Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism

Article (Accepted Version)

This is a pre-print version of the article. The final version is available here:
https://www.osapublishing.org/josaa/abstract.cfm?uri=josaa-33-3-A22

Accurate rapid averaging of multi-hue ensembles is due to a limited capacity sub-sampling mechanism

JOHN MAULE,[1,*] ANNA FRANKLIN,[1]

[1]*The Sussex Colour Group, School of Psychology, University of Sussex, Brighton, UK*
**Corresponding author: J.Maule@sussex.ac.uk*

Received XX Month XXXX; revised XX Month, XXXX; accepted XX Month XXXX; posted XX Month XXXX (Doc. ID XXXXX); published XX Month XXXX

It is claimed that the extraction of average features from rapidly-presented ensembles is holistic, with attention distributed across the whole set. We investigated whether observers' extraction of mean hue is holistic or could reflect sub-sampling. Analysis of selections for the mean hue revealed a distribution which peaked at the expected mean hue. However, an ideal observer simulation suggested that a sub-sampling mechanism incorporating just two items from each ensemble would suffice to reproduce the precision of most observers. The results imply that hue may not be averaged as holistically and efficiently as other attributes.

OCIS codes: *(330.0330) Vision, color and visual optics; (330.5020) Perception psychology; (330.5510) Psychophysics*

http://dx.doi.org/10.1364/xxxxxxxx

1. INTRODUCTION

Ensemble perception describes the extraction of summary statistics from a set of items varying in some stimulus dimension, typically in the absence of representation of the individual items and with very short stimulus presentation time [1]. For example, observers can extract the mean size from a set of circles of different sizes seen for 500ms, but are relatively poor at identifying individual members of the set [2]. Sensitivity to summary statistics has been demonstrated in a variety of perceptual domains, including size [e.g., 2, 3-5], orientation [e.g., 6, 7], facial expression [e.g., 8, 9, 10], facial identity [e.g., 10, 11, 12], and color [e.g., 13, 14, 15]. Much of this research has focused on whether the mean value of an ensemble has a special perceptual salience. The encoding of a mean in spite of a lack of individual item representation has led to suggestions that the ensemble perception mechanism could operate outside of the limits of focused attention, instead using distributed attention to process sets holistically [see 16, 17]. However this mechanism is subject to debate, with various researchers pointing out that a mechanism combining focused attention with sub-sampling might be adequate to explain observers' performance on perceptual averaging tasks, without the need to postulate a new holistic processing mechanism [18-21].

Summary statistics of color are likely to be of relevance to the visual system. For example, the color variance in surrounds is known to modulate the appearance of individual colors [22, 23], and priming by the variance of color present in a rapidly-presented ensemble has also been reported [24]. The mean color of a scene may also play a role in the estimation of the illuminant, necessary for color constancy ["gray world hypothesis", 25], and in color memory [26]. We have previously shown that, when observers are presented with an ensemble of two different hues for a short time (500ms), they tend to have a bias in their memory of which hues were present in the ensemble towards the mean hue even if that mean hue was not present [14]. We have also shown that observers can reliably identify the unseen mean hue of an ensemble when that hue is paired with a similar distractor hue [13]. Both of these studies found an effect of the range of hues in the ensemble where the mean bias and mean identification accuracy were both reduced when the range of ensemble hues was increased [see also 15]. We further found that there is no impact of increasing the number of elements in an ensemble – observers were able to identify the mean equally reliably whether required to average 4, 8 or 16 patches of color [13]. The robustness of mean identification ability to changes in number of elements has also been demonstrated for ensemble perception of size [e.g., 2, 18, 27-29] and faces [12, 30], and is suggestive of an efficient mechanism where processing occurs in parallel, across the whole display and all items [17].

Although we have shown that mean identification is above chance on a two-alternative forced-choice (2AFC) task [13], no study has directly investigated the precision of mean representation following rapidly-presented ensembles. Kuriki [31] has shown that adjustments to mosaics with many tiny elements were not reflective of the colorimetric mean, being biased towards the most saturated element. However this was under continuous viewing conditions [see also 32], rather than the rapid-exposure of the ensemble perception paradigm.

We know that number of elements has no effect on the identification of the mean color given a 2AFC [13] and Ariely [2, 33] reasoned that a sub-sampling mechanism with a fixed sample size, should extract the mean with precision proportional to the set size. However, ideal-observer simulations have suggested that the performance of actual observers in a number of experiments showing rapid extraction of mean size [2, 3, 27, 34] could be explained by a limited sub-sampling mechanism with a sample size as small as just one or two items from each set [18, 19]. Similarly, Marchant and de Fockert [18] showed that their finding that mean size estimates are affected by set size, for irregular sets (ensembles in which all elements have a unique size) but not regular sets (where some elements are the same size), can be predicted by a limited sub-sampling model. Other simulations [e.g., 8, 35] have attempted to better characterize the process of mean estimation by including internal noise into simulations – i.e. the "judgment error" [33] present in all psychophysical measurements. Simulations of sub-sampling which incorporate or estimate internal noise as part of the model tend to perform less well compared to real observer data and suggest that larger sub-samples (around seven items), would in fact be required to simulate the averaging performance of human observers [35]. Likewise, experimental evidence also suggests that observers still outperform subsampling expectations even when explicitly instructed to use such strategies in ensemble perception tasks [36], and when ensembles contain a manageable range of stimuli [29]. Such simulations have therefore shown that a sub-sampling mechanism or strategy cannot account for the level of performance observed on ensemble perception tasks in these domains (faces and size).

The present study investigated whether adjustments to the mean hue for rapidly-presented ensembles are equivalent to adjustments to a single hue. We assess whether settings converge at the expected value (the exact hue shown for single hues, or the mean hue for heterogeneous ensembles), as indicated by the position of the peak in the distribution of settings relative to the expected mean/actual color. We analyze the average amount of error in these settings to indicate how variable these settings are across trials. This measure of variability tells us how precise settings are. Although an observer's settings of the mean hue may converge at the expected hue, there might also be large variability in their settings indicating that hue averaging is not precise. If the precision of settings around the expected hue is similar when setting to the average of a set of hues or setting to a single hue, it will be a strong indicator that the mean hue is encoded as strongly as individual hues, suggesting that the ensemble is represented by a single average hue. These measurements of precision were also used to address the question of whether the observed hue averaging precision could be the result of a limited sub-sampling mechanism or whether the performance could support the proposal of an efficient holistic mechanism, integrating attention from the whole ensemble and circumventing the limited capacity of visual working memory [16, 37-39]. Measurements of internal noise (based on adjustments to single hues) were incorporated into a simulation estimating the precision of mean adjustments based on a random sub-sample of n elements. By comparing the simulation results to the precision of real observers in the ensemble adjustment task it was possible to estimate the sample size required to explain their performance by sub-sampling.

2. METHOD

A. Participants

Fifteen observers (three males) of average age 20.5 years (SD = 2.97) took part in the experiment. All reported normal or corrected-to-normal visual acuity and were assessed as having normal color vision using the Ishihara plates [40] and City University test [41]. All spoke English as their first language. Participants received either payment at a rate of £7.50 per hour or course credit. The research protocol was approved by the University of Sussex Sciences and Technology Cross-Schools Ethics Committee.

B. Stimuli

All colors were taken from a set of 48 hues specified from a circle on an equiluminant plane in Derrington-Krauskopf-Lennie (DKL) space [42, 43]. In order to ensure that the colors were approximately equally discriminable, and thus provide uniform perceptual differences between the hues presented in ensembles, the hue discrimination threshold data from Witzel and Gegenfurtner [44] were used to space the selected hues by 1 just-noticeable difference (JND) (see figure 1). Throughout the experiment the background was a uniform grey (xyY (1931): 0.310, 0.337, 30.039), as used by Witzel and Gegenfurtner.

C. Apparatus

A 22-inch Mitsubishi DiamondPlus 2070SB Diamondtron CRT monitor, set to a resolution of 1600 x 1200 pixels, 24-bit color resolution, and a refresh rate of 100 Hz was used. A Cambridge Research Systems ColorCal colorimeter was used to measure the monitor gamut and primary outputs, gamma correction applied, and look-up tables generated to automatically estimate the RGB primary values required to render each desired stimulus color. 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 obscured peripheral objects from the participants' view and a chin rest was used to maintain a viewing distance of approximately 57cm. Participants gave their responses using the keyboard.

D. Design

Ensembles consisted of sixteen colored circles (elements) each allocated to a cell in an invisible 4-by-4 grid centered on the screen. Elements subtended 1.75° visual angle and were spatially jittered by up to 0.25° visual angle horizontally and vertically from the center point of the cell to remove the appearance of a regular structure in the ensemble. Ensembles contained either elements all of one hue (homogeneous trials) or four hues, i.e. four elements of each hue, arranged randomly (heterogeneous trials). Hues presented together in heterogeneous trials were separated by 2 JNDs.

The task used a method of adjustment, in which participants first saw an ensemble and then attempted to match the average color of the ensemble in a subsequent display. The adjustment display was an ensemble of 16 elements arranged in a 4-by-4 grid (un-jittered), and all elements of the adjustment display were the same color. Trials began with the presentation of a black fixation point in the center of the display for 1000ms, immediately followed by the presentation of a 'study' ensemble for 500ms. An inter-stimulus interval lasting 1000ms was indicated by a white fixation point and then replaced by the adjustment ensemble. The initial color of the elements of the adjustment ensemble was selected at random from a range +/-7 JNDs from the actual mean of the ensemble. By pressing the left and right arrow keys participants were able to adjust the color of the elements in the ensemble, around the hue circle in 1 JND steps. The space bar was used to confirm the selection for that trial.

Participants took part in five blocks of trials. Each block presented trials from a list comprising 48 heterogeneous ensembles (i.e. ensembles with a mean corresponding to one of the 48 hues in the stimulus set) and all 48 homogeneous ensembles, in a random order.

E. Procedure

Participants read instructions on the screen prior to the task. The time spent reading instructions and completing practice trials ensured adaptation to the white point. The instructions stated that participants should pay attention to the initial ensemble and then "adjust the dots until they match the average color of the first set" (for heterogeneous ensembles) or "match the color exactly" (for homogeneous ensembles). It is possible that the concept of a perceptual average of

color is less intuitive than the average of stimuli such as size or facial expressions which have clearly observable features that vary on quantifiable dimensions rather than being simple patches of light. Therefore, in order to help explain the concept of the 'average' participants were given practice trials where they were asked to average achromatic stimuli varying in lightness. It was felt that training using ensembles that vary in lightness was appropriate for helping participants understand the concept of an average hue as both types of stimuli are simple patches of light. Participants completed 10 lightness practice trials using ensembles of achromatic stimuli, varying in lightness relative to the background (8 – 48 cd/m², in 4 cd/m² steps). The practice included feedback to indicate if the participant's selection was "correct" (at the mean lightness, also the mid-point of the range shown in the ensemble), "close" (within one step of the correct response) or "incorrect". Participants were required to be "correct" or "close" on each of the last three practice trials in order to proceed to the main task, otherwise the practice was repeated. The task appeared to effectively explain the concept of a perceptual average as the majority of participants proceeded to the color task after one run of 10 practice trials, apart from one who required 20 trials in total. No feedback on performance was given during the color task. At the end of the color task participants were asked about the strategy that they used on the task. No participant reported adopting a conscious strategy other than looking at the set of colors and deciding on the average and none reported difficulty in understanding the concept of an average color.

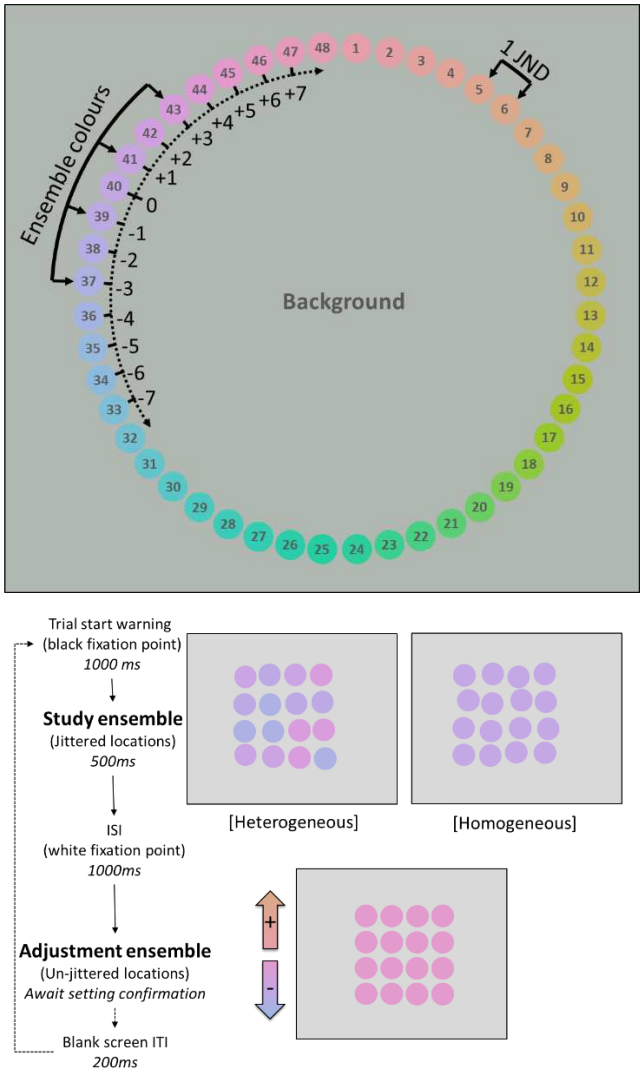

Fig. 1. Upper panel – an approximate rendering of the 48 hues, and the background, as used in the experiment, arranged in a continuous hue circle. Adjacent hues are separated by 1 JND. The solid black line towards the top left indicates the selection of hues for ensembles – each ensemble had four different hues, drawn from a 6-JND span with 2-JNDs between each exemplar. This arrangement moved at random around the hue circle on each trial to present ensembles with different mean hues, but with the spacing of the element colors yoked in the way shown. The dotted line inside represents the adjustment phase at which participants could select any hue from the circle, moving in single JND steps in the positive (clockwise) or negative (counterclockwise) direction. These responses are coded according to their JND-distance from the ensemble mean, which was assumed to fall at the middle of the distribution of ensemble colors. Lower panel – the order and timing of events in a single trial of the ensemble task. JND = Just-noticeable difference; ISI = inter-stimulus interval; ITI = inter-trial interval.

3. RESULTS

A. Homogeneous vs Heterogeneous ensembles

In all cases and conditions the settings peaked at the expected mean, indicating that observers were able to average the hues. Observer settings were coded by their absolute error from the actual mean of ensembles, in terms of 1 JND steps around the hue circle. For heterogeneous ensembles this was assumed to be the mid-point of the distribution of hues which were present in the ensemble, for homogeneous ensembles this was the hue matching those used in the ensemble. Mean absolute error (i.e. error in either hue direction from the correct mean) was significantly greater for the heterogeneous ensembles (M = 2.02, SD = 0.25) than the homogeneous (M = 1.34, SD = 0.21) ($t(14) = 9.44$, $p < .001$). This can be seen in the data presented in figure 2 (selected individuals) and figure 3 (average observer) - the distribution of selection errors around the mean is greater (a wider normal curve with a greater standard deviation) in the heterogeneous condition compared to the homogeneous.

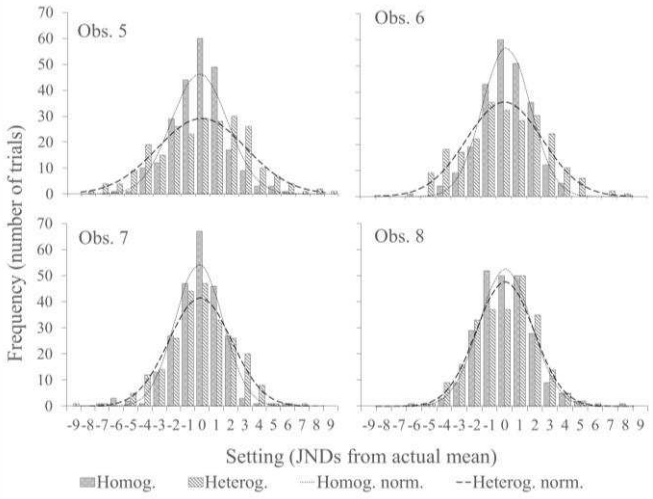

Fig. 2. Error distribution histograms for homogeneous and heterogeneous conditions for four example observers typical of the whole sample. Dashed curves indicate normal distributions with a mean and standard deviation (SD) equal to that for each observer and condition. N.B. Settings outside the range of +/-9 JNDs are not displayed by the histogram, but do contribute to the mean and SD of the normal curves.

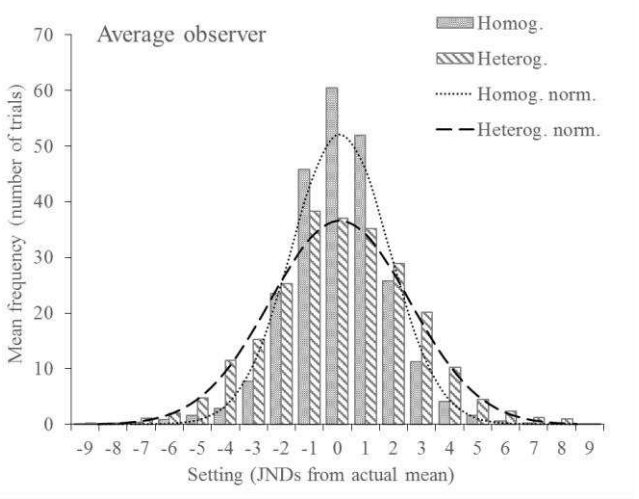

Fig. 3. Error distribution histograms for homogeneous and heterogeneous conditions for the average observer. Bars are based on mean frequency of response across observers. Dashed curves indicate normal distributions with a mean and standard deviation (SD) equal to the mean for each observer and condition. N.B. Settings outside the range of +/-9 JNDs are not displayed by the histogram, but do contribute to the mean and SD of the normal curves.

B. Simulation of Limited-Capacity Sampling Strategies

In order to evaluate observers' performance in the heterogeneous condition (when they are required to pick a single color to represent a multi-color ensemble), relative to the homogeneous condition (where they needed simply to match the single color present in the ensemble), an ideal observer simulation was carried out. This analysis sought to ascertain how many single ensemble elements an observer would have to sample in order to achieve performance at the level observed in the heterogeneous condition.

Two models were used for the simulation (figure 4). Both involved a *sampling* of elements from an ensemble composed exactly as in the adjustment experiment, followed by *averaging* of that sample, and finally *selection* from the available hues. The *early noise* model [8] applied noise to the representation of the colors at the sampling stage, such that each sampled element would be represented by a value selected from a normal distribution with a mean equal to the true element value and a standard deviation (SD) equal to that observed for settings in the homogeneous condition. The *late noise* model [8] applied noise after the averaging stage, such that the color representing the whole ensemble was subject to noise prior to selection. In both models selection involved rounding to the nearest integer. Simulations were run for each observer, using their individual SD from the homogeneous condition, and performance 10,000 trials.

These simulations, like those used in similar studies [e.g., 8, 35, 45], assume that the level of internal noise is constant across sample size. It is possible that a sub-sampling mechanism would have noise which increases with sample size, meaning that the benefit of larger samples being more representative is diminished by increased noise [see 16]. As there is no readily available estimate of how internal noise might be affected by sample size to include in the model, the results should be considered with this assumption in mind.

The results of the simulation are considered in terms of precision of performance, summarized by the standard deviation of error from the true mean in the simulated adjustment settings. By comparing these to the standard deviation of adjustment settings in the observer data it is possible to evaluate, given a limited-capacity sampling strategy or mechanism, how many elements an ideal observer would need to sample in order to reach or exceed the level of performance exhibited by the observers during the mean adjustment task.

The simulation revealed that most observers were performing at a level equivalent to sampling between one and two elements from each ensemble. This was true of both the early and late noise models. Figure 5 shows the simulated data for four observers with the actual performance also plotted for comparison. The simulation data show that there are diminishing returns from taking more and more samples, and in the case of the late noise simulation, an optimum number of samples is reached at around six or seven elements. Importantly, however, only one observer (obs. 8) exhibited performance near this optimal level.

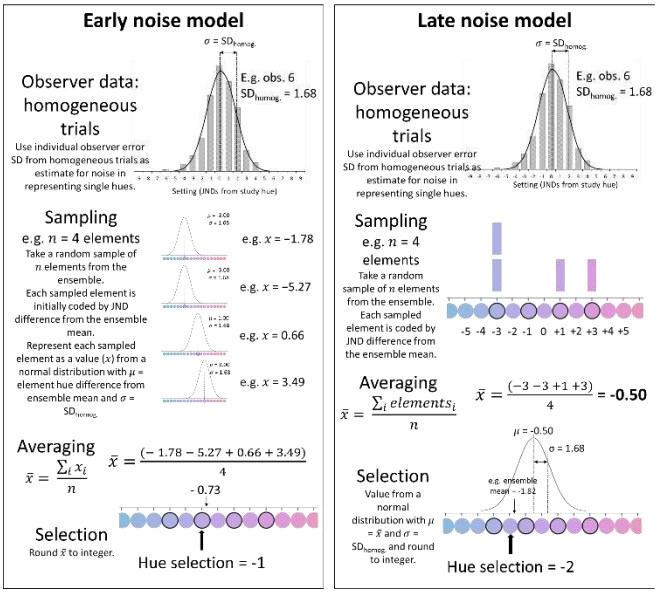

Fig. 4: Schematic representation of the early and late noise simulations. In the early noise model noise is added to each sampled element prior to averaging. In the late noise model each element is first averaged, after which noise is added to the mean representation. Selection of the eventual mean hue response requires rounding to the nearest integer. Noise is equivalent to the observed standard deviation of settings from each real observer's responses to homogeneous (i.e. single-hue) ensembles. Both panels represent a single exemplar trial where the same 4 elements are sampled from an ensemble and noise is based on observer 6. Note that the simulation was run for each observer and at sample sizes of 1-16 elements, for 10,000 trials each. n = number of samples; μ = mean of normal distribution indicating a noisy representation of a hue; σ = standard deviation of normal distribution; x = value assigned to a sampled element prior to averaging; $\bar{x}$ = calculated value for mean hue. All values given are in terms of JNDs from the true mean hue of the ensemble.

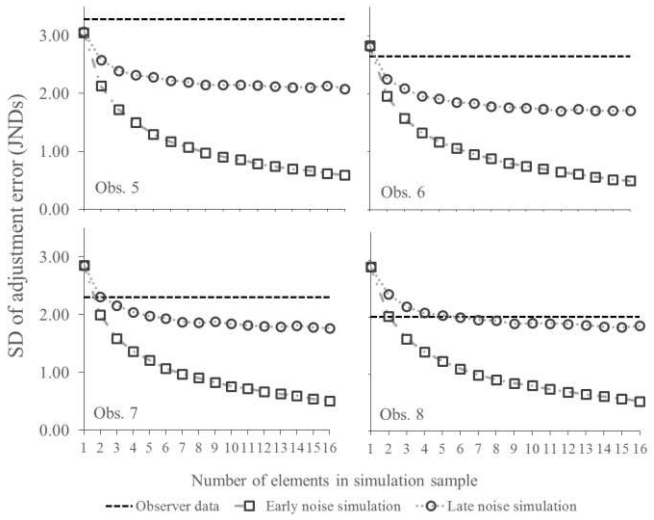

Fig. 5: Simulation results and actual data for four observers. The top left panel shows observer 5, who performs worse in the ensemble condition than would be predicted if they made their settings based on sampling just one ensemble element. Observer 6 (top right panel) and 7 (bottom left panel) perform at a level equivalent to sampling 1-2 ensemble elements (based on either late or early noise). Observer 8 (bottom right panel) performs at a level equivalent to sampling two elements in the early noise model, or between 5 and 6 in the late noise model.

4. DISCUSSION

This study had two main aims. The first was to compare the precision of settings of the mean hue of a rapidly-presented ensemble of different hues to the settings for a single hue.

The data show that, on average, observer settings tended to the mean color – settings peaked around the expected mean hue, with error distributed symmetrically either side indicating no bias or skew to the settings. The same pattern was found for the homogeneous condition. However the variance of settings was greater in the heterogeneous condition than in the homogeneous, indicating a difference in precision between these conditions. Settings of a mean hue were less precise than for single hue, indicating that reproducing the mean hue was subject to more error than reproducing a match for a single hue.

The second aim of this study was to establish whether a limited sub-sampling mechanism could explain the observers' performance on the hue averaging task. If the observer estimates of the average hue appear to be based on integrating the colors from more than three or four elements then the process would appear to exceed the limits of visual working memory [16], thus implying that an efficient, holistic mechanism may be responsible for the formation of ensemble representations of color.

The results of the simulation suggest that a sub-sampling mechanism where attention is devoted to encoding and averaging no more than two elements would be sufficient to provide estimates of the mean with precision equal to, or better than, most observers. In other words, the within-subject variance in responses around the mean is no better than would be expected from a limited-capacity, sub-sampling mechanism or strategy involving focused attention. Therefore, while the observers are clearly able to pick a mean hue following a rapidly-presented ensemble, and those selections converge on the true mean across trials, our data do not provide support for the proposal of a holistic hue averaging mechanism using distributed attention, or a mechanism with a capacity beyond the limits of visual short-term memory [16].

Our early noise simulation is most similar to that used by Haberman and Whitney [8], using measurements of error in setting a homogeneous ensemble as noise applied to each sampled element. We also included a late noise simulation, where the internal noise was applied after averaging had taken place. In reality, noise is present at both of these stages, but the measure of internal noise taken from the homogenous condition conflates these two sources of error, meaning that it would not be possible to include an accurate estimate of the noise at both stages in a single simulation. As can be seen from figure 5, with increasing set sizes, the late noise simulations asymptote at a higher level of error than the early noise simulations. This is because when noise is applied independently to each sampled element it is then subject to noise cancellation, where noise in the positive direction for one element is cancelled out by noise in the negative direction for another. In contrast there is no noise cancellation in the late noise model. As the sample size approaches the whole set, the late noise simulation predicts that precision will improve to equal that observed in the homogeneous condition, while the early noise simulation predicts that averaging precision will be higher than for the homogeneous ensembles. Therefore, if it was necessary to prefer one of the two, the late noise simulation would seem to make more realistic predictions than the early noise simulation. This difference notwithstanding, the conclusion with regard to observer performance is similar for both simulations – observers are sufficiently imprecise in their mean hue settings that the difference between simulations is trivial.

The simulation results should be interpreted with the assumption of fixed internal noise borne in mind. It has previously been suggested that internal noise (i.e., the precision of representations) may increase with larger samples [16]. Noisier representations for larger samples would reduce the advantages of taking a more representative sample. As a result, a simulation applying variable noise which increases with larger samples would return less precise mean estimations than a simulation applying a fixed amount of noise to mean representations. The magnitude of the difference between these possible models of internal noise is impossible to assess without an available estimate of how internal noise might change with sub-sample size. Although the assumption of fixed internal noise is conservative, in that it may over-estimate the precision of the null hypothesis of sub-sampling (the alternative being holistic averaging), the same assumption is made in other ensemble perception studies which do support the suggestion of holistic averaging for faces [8] and size [35]. Separating the effects of internal noise from sample size is a major theoretical challenge in understanding the possible mechanisms behind ensemble perception [45]. One way in which future research may address this would be to gather data which could enable the estimation of the level of internal noise which would be present for different sizes of sub-sample. This may be through cueing sub-sampling in an ensemble task, or simply gathering individual color estimations for more than one color, in a manner similar to the homogeneous condition of this study, but where the observer must attempt to retain and reproduce two (three, four, etc.) hues. Measured changes in internal noise that occur with increased sample size could then be included in further simulations. Such work is beyond the scope of the present study, however the approach of modelling the application of internal noise more realistically would be beneficial not just for understanding ensemble perception of color, but also other attributes.

In spite of the similarity in simulation assumptions and structure, our finding is somewhat at-odds with other simulations which also incorporate fixed internal noise, which have found sub-sampling models of ensemble perception of faces [8] and size [35] underperform on averaging precision relative to real observers with sample sizes fewer than seven elements. Notably, Haberman and Whitney [30] report that discrimination for the mean emotional expression from an ensemble was at least as good as discrimination for individual expressions – a trend not evident in our data, where precision for homogeneous ensembles (single colors) was better than for heterogeneous.

It is unlikely that a sub-sample would be taken from an ensemble at random, i.e. some elements may contribute disproportionately to the mean estimation [weighted-averaging, e.g., 46], or be more likely to be selected for a sub-sample. Attention cued to individual items has been shown to affect mean size estimates [21], and averaging of size over time has been shown to be biased towards looming phases, perhaps because these are more salient [47]. Variations in the salience of individual hues would not affect the overall central tendency (the peak of the distribution of selections) of mean adjustment in this experiment as the position of the ensemble elements rotates through every possible color in the stimulus circle (in a random order), however it could exaggerate deviations from the expected mean if those colors are assigned higher weights when averaging. The salience of each hue in the present study should be approximately equal in this study (stimuli are equated for luminance and equally distant from the white-point in DKL space). As DKL space is not scaled to equate saturation there is some residual variation in salience of hues around the hue circle. However this variation is gradual around the hue circle, meaning that local saturation differences (i.e. the difference between neighboring hues) is very subtle.

It is possible that hue is not as apt to be averaged using holistic sampling as size or faces. There are some differences between hue and other domains which may be responsible for differences in ensemble processing, but also several similarities. Unlike size, hue is a matter of qualitative experience, rather than magnitude. Saturation and lightness may both be described in terms of magnitude or intensity, so one color can be said to be "more saturated" or "lighter" than another. In contrast, hue is a circular dimension, requiring reference to color categories to describe relationships. Therefore, given highly distant exemplars (e.g., red and green) it may not be easy to imagine what the mean should look like. As the angle (e.g., in DKL space) between hues to be averaged approaches 180 degrees averaging could become increasingly difficult, or impossible, as the elements now represent opponent colors with qualitatively different sensations which do not blend into a meaningful average. Although the qualitative and circular nature of hue perception seems a plausible reason that averaging would be harder, these do not necessarily preclude hue from rapid, holistic averaging and there remain similarities with other domains. For example, hue averaging ability is reduced by increased ranges of hue in ensembles [13, 15], but this is also the case for size [29], and the stimuli in the present experiment were within the range at which mean selection from a 2AFC is reliable [13]. Hue is subject to categorization [e.g., 48], however our previous study has demonstrated that there is no effect of color categories on mean selection [13]. Face perception is also somewhat qualitative (in terms of emotions and identity), and is widely understood in terms of norm-based coding accounts, which rely on extraction of the mean [for a review see 49]. Like norm-based models of face coding, color perception is subject to white-point adaptation which supports color constancy [50]. In short, there do not appear to be particular features of hue perception which can be said to account for the high variance in average hue settings, particularly given the evidence for holistic ensemble perception in other domains.

It should be noted that just because sub-sampling could explain the results in this experiment, it does not necessarily imply that holistic averaging of hue cannot or does not take place [30]. Evidence from other domains suggests that averaging may be most reliable when the summary statistics are incidental to the main task. Summary statistics can have effects on response times and performance even when observers are not instructed to judge the mean or extract the gist at all. For example, response times for ensemble classification ("red" or "blue" average) can be reduced when a prime ensemble with the same variance is presented beforehand, even when the prime ensemble has a different mean color [24]. There are many other examples of tasks in which implicit processing of the summary statistics of sets appears to influence responses [2, 7, 14, 51-58]. It may be that instructing observers to consider and retain an average hue results in the use of a sub-sampling strategy, whereas observers may perform better relative to sub-sampling when the encoding of mean hue is an implicit part of the task.

The present study has shown that observers are able to reproduce the average hue following a rapidly-presented multi-hue ensemble, and their settings tend to the expected mean over many trials. However, these settings are distributed noisily around the mean, showing imprecision in the representation of mean hue. This imprecision is far greater than observed for reproduction of a single hue setting for single hues presented in the same way. The ideal observer simulation suggests that a sub-sampling mechanism integrating just two items from the set would outperform most observers on the task. This implies that holistic averaging of the whole set may not occur for ensembles of hue and that our percept of color gist may be biased towards the particular sub-sampled colors of an ensemble. Further research is needed to clarify what factors drive certain elements to be included in the sub-sample (e.g., salience, spatial position/fixation), and whether holistic ensemble representations of hue can be promoted during tasks where color summary statistics are not the focus of attention. It would also be informative to encourage sub-sampling strategies (e.g., random, spatial, or pre-cued elements) in participants as a real-observer analogue to the ideal-observer simulation reported here, in order to assess whether the simulation could be a realistic model for how the task is done. The present study suggests that average hue may not be a summary statistic which is automatically and efficiently encoded by observers, and that the perception of a rich world of color may be biased by the hue of individual elements in a scene.

Funding. The research was supported by an Economic and Social Research Council grant to JM (ES/J500173/1), and a European Research Council grant to AF (CATEGORIES: 283605).

Acknowledgment. We would like to thank the Colour Group GB who supported the presentation of this work at the 23rd Symposium of the International Colour Vision Society through the WD Wright award (to JM).

References

1. J. Haberman and D. Whitney, "Ensemble Perception: Summarizing the scene and broadening the limits of visual processing," in *From Perception to Consciousness: Searching with Anne Treisman,* J. Wolfe and L. Robertson, eds. (Oxford University Press, Oxford, 2012), pp. 339-349.
2. D. Ariely, "Seeing sets: Representation by statistical properties," Psychol Sci **12**, 157-162 (2001).
3. S. C. Chong and A. Treisman, "Representation of statistical properties," Vision Res **43**, 393-404 (2003).
4. J. E. Corbett and C. Oriet, "The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation," Acta Psychol **138**, 289-301 (2011).
5. A. P. Marchant and J. W. de Fockert, "Priming by the mean representation of a set," Q J Exp Psychol **62**, 1889-1895 (2009).
6. S. C. Dakin, "Information limit on the spatial integration of local orientation signals," J Opt Soc Am A Opt Image Sci Vis **18**, 1016-1026 (2001).
7. L. Parkes, J. Lund, A. Angelucci, J. A. Solomon, and M. Morgan, "Compulsory averaging of crowded orientation signals in human vision," Nat Neurosci **4**, 739-744 (2001).
8. J. Haberman and D. Whitney, "The visual system discounts emotional deviants when extracting average expression," Atten Percept Psycho **72**, 1825-1838 (2010).

9. A. Y. Leib, A. M. Puri, J. Fischer, S. Bentin, D. Whitney, and L. Robertson, "Crowd perception in prosopagnosia," Neuropsychologia **50**, 1698-1707 (2012).
10. J. W. de Fockert and C. Wolfenstein, "Rapid extraction of mean identity from sets of faces," Q J Exp Psychol **62**, 1716-1722 (2009).
11. C. Fiorentini, L. Gray, G. Rhodes, L. Jeffery, and E. Pellicano, "Reduced face identity aftereffects in relatives of children with autism," Neuropsychologia **50**, 2926-2932 (2012).
12. A. Y. Leib, J. Fischer, Y. Liu, S. Qiu, L. Robertson, and D. Whitney, "Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity," J Vision **14**(2014).
13. J. Maule and A. Franklin, "Effects of ensemble complexity and perceptual similarity on rapid averaging of hue," J Vision **15**, 1-18 (2015).
14. J. Maule, C. Witzel, and A. Franklin, "Getting the gist of multiple hues: metric and categorical effects on ensemble perception of hue," J Opt Soc Am A **31**, A93-A102 (2014).
15. J. Webster, P. Kay, and M. A. Webster, "Perceiving the average hue of color arrays," J Opt Soc Am A **31**, A283-A292 (2014).
16. G. A. Alvarez, "Representing multiple objects as an ensemble enhances visual cognition," Trends Cogn Sci **15**, 122-131 (2011).
17. A. Treisman, "How the deployment of attention determines what we see," Vis Cogn **14**, 411-443 (2006).
18. A. P. Marchant, D. J. Simons, and J. W. de Fockert, "Ensemble representations: Effects of set size and item heterogeneity on average size perception," Acta Psychol **142**, 245-250 (2013).
19. K. Myczek and D. J. Simons, "Better than average: Alternatives to statistical summary representations for rapid judgments of average size," Percept Psychophys **70**, 772-788 (2008).
20. D. J. Simons and K. Myczek, "Average size perception and the allure of a new mechanism," Percept Psychophys **70**, 1335-1336 (2008).
21. J. W. de Fockert and A. P. Marchant, "Attention modulates set representation by statistical properties," Percept Psychophys **70**, 789-794 (2008).
22. R. O. Brown and D. I. A. MacLeod, "Color appearance depends on the variance of surround colors," Curr Biol **7**, 844-849 (1997).
23. S. Ratnasingam and B. L. Anderson, "The role of chromatic variance in modulating color appearance," J Vision **15**(2015).
24. E. Michael, V. de Gardelle, and C. Summerfield, "Priming by the variability of visual information," P Natl Acad Sci USA **111**, 7873-7878 (2014).
25. G. Buchsbaum, "A Spatial Processor Model for Object Color-Perception," J Franklin I **310**, 1-26 (1980).
26. M. Olkkonen, P. F. McCarthy, and S. R. Allred, "The central tendency bias in color perception: effects of internal and external noise," J Vision **14**(2014).
27. S. C. Chong and A. Treisman, "Statistical processing: computing the average size in perceptual groups," Vision Res **45**, 891-900 (2005).
28. N. Robitaille and I. M. Harris, "When more is less: Extraction of summary statistics benefits from larger sets," J Vision **11**(2011).
29. I. S. Utochkin and N. A. Tiurina, "Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert," Acta Psychol **146**, 7-18 (2014).
30. J. Haberman and D. Whitney, "Seeing the Mean: Ensemble Coding for Sets of Faces," J Exp Psychol Human **35**, 718-734 (2009).
31. I. Kuriki, "Testing the possibility of average-color perception from multi-colored patterns," Opt Rev **11**, 249-257 (2004).
32. S. Sunaga and Y. Yamashita, "Global color impressions of multicolored textured patterns with equal unique hue elements," Color Res Appl **32**, 267-277 (2007).
33. D. Ariely, "Better than average? When can we say that subsampling of items is better than statistical summary representations?," Percept Psychophys **70**, 1325-1326 (2008).
34. S. C. Chong and A. Treisman, "Attentional spread in the statistical processing of visual displays," Percept Psychophys **67**, 1-13 (2005).
35. H. Y. Im and J. Halberda, "The effects of sampling and internal noise on the representation of ensemble average size," Atten Percept Psycho **75**, 278-286 (2013).
36. S. C. Chong, S. J. Joo, T. A. Emmanouil, and A. Treisman, "Statistical processing: not so implausible after all," Percept Psychophys **70**, 1327-1334; discussion 1335-1326 (2008).
37. M. Attarha, C. M. Moore, and S. P. Vecera, "Summary Statistics of Size: Fixed Processing Capacity for Multiple Ensembles but Unlimited Processing Capacity for Single Ensembles," J Exp Psychol Human **40**, 1440-1449 (2014).
38. S. Baijal, C. Nakatani, C. van Leeuwen, and N. Srinivasan, "Processing statistics: An examination of focused and distributed attention using event related potentials," Vision Res **85**, 20-25 (2013).
39. H. Y. Im and S. C. Chong, "Mean size as a unit of visual working memory," Perception **43**, 663-676 (2014).
40. S. Ishihara, *Ishihara's test chart for colour deficiency* (Kanehara Trading INC, Tokyo, 1973).
41. R. Fletcher, *The City University Colour Vision Test, 2nd Edition.* (Keeler, London, 1980).
42. A. M. Derrington, J. Krauskopf, and P. Lennie, "Chromatic Mechanisms in Lateral Geniculate-Nucleus of Macaque," J Physiol-London **357**, 241-265 (1984).
43. J. Krauskopf, D. R. Williams, and D. W. Heeley, "Cardinal Directions of Color Space," Vision Res **22**, 1123-1131 (1982).
44. C. Witzel and K. R. Gegenfurtner, "Categorical sensitivity to color differences," J Vision **13**(2013).
45. Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. Vision Research, 83, 25-39. doi: 10.1016/j.visres.2013.02.018
46. V. de Gardelle and C. Summerfield, "Robust averaging during perceptual judgment," P Natl Acad Sci USA **108**, 13341-13346 (2011).
47. A. R. Albrecht and B. J. Scholl, "Perceptually Averaging in a Continuous Visual World: Extracting Statistical Summary Representations Over Time," Psychol Sci **21**, 560-567 (2010).
48. C. M. Bird, S. C. Berens, A. J. Horner, and A. Franklin, "Categorical encoding of color in the brain," P Natl Acad Sci USA **111**, 4590-4595 (2014).
49. D. Y. Tsao and W. A. Freiwald, "What's so special about the average face?," Trends Cogn Sci **10**, 391-393 (2006).
50. M. A. Webster and D. I. MacLeod, "Visual adaptation and face perception," Philosophical transactions of the Royal Society of London: B **366**, 1702-1725 (2011).
51. J. Allik, M. Toom, A. Raidvee, K. Averin, and K. Kreegipuu, "Obligatory averaging in mean size perception," Vision Res **101**, 34-40 (2014).
52. G. A. Alvarez and A. Oliva, "The representation of simple ensemble visual features outside the focus of attention," Psychol Sci **19**, 392-398 (2008).
53. G. A. Alvarez and A. Oliva, "Spatial ensemble statistics are efficient codes that can be represented with reduced attention," P Natl Acad Sci USA **106**, 7345-7350 (2009).
54. C. Oriet and J. Brand, "Size averaging of irrelevant stimuli cannot be prevented," Vision Res **79**, 8-16 (2013).
55. J. E. Corbett and D. Melcher, "Stable Statistical Representations Facilitate Visual Search," J Exp Psychol Human **40**, 1915-1925 (2014).
56. J. E. Corbett and D. Melcher, "Characterizing ensemble statistics: mean size is represented across multiple frames of reference," Atten Percept Psycho **76**, 746-758 (2014).
57. J. E. Corbett, N. Wurnitsch, A. Schwartz, and D. Whitney, "An aftereffect of adaptation to mean size," Vis Cogn **20**, 211-231 (2012).
58. L. Lanzoni, D. Melcher, G. Miceli, and J. E. Corbett, "Global statistical regularities modulate the speed of visual search in patients with focal attentional deficits," Front Psychol **5**, 1-12 (2014).