed distance apart regardless of the number of intermediate colors that are included. Note the difference between "tight" and "loose" arrangements for the 20-JND, four-color ensembles. In Experiment 2, all ensembles had eight elements.

different numbers of colors (similar to Utochkin & Tiurina, 2014). By varying each factor independently, we can examine which is having the stronger effect on accuracy in the task.

Experiment 2a

Method

Participants

Nineteen observers (12 female, mean age = 22.7 years, $SD = 2.8$ years) took part. Observers were naive to the purpose of the experiment and none had taken part in Experiment 1. Visual acuity, color vision tests, and payment were as for Experiment 1.

Stimuli and apparatus

Range of hues and apparatus were as described for Experiment 1.

Design

The design of ensembles was similar to Experiment 1. Number of elements was fixed to eight throughout. Ensembles again consisted of two, four, or eight different hues with a range (i.e., the distance in JNDs from the hues at the extreme ends of the distribution for a given ensemble) fixed at 12, 20, or 28 JNDs. These conditions are summarized in Figure 6.

Fixing the range meant that the spacing between element hues was variable, depending on the range used. In the case of 20-JND, four-color ensembles, the resolution of our stimulus set (hues in two JND steps) meant that some ensembles had intermediate elements close to the mean ("tight"), and others had the intermediate elements closer to the extrema ("loose"—see Figure 6). These trials were counterbalanced through the experiment and are pooled together for analysis. Averaging these two arrangements results in an approximately equal perceptual spacing between elements without the need to present a distractor that matches an ensemble member. The main conditions for analysis were the two- and four-color ensembles across the three ranges, but an additional eight-color, 28-JND condition was included for comparison.

Procedure

Procedure, instructions, and the total number and arrangement of trials, conditions, and blocks were as described for Experiment 1.

Results

Data were screened using the same procedure as described for Experiment 1, resulting in the removal of

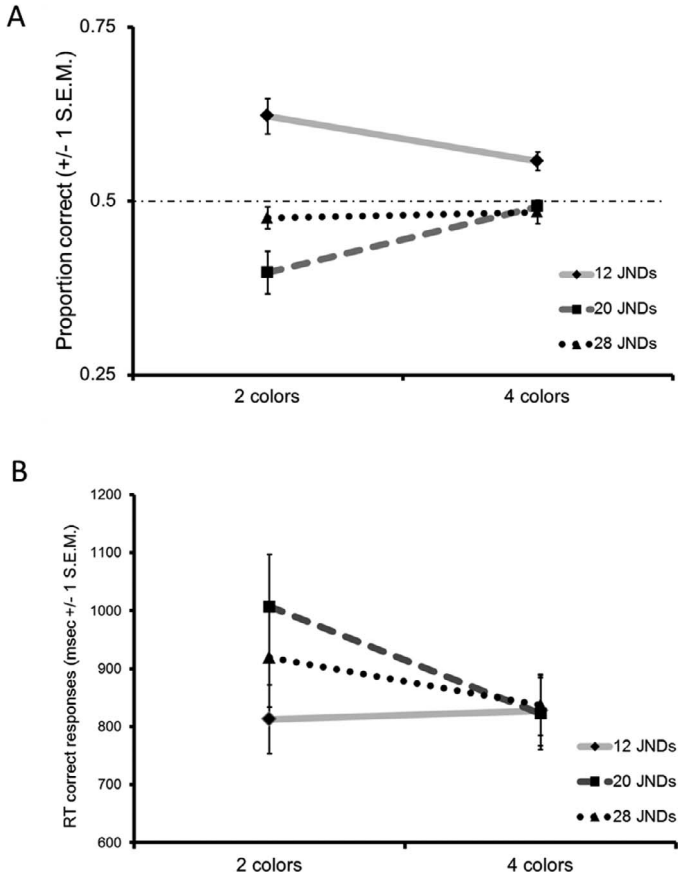

Figure 7. Results of Experiment 2a. (A) Accuracy: mean proportion of trials in which the mean color was correctly chosen from the 2AFC. The dotted line represents chance (0.5). (B) RTs: mean response latency for correct trials. All error bars represent 1 *SEM*.

191 trials (out of 7,296, approximately 3%). The interpretation of the statistical analysis that follows is the same whether these trials are removed or not.

Figure 7 presents the mean proportion and RT of correct responses for the two- and four-color ensembles across the different ranges. A 3 (range: 12, 20, or 28 JNDs) × 2 (number of colors: two or four) repeated-measures ANOVA found a significant main effect of range on accuracy, $F(2, 36) = 17.99$, $p < 0.001$, $\eta^2_G = .35$, and a significant range × colors interaction, $F(2, 36) = 8.08$, $p = 0.001$, $\eta^2_G = .13$. In contrast to the results of Experiment 1, there was no main effect of number of colors, $F(1, 18) = 1.78$, $p = 0.20$, $\eta^2_G = .01$.

Performance was generally worse than observed in Experiment 1 with several conditions appearing at or below chance. Although it appears that participants were most successful in the 12-JND condition, the relationship between performance and range in the 20- and 28-JND conditions is less clear. One-sample t tests (two-tailed) confirmed that, when collapsed across number of colors, only performance in the 12-JND condition was significantly above chance (50%), $M = 0.59$, $SD = 0.08$, $t(18) = 5.04$, $p < 0.001$. Performance was significantly below chance for 20-JND ensembles, $M = 0.44$, $SD = 0.07$, $t(18) = 3.29$, $p = 0.004$, and at chance for 28-JND ensembles, $M = 0.48$, $SD = 0.05$, $t(18) = 1.93$, $p = 0.07$. A one-sample t test on the eight-color, 28-JND condition (included for comparison; not shown in Figure 7) was also significant, $M = 0.53$, $SD = 0.07$, $t(18) = 2.12$, $p = 0.049$.

The below-chance performance for ensembles with a range of 20 JNDs warrants further attention as it indicates a systematic bias away from the colorimetric mean in this condition. The lack of a consistent effect found in the more wide-ranging ensembles suggests this could be an aberration caused by some feature of that condition (a potential explanation for this will be given in the General discussion). From inspection of Figure 7, it seems possible that this condition could be masking a main effect of number of colors and causing the appearance of an interaction. Therefore, in an attempt to verify the strength and nature of the main effects found, a further analysis was performed, this time excluding the below-chance 20-JND conditions. A 2 (range: 12 or 28) × 2 (number of colors: two or four) repeated-measures ANOVA found a significant main effect of number of colors, $F(1, 18) = 6.31$, $p = 0.02$, $\eta^2_G = .03$; a significant main effect of range, $F(1, 18) = 31.64$, $p < 0.001$, $\eta^2_G = .50$; and a nonsignificant interaction, $F(1, 18) = 3.60$, $p = 0.07$, $\eta^2_G = .05$.

A 2 (range: 12 or 28) × 2 (number of colors: two or four) repeated-measures ANOVA on the RTs for correct trials also found a significant main effect of range, $F(1, 18) = 9.33$, $p = 0.007$, $\eta^2_G = .01$, but no effect of colors, $F(1, 18) = 1.52$, $p = 0.233$, $\eta^2_G < .01$, and no interaction, $F(1, 18) = 1.84$, $p = 0.192$, $\eta^2_G < .01$. From Figure 7, it would appear that selecting the mean for the smallest range ensembles (12-JND) was quicker at least for the two-color condition.

Finally, the "tight" and "loose" variations of the 20-JND, four-color condition were compared. Accuracy in the "tight" condition ($M = 0.55$, $SD = 0.08$) was significantly higher than that in the "loose" condition ($M = 0.44$, $SD = 0.10$), $t(18) = 2.89$, $p = 0.010$. There was no difference in RTs for these two conditions ("tight": $M = 824$ ms, $SD = 347$ ms; "loose": $M = 828$ ms, $SD = 266$ ms), $t(18) = .06$, $p = 0.956$.

Interim discussion

The results of the second experiment indicate observers are able to extract the mean for small ranges (12 JNDs) but are unable to do so accurately for larger ranges (20 JNDs or more). This suggests that the effect of colors observed in Experiment 1 was not due to the difficulty of averaging a greater number of unique exemplars but rather is a result of greater perceptual

difference between the most extreme elements of the ensemble.

Due to the large interelement perceptual differences in this experiment, there is an additional variable that may be affecting the representation of the mean. Wider ranges may be more likely to be associated with ensembles containing elements from multiple color categories, and this may have a detrimental effect on the accuracy of hue averaging.

Because categories seem to affect postperceptual processes, rather than early, unconscious processing (He et al., 2014; Bird et al., 2014; Roberson et al., 2008), it will be of interest to see whether, for ensembles of equal range and number of colors, the presence of one or more categorical boundaries separating the elements affects the accuracy of mean hue encoding. If ensemble coding occurs early in visual processing, the categorical complexity of the ensemble (number of categories in the ensemble) should have no effect on the accuracy of mean encoding. If there is an effect of categories on mean accuracy, this may challenge views that perceptual averaging is automatic and compulsory (e.g., Parkes et al., 2001), instead reflecting cognitively effortful averaging, at least for color.

Experiment 2b

To address the question of the impact of categories on mean extraction, a subexperiment was conducted, applying color-naming data from an extra sample of observers to a reanalysis of the trials from Experiments 1 and 2a.

Method

Participants

Ten observers (eight female, mean age = 29.5 years, $SD = 8.2$ years) took part in the color-naming task. None had taken part in Experiments 1 or 2a. Visual acuity and color vision tests were as for Experiments 1 and 2a, and participants were paid £2 for their time. All were native British English speakers.

Stimuli and apparatus

The range of hues, background, and apparatus were as described for Experiment 1, but responses were given using a number pad.

Design

Each trial involved the presentation of a single, uniform, circular color patch—one of the 24 hues used in Experiments 1 and 2a—positioned centrally on the monitor and the same size as elements from ensembles (1.2° visual angle). The color patch was displayed for 500 ms, after which the patch disappeared and a key legend indicating which key on the number pad corresponded to which color name appeared in the bottom left-hand corner of the monitor. In every trial, observers had a choice of any of the eight English basic color terms: "green" (1), "brown" (2), "yellow" (3), "blue" (4), "orange" (6), "purple" (7), "pink" (8), and "red" (9) and gave their response on a USB number pad.

Procedure

Observers were briefed on the task with instructions emphasizing that they should respond based on their first reaction to the color because the patch would not be visible for long. Each observer completed 96 trials. These trials were split into four blocks (although there was no break between blocks for observers) with each block presenting each of the 24 hues once in a pseudorandom order.

Results

Figure 8 presents the arrangement of the consensus color category boundaries from the naming data collected. By taking the modal naming response to each hue across all trials and participants, a consensus naming map was determined. Across participants, agreement over the name given to each hue was high (mean agreement = 86%).

Following the establishment of the consensus position of category boundaries, the data from Experiments 1 and 2a were reanalyzed. First, the ensemble elements from every trial were recoded according to their consensus color category from the participants of Experiment 2b. The number of different categories present in that ensemble was then calculated for use as an independent variable in the analyses that follow. Conditions that were at or below chance performance in the original noncategorical analyses were excluded from the categorical analysis. Ensembles in the 12-JND, two-color condition of Experiment 2a were all two-category, so they have not been included. In both experiments, the four-color condition also yielded a small number of four-category ensembles; however, this represented too few trials to provide a reliable indication of performance (on average, seven trials per observer during Experiment 1 and two trials per observer during Experiment 2a). Therefore, the remaining conditions upon which the category analysis was performed were as follows: two-color ensembles with one or two categories (Experiment 1); four-color

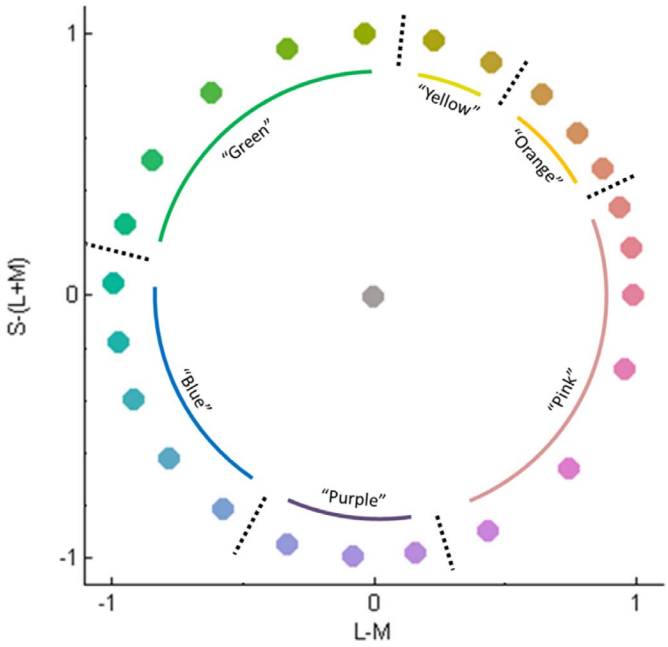

Figure 8. Reproduction of stimuli in DKL space (see Figure 1 caption for details) with consensus color categories obtained from the naming task of Experiment 2b, labeled for the 24-hue circle (see approximate color rendering) used throughout the experiments.

ensembles with two or three categories (Experiment 1); four-color ensembles with a range of 12-JNDs and two or three categories (Experiment 2a). Due to the main effects of number of color and range already established, comparisons across these parameters should be avoided, so the focus of the analysis is on the difference between levels of category only when other factors (number of colors, range, experiment) are the same.

Figure 9 shows the mean proportion and RTs of correct trials for ensembles with one, two, and three categories across the two- and four-color conditions from Experiment 1 and the 12-JND, four-color condition from Experiment 2a. Contrary to the expected disadvantage of averaging across a category boundary, in the two-color condition, the proportion of correct responses to ensembles with two categories ($M = .68$, $SEM = .02$) was slightly more accurate than responses to ensembles with one category ($M = .64$, $SEM = .02$) although this difference was not significant, $t(18) = 1.70$, $p = 0.106$. Likewise, in the four-color condition of Experiment 1, there was no significant difference between accuracy for two-category ($M = .61$, $SEM = .02$) and three-category ($M = .61$, $SEM = .01$) ensembles, $t(18) = .39$, $p = 0.699$, nor was there any difference between category levels in the 12-JND, four-color condition of Experiment 2a (two-category $M = .54$, $SEM = .03$; three-category $M = .55$, $SEM = .02$), $t(18) = 0.263$, $p = 0.795$. RTs revealed the same pattern with no significant differences due to number of

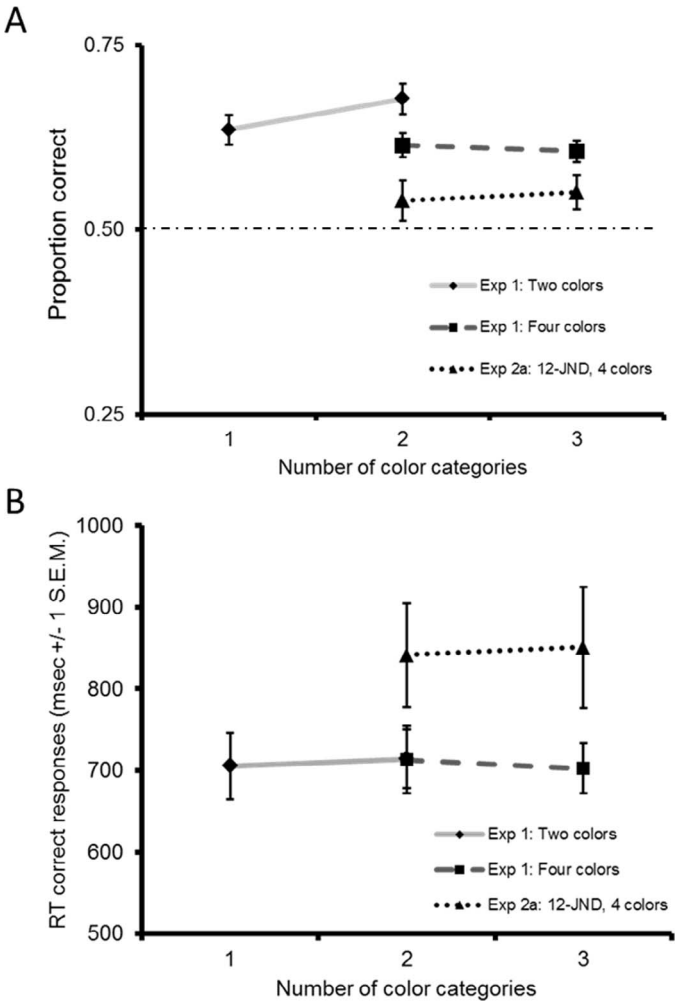

Figure 9. Results of Experiment 2b: reanalysis of selected conditions from Experiments 1 and 2a with trials recoded by number of color categories present in each ensemble. (A) Accuracy: mean proportion of trials in which the mean color was correctly chosen from the 2AFC. The dotted line represents chance (0.5). (B) RTs: mean response latency for correct trials. All error bars represent 1 *SEM*.

categories (largest $t = .424$, smallest $p = 0.677$). These results were replicated even when trials containing any color with low agreement on naming (<75%) are removed from the analysis.

Interim discussion

The results of the post hoc category analysis on the data from Experiments 1 and 2a show that the number of color categories represented by the elements of a multihue ensemble affects neither the accuracy of mean selection nor the time taken to identify the mean. This finding is concordant with both the postperceptual role for color categories in cognition (He et al., 2014) and the early and automatic nature of ensemble encoding

(e.g., Allik, Toom, Raidvee, Averin, & Kreegipuu, 2014; Corbett & Oriet, 2011; Im & Chong, 2014) and also supports our previous findings of the effect of categories on mean color familiarity (Maule et al., 2014).

General discussion

Overview of findings

The present study aimed to establish whether observers can extract an accurate mean hue from a rapidly presented multicolor ensemble and how this ability is affected by ensembles of varying number of elements, number of colors, and the amount of variation in hues in the ensemble.

In Experiment 1, both number of elements and number of different colors in ensembles were varied in order to explore the effects of these parameters separately as well as any interaction between them. The results revealed that observers are able to reliably extract the mean hue from the multicolored ensemble and that increasing the number of elements (essentially giving the observer more exemplars, more of the same information) had no effect on observers' ability to identify the mean hue following rapid exposure to an ensemble of different hues. However, increasing the number of colors in the ensemble was detrimental to the observers' ability to identify the mean hue. This could be due to there being more unique stimuli to average or, alternatively, could be caused by the stimuli being more different from one another. In Experiment 2a, the range of hues (i.e., the perceptual difference between the most extreme elements of an ensemble) and the number of colors were disassociated, making it possible to observe the effects of variation in stimuli and number of unique stimuli independently. The results demonstrated a crucial role for hue variation in observers' ability to extract the mean with selection for ranges greater than 20 JNDs being made at chance levels. Even with range of hue controlled, increasing the number of colors still affected observers' ability to select the mean hue although this effect was small. Finally, in Experiment 2b, based on color-naming data from an additional sample, a post hoc category classification of ensemble elements was performed in order to assess the accuracy of averaging across color category boundaries compared to ensembles consisting of elements of the same category. Results demonstrated no effect of categories on the accuracy of mean selection, suggesting that the averaging process occurs prior to the influence of categories on perception and cognition of color. The findings of this study have a number of implications for the literature on both ensemble perception and color cognition. These implications will each be discussed in turn.

Exhaustive processing versus subsampling

In Experiment 1, the observed consistency in averaging ability despite changes in the number of elements in ensembles supports the proposition that mean hue may be extracted from colorful ensembles using distributed attention. This is contrasted with the prediction regarding serial, focused attention on individual elements. A basic prediction of a focused-attention account is that, assuming the number of elements sampled is constant (but see Allik et al., 2014), averaging accuracy should decline with more elements as the subsample of elements to which attention is paid is more prone to bias. Marchant et al. (2013) found that adding more elements decreased accuracy of mean settings for ensembles in which each element had a unique size. Comparing to simulation data, they interpreted this finding as supporting a focused, rather than distributed, mode of attention and, hence, a subsampling mechanism guiding observer judgments. However, their design meant that ensembles with more elements also had more sizes present. In the current study, when number of elements and number of colors varied independently, this effect was not observed. Thus, we find better support for a distributed-attention account of the ability for averaging hue. Our results combined with results from similar manipulations with ensembles of different size elements (Utochkin & Tiurina, 2014) suggest that Marchant et al.'s finding may have been an artifact of the conflation of number of elements and number of unique sizes in their ensembles. It should be noted that there was no evidence for the improvement in performance, or reduction in RT, associated with having more elements as previously reported for ensembles of different size elements (Robitaille & Harris, 2011). Whether this is a result of a difference in the ensemble processing of these types of stimuli or some experimental factor is an area for further research.

It is important to note that although focused attention would entail a subsampling mechanism, distributed attention is not incompatible with a subsampling explanation for the mechanism in general. Although the observed insensitivity to number of elements lends weight to holistic processing explanations (e.g., Robitaille & Harris, 2011), a subsampling process immune to increasing elements is still possible. For example, attention might be distributed across all elements, but only a subsample is included when encoding summary statistics (see also Utochkin & Tiurina, 2014, p. 17). Indeed such a mechanism might help account for down-weighting or extraction of

outlying exemplars in other studies of ensemble perception (e.g., de Gardelle & Summerfield, 2011; Haberman & Whitney, 2010; Myczek & Simons, 2008).

Importance of range/variance to ensemble representations

From the current and previous studies, it is clear that the mean is not all-important in ensemble perception. Although many studies allude to summary statistics other than the mean (e.g., variance, range, presences of outliers), few have systematically varied those characteristics of ensembles (Haberman & Whitney, 2010; Im & Halberda, 2013; Maule et al., 2014; Utochkin & Tiurina, 2014; Webster et al., 2014). Summary statistics that represent the variance (or other measure of spread around the mean, such as standard deviation) are clearly of relevance to ensemble perception. Although it is now relatively uncontroversial to suggest that the mean characteristics (or an estimate of the mean) are often encoded in response to rapidly presented ensembles of various types of stimuli, it also seems likely that information pertaining to the variation around that mean is also encoded and used to guide perceptual judgments.

The present study has also added weight to the argument that the variance, as well as the mean, has a crucial role in the extraction of summary statistics, demonstrating that the range of colors present in an ensemble has a strong effect on the accuracy of mean judgments. Experiment 2a adds to growing evidence for the importance of range in summary statistical visual processing, and the advantage for accuracy in the "tight" compared to "loose" variations of the 20-JND, four-color ensembles give a further indication of the role that variance (i.e., interelement difference) might have in perceptual averaging, independently of range. Similarly, it has been shown that adaptation to mean size is weaker when ensembles contain more variance (Corbett et al., 2012), and models of mean size judgments are closer to actual observer performance when internal (i.e., judgment error) and external (i.e., ensemble variance) noise are included as factors affecting the judgment (Im & Halberda, 2013).

The results of the present study suggest that although multiple hues can be represented by their mean hue, this is subject to modulation by the variance (external noise) of the hues in the ensemble. As such, the pattern of results is similar to that expected from models representing ensembles (holistically and their elements individually) as a probability distribution or set of probability distributions, each subject to internal noise (e.g., Alvarez, 2011; Haberman & Whitney, 2012). Similarly, it has been shown that visual judgments of which of two groups of bars has the greater mean height involves assessment of the relative variance in each set as well as the mean difference, the process of which appears to follow that of Student's *t* test (Fouriezos, Rubenfeld, & Capstick, 2008).

Threshold of segmentation

The idea that variance, effectively the perceptual similarity of elements, drives the bias, accuracy, and strength of mean representations has previously been expressed in terms of the "threshold of segmentation" (Utochkin & Tiurina, 2014). According to this theory, the variance of an ensemble is crucial in the coding of an accurate mean representation. When the variance is high, elements are very different perceptually, and if no intermediate elements are included, segmentation occurs, the averaging mechanism fails to form a unifying mean (e.g., Webster et al., 2014), and hence, alternative decision processes take over when selecting a mean.

Our experiment is well placed to estimate the threshold of segmentation for ensemble perception of hue as our stimuli are controlled in JNDs. In this case, it appears that the threshold at which segmentation occurs and the mean can no longer be extracted from a two-color ensemble is between 12 and 20 JNDs. This is also reflected by the difference observed between the "tight" and "loose" variations of the 20-JND, four-color condition. The superior performance in the "tight" version (in which inner colors are separated by four JNDs compared to 12 JNDs in the "loose" version; see Figure 6) of this ensemble suggests an important role for variance in whether segmentation or averaging occurs; when the "inner" colors are more similar to each other, averaging is stronger compared to when they are more different. This suggests that, at least in ensembles with four colors, the extrema may have less impact on the accuracy of average representation—a process that, if true, would imply a mechanism identifying which exemplars differ greatly from the mean (see also outlier exclusion in face ensemble representation in Haberman & Whitney, 2010) serving to down-weight the extrema (e.g., "precision-weighted averaging," Alvarez, 2011; see also de Gardelle & Summerfield, 2011). Understanding which elements contribute most strongly to the representation of the set is certainly a matter that further ensemble perception research should seek to address through systematic manipulation of the interelement perceptual difference.

It should be noted that in Experiment 2a accuracy was slightly above chance in the 28-JND, eight-color condition, and it was at chance for the two- and four-color conditions at the same range. The low accuracy in the 28-JND conditions limits the conclusions somewhat, but such a pattern would point to further support

for the idea that smaller interelement differences support more accurate mean representation regardless of total range.

Some of the results of Experiment 2a run contrary to the findings and theories of mean size perception. In the 12-JND condition, rather than supporting mean encoding, the presence of intermediate colors (compare four-color to two-color conditions) was somewhat detrimental to performance. The cause of this decrement is unclear because holistic averaging, weighted-averaging, threshold of segmentation, and subsampling models all predict an improvement in the accuracy of mean representation given exemplars that fall closer to the mean. It is possible that this is an effect of the decision-making process when offered the mean and a distractor, both of which bear a quite close resemblance to some of the ensemble elements. As such, it may be that the responses to two-color ensembles reflect different processes due to the simplicity of the ensemble, and the more challenging four-color ensemble shows a pattern that is better explained by models entailing ensemble coding. In any case, the performance is still above chance, and the main finding, that range has a stronger effect than number of colors on the accuracy of mean encoding, is unaffected.

Below-chance performance

The below-chance level of accuracy for 20-JND, two-color ensembles deserves further attention. In this condition, the range and number of colors interact to bias the observer systematically to selecting the distractor hue over the mean hue. As the experiment used a full-hue circle with a mean color for ensembles sampled randomly in each trial and counterbalanced distractor locations and perceptual relationships to the mean, irregularities in the salience, perceived lightness, or perceptual spacing of the hues cannot account for a tendency toward choosing the distractor (and, particularly, it would not account for this phenomenon occurring in this condition only). A speculative explanation for this finding is that, if the colors are sufficiently different, deciding on a mean color is made more effortful, and because there are only two element colors to integrate, judgments are subject to be being biased more strongly toward the individual colors in the ensemble itself. Because the distractor is always more similar to one of the two element colors than the mean color is to either one individually, it may be that the distractor is erroneously chosen due to a bias toward the individual representations of the ensembles hues rather than the ensemble mean. The fact that this effect disappears when there are intermediate hues also present (i.e., in the 20-JND, four-color condition) could be taken to suggest that individual representations no longer bias the choice, perhaps due to the constraints of visual working memory inhibiting the encoding of individual items (e.g., Alvarez & Oliva, 2008; Attarha, Moore, & Vecera, 2014; Baijal et al., 2013; Chong & Treisman, 2005a; Corbett & Oriet, 2011; De Fockert & Marchant, 2008). When the hue range is even greater (i.e., 28 JNDs, two colors), we suggest that the difference between distractor and element hues is sufficiently large that similarity to element colors is not strong enough to bias responses toward either option of the 2AFC, so mean accuracy is near chance.

Circularity

An additional problem for the visual system in the averaging of such radically different hues is the circularity of hue space. In a perceptually uniform hue circle, averaging the angular position of any two colors (i.e., averaging along the perimeter of the hue circle) has two possible solutions: the "clockwise" midpoint between those hues and the "counterclockwise" midpoint. In Experiment 2a, the 28-JND, two-color ensembles recreated this problem; although the hues were separated by 28 JNDs in the direction of the colors presented for the 2AFC mean selection task, they were actually separated by only 20 JNDs in the other direction. It would be expected that, given only two hues, any hue-averaging mechanism would average across the smaller distance, and thus, the options given in the 2AFC are unlikely to have reflected the mean encoded. The fact that observers were at chance for this condition may reflect this. However, the presence of intermediate hues (in the four-color condition of the 28-JND range) should anchor the direction of hue averaging, providing a better guide as to from which direction the mean should be approached. Observers were equally unsure in this condition, however, with accuracy again falling at chance levels. The circularity of hue space, therefore, is probably partly responsible for the breakdown of averaging at wider ranges. However, the present data show that averaging along the hue circle is still reliable when the range is small.

Common mechanism of ensemble perception

In spite of the circular representation of hue and application of categorical labels to color perception, the results of this study suggest a similarity between the mechanism responsible for ensemble coding of color and ensemble coding of other features, such as size, length, orientation, and facial expression. Color averaging appears to be rapid; is insensitive to changes in number of elements, indicating distributed attention; and is range-limited but with a sensitivity to variance or

interelement difference. We find no evidence that the categorical relationships of ensemble elements affect mean encoding. When stimuli are appropriately controlled for perceptual difference, color categories only affect postperceptual processing (He et al., 2014), so the lack of a category effect is further support for early encoding of the mean.

Although the circularity of color space may be partly responsible for the breaking down of averaging at higher ranges, the results can still be compatible with models incorporating both internal and external noise, such as the threshold of segmentation (Utochkin & Tiurina, 2014). Therefore, in spite of the differences between color perception and that of other, more linearly represented features, it does appear that ensemble coding for color shares a common mechanism. Whether this mechanism is located in a specific part, or parts, of the brain that deals with all summary statistics or whether it is an emergent property of the organization of the visual cortex is a subject for further research.

Purpose of color averaging

Evidence that consistent summary statistics help speed visual judgments (Michael et al., 2014) and facilitate visual search (Corbett & Melcher, 2014b) support the idea that the function of an ensemble coding mechanism able to rapidly extract summary statistics is to tune the visual system to the characteristics of the environment. This proposition is also supported by evidence for adaptation to the mean of an ensemble (Corbett et al., 2012), evidence that ensemble means may be used as "units" of working memory (Im & Chong, 2014; Im, Park, & Chong, 2014) and are represented across multiple frames of reference (Corbett & Melcher, 2014a). Recent evidence also suggests that ensemble means can be used by 4- to 5-year-olds to guide their perceptual judgments (Sweeny, Wurnitsch, Gopnik, & Whitney, 2014), adding weight to the argument that ensemble coding is a pervasive feature of the visual system, and in some cases, averaging appears to be an obligatory process (e.g., Allik et al., 2014; Parkes et al., 2001), requiring little or no attention to the features in question (Alvarez & Oliva, 2008, 2009; Baijal et al., 2013).

For color perception, ensemble coding may also help tune the visual system to the environment. It has long been proposed that the average color of a scene could be used by the visual system to estimate the characteristics of the light illuminating the scene and thus support color constancy (e.g., "gray-world" hypothesis; Buchsbaum, 1980). It has also been demonstrated that modulation of the variance of colors surrounding another is sufficient to induce a change in color appearance of that focal patch (Brown & MacLeod, 1997); therefore, representation of the mean color of a scene may play a role in color constancy. Color summary statistics could also be relevant to the perception and memory of surface colors (e.g., Milojevic, Ennis, & Gegenfurtner, 2014). It is clearly inefficient and unnecessary to recall all of the precisely and subtly varying colors of an object in order to recognize it in the future. Instead, a summary representation could provide an adequate prior to aid these processes (Olkkonen & Allred, 2014; Olkkonen, McCarthy, & Allred, 2014). Additionally, although categories do not appear to affect ensemble coding of color, it is plausible that ensemble coding may inform our categorical structure. Infants as young as 10 months are known to be able to form category prototypes based on the mean features of a series of successively presented stimuli (Younger, 1985). Similarly, adult color category boundaries have been shown to shift toward the center of the range of stimuli offered (Wright, 2011). Therefore, it may be that ensemble statistics are used within the context of learning color categories and tuning them to the environment.

Conclusions

The present study has shown that a mean color can be selected with above-chance accuracy from rapidly presented ensembles containing multiple hues. The veracity of the process appears to be limited by the range or variance of the colors shown in the ensemble. Further research into ensemble coding (in any domain) should focus not solely on the mean (as has been common), but should also consider the role of the range and variance of the scene in the extraction of summary statistics with particular view to its contribution to segmentation of ensembles and the contribution of ensemble coding to visual stability. Future research on ensemble perception of color will present a unique opportunity to better understand the extent to which ensemble coding might be influenced by categorical relationships between ensemble elements and how ensemble coding might operate in the real world with regard to surface color perception and memory.

Keywords: ensemble coding, color, perceptual averaging

Acknowledgments

Thanks are due to Jessica Banks for assistance with data collection and to Christoph Witzel for sharing the hue discrimination data and assisting in accurate

rendering of colors. This research was funded by an ESRC grant to JM (ES/J500173/1) and an ERC grant to AF (CATEGORIES: 283605).

Commercial relationships: none.
Corresponding author: John Maule.
Email: j.maule@sussex.ac.uk.
Address: The Sussex Colour Group, School of Psychology, University of Sussex, Brighton, UK.

References

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2014). Obligatory averaging in mean size perception. *Vision Research, 101,* 34–40, doi:10.1016/j.visres.2014.05.003.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131, doi:10.1016/j.tics.2011.01.003.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392–398, doi:10.1111/j.1467-9280.2008.02098.x.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences, USA, 106*(18), 7345–7350, doi:10.1073/pnas.0808981106.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162, doi: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*(7), 1325–1326, doi:10.3758/Pp.70.7.1325.

Attarha, M., Moore, C. M., & Vecera, S. P. (2014). Summary statistics of size: Fixed processing capacity for multiple ensembles but unlimited processing capacity for single ensembles. *Journal of Experimental Psychology-Human Perception and Performance, 40*(4), 1440–1449, doi:10.1037/A0036206.

Baijal, S., Nakatani, C., van Leeuwen, C., & Srinivasan, N. (2013). Processing statistics: An examination of focused and distributed attention using event related potentials. *Vision Research, 85,* 20–25, doi:10.1016/j.visres.2012.09.018.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379–384, doi:10.3758/Bf03192707.

Bauer, B. (2009a). The danger of trial-by-trial knowledge of results in perceptual averaging studies. *Attention Perception & Psychophysics, 71*(3), 655–665, doi:10.3758/App.71.3.655.

Bauer, B. (2009b). Does Stevens's power law for brightness extend to perceptual brightness averaging? *Psychological Record, 59*(2), 171–185.

Bird, C. M., Berens, S. C., Horner, A. J., & Franklin, A. (2014). Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences, USA, 111*(12), 4590–4595, doi:10.1073/pnas.1315275111.

Brown, R. O., & MacLeod, D. I. A. (1997). Color appearance depends on the variance of surround colors. *Current Biology, 7*(11), 844–849, doi:10.1016/S0960-9822(06)00372-1.

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

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404.

Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics, 67*(1), 1–13, doi:10.3758/Bf03195009.

Chong, S. C., & Treisman, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900, doi:10.1016/j.visres.2004.10.004.

Clifford, A., Franklin, A., Holmes, A., Drivonikou, V. G., Ozgen, E., & Davies, I. R. L. (2012). Neural correlates of acquired color category effects. *Brain and Cognition, 80*(1), 126–143, doi:10.1016/j.bandc.2012.04.011.

Clifford, A., Holmes, A., Davies, I. R. L., & Franklin, A. (2010). Color categories affect pre-attentive color perception. *Biological Psychology, 85*(2), 275–282, doi:10.1016/j.biopsycho.2010.07.014.

Corbett, J. E., & Melcher, D. (2014a). Characterizing ensemble statistics: Mean size is represented across multiple frames of reference. *Attention Perception & Psychophysics, 76*(3), 746–758, doi:10.3758/s13414-013-0595-x.

Corbett, J. E., & Melcher, D. (2014b). Stable statistical representations facilitate visual search. *Journal of Experimental Psychology-Human Perception and Performance, 40*(5), 1915–1925, doi:10.1037/A0037375.

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual

averaging in the absence of individual item representation. *Acta Psychologica, 138*(2), 289–301, doi:10.1016/j.actpsy.2011.08.002.

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition, 20*(2), 211–231, doi:10.1080/13506285.2012.657261.

Daoutis, C. A., Pilling, M., & Davies, I. R. L. (2006). Categorical effects in visual search for colour. *Visual Cognition, 14*(2), 217–240, doi:10.1080/13506280600158670.

De Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics, 70*(5), 789–794, doi:10.3758/Pp.70.5.789.

De Fockert, J. W., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology, 62*(9), 1716–1722, doi:10.1080/17470210902811249.

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences, USA, 108*(32), 13341–13346, doi:10.1073/pnas.1104517108.

Demeyere, N., Rzeskiewicz, A., Humphreys, K. A., & Humphreys, G. W. (2008). Automatic statistical processing of visual properties in simultanagnosia. *Neuropsychologia, 46*(11), 2861–2864, doi:10.1016/j.neuropsychologia.2008.05.014.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate-nucleus of macaque. *Journal of Physiology-London, 357*, 241–265.

Drivonikou, G., Clifford, A., Franklin, A., Ozgen, E., & Davies, I. R. L. (2011). Category training affects colour discrimination but only in the right visual field. In C. P. Biggam, C. A. Hough, C. J. Kay, & D. R. Simmons (Eds.), *New directions in colour studies* (pp. 251–264). Amsterdam: John Benjamin.

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics, 70*(3), 456–464, doi:10.3758/PP.70.3.456.

Giesel, M., & Gegenfurtner, K. R. (2010). Color appearance of real objects varying in material, hue, and shape. *Journal of Vision, 10*(9):10, 1–21, doi:10.1167/10.9.10. [PubMed] [Article]

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology-Human Perception and Performance, 35*(3), 718–734, doi:10.1037/A0013899.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention Perception & Psychophysics, 72*(7), 1825–1838, doi:10.3758/App.72.7.1825.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.), *From perception to consciousness: Searching with Anne Treisman* (pp. 339–349). Oxford: Oxford University Press.

He, X., Witzel, C., Forder, L., Clifford, A., & Franklin, A. (2014). Color categories only affect post-perceptual processes when same- and different-category colors are equally discriminable. *Journal of the Optical Society of America a-Optics Image Science and Vision, 31*(4), A322–A331, doi:10.1364/Josaa.31.00a322.

Im, H. Y., & Chong, S. C. (2014). Mean size as a unit of visual working memory. *Perception, 43*(7), 663–676, doi:10.1068/P7719.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention Perception & Psychophysics, 75*(2), 278–286, doi:10.3758/s13414-012-0399-4.

Im, H. Y., Park, W. J., & Chong, S. C. (2014). Ensemble statistics as units of selection. *Journal of Cognitive Psychology, 27*(1), 114–127, doi:10.1080/20445911.2014.985301.

Ishihara, S. (1973). *Ishihara's test chart for colour deficiency*. Tokyo: Kanehara Trading Inc.

Krauskopf, J., Williams, D. R., & Heeley, D. W. (1982). Cardinal directions of color space. *Vision Research, 22*(9), 1123–1131, doi:10.1016/0042-6989(82)90077-3.

Kuehni, R. G. (2014). Unique hues and their stimuli-State of the art. *Color Research and Application, 39*(3), 279–287, doi:10.1002/Col.21793.

Kuriki, I. (2004). Testing the possibility of average-color perception from multi-colored patterns. *Optical Review, 11*(4), 249–257, doi:10.1007/s10043-004-0249-2.

Lanthony, P. (1998). *Album Tritan* (2nd ed.). Paris: Laboratoire de la Vision des Couleurs.

Lanzoni, L., Melcher, D., Miceli, G., & Corbett, J. E. (2014). Global statistical regularities modulate the speed of visual search in patients with focal attentional deficits. *Frontiers in Psychology, 5*(514), 1–12, doi:10.3389/Fpsyg.2014.00514.

Leib, A. Y., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent

average crowd identity. *Journal of Vision, 14*(8):26, 1–13, doi:10.1167/14.8.26. [PubMed] [Article]

Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia, 50*(7), 1698–1707, doi:10.1016/j.neuropsychologia.2012.03.026.

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*(2), 245–250, doi:10.1016/j.actpsy.2012.11.002.

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: Metric and categorical effects on ensemble perception of hue. *Journal of the Optical Society of America A, 31*(4), A93–A102, doi:10.1364/Josaa.31.000a93.

Michael, E., de Gardelle, V., & Summerfield, C. (2014). Priming by the variability of visual information. *Proceedings of the National Academy of Sciences, USA, 111*(21), 7873–7878, doi:10.1073/pnas.1308674111.

Milojevic, Z., Ennis, R. J., & Gegenfurtner, K. R. (2014). Color classification of leaves [Abstract]. *Perception, 43*, 146.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70*(5), 772–788, doi:10.3758/Pp.70.5.772.

Olkkonen, M., & Allred, S. R. (2014). Short-term memory affects color perception in context. *Plos One, 9*(1), e86488, doi:10.1371/journal.pone.0086488.

Olkkonen, M., McCarthy, P. F., & Allred, S. R. (2014). The central tendency bias in color perception: Effects of internal and external noise. *Journal of Vision, 14*(11):5, 1–15, doi:10.1167/14.11.5. [PubMed] [Article]

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744, doi:10.1038/89532.

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition, 28*(6), 977–986, doi:10.3758/Bf03209345.

Roberson, D., Pak, H., & Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition, 107*(2), 752–762, doi:10.1016/j.cognition.2007.09.001.

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision, 11*(12):18, 1–8, doi:10.1167/11.12.18. [PubMed] [Article]

Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics, 70*(7), 1335–1336, doi:10.3758/Pp.70.7.1335.

Sunaga, S., & Yamashita, Y. (2007). Global color impressions of multicolored textured patterns with equal unique hue elements. *Color Research and Application, 32*(4), 267–277, doi:10.1002/Col.20330.

Sweeny, T. D., Haroz, S., & Whitney, D. (2012). Reference repulsion in the categorical perception of biological motion. *Vision Research, 64,* 26–34, doi:10.1016/j.visres.2012.05.008.

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2014). Ensemble perception of size in 4-5-year-old children. *Developmental Science, 18*(4), 556–568, doi:10.1111/desc.12239.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*(4–8), 411–443, doi:10.1080/13506280500195250.

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, doi:10.1016/j.actpsy.2013.11.012.

Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual-system averages speed information. *Vision Research, 32*(5), 931–941, doi: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*(1), 47–59, doi: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*(4), A283–A292, doi:10.1364/Josaa.31.00a283.

Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review, 18*(3), 484–489, doi:10.3758/s13423-011-0071-3.

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision, 13*(7):1, 1–33, doi:10.1167/13.7.1. [PubMed] [Article]

Wright, O. (2011). Effects of stimulus range on color categorization. In C. P. Biggam, C. A. Hough, C. J. Kay, & D. R. Simmons (Eds.), *New directions in colour studies* (pp. 265–276). Amsterdam: John Benjamin.

Younger, B. (1985). The segregation of items into categories by ten-month-old infants. *Child Development, 56*(6), 1574–1583.