Variability leads to overestimation of mean summaries

As Published: https://doi.org/10.3758/s13414-021-02269-2

Publisher: Springer US

Persistent URL: https://hdl.handle.net/1721.1/136762

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Variability Leads to Overestimation of Mean Summaries

Cite this article as: Yelda Semizer, Aysecan Boduroglu, Variability Leads to Overestimation of Mean Summaries, *Attention, Perception, &amp; Psychophysics*, doi: 10.3758/s13414-021-02269-2

Variability Leads to Overestimation of Mean Summaries

Yelda Semizer[1] & Aysecan Boduroglu[2]

[1]Massachusetts Institute of Technology

[2]Bogazici University

Word Count: 6372

Yelda Semizer

New Jersey Institute of Technology

Department of Humanities

University Heights, Newark, NJ 07102

yelda.semizer@njit.edu

Abstract

Research on ensemble perception has shown that people can extract both mean and variance information but much less is understand how these two different types of summaries interact with one another. There has been some research arguing that people are more erroneous in extracting the mean of displays that have greater variability. In all three experiments, we manipulated the variability in the displays. Participants reported the mean size of a set of circles (Experiment 1) and mean length of horizontally placed (Experiment 2A) and randomly oriented lines (Experiment 2B). In all experiments, we found that mean size estimations were more erroneous for higher than smaller variance displays. More critically, there was a tendency to overestimate the mean, driven by variance in both task-relevant and task-irrelevant features. We discuss these findings in relation to limitations in concurrent summarization ability and outlier discounting in ensemble perception.

Keywords: summary statistics, ensemble perception, variance, overestimation

Variability Leads to Overestimation of Mean Summaries

Since the seminal paper of Ariely (2001) on the representation of statistical properties, there have been numerous studies on ensemble perception, demonstrating that viewers can efficiently extract the mean of numerous static and dynamic perceptual features (for reviews see Alvarez, 2011, Whitney & Yamanashi-Leib, 2018), supported by multiple, feature specific mechanisms (Haberman et al. 2015; Yörük & Boduroglu, 2020). More recent work has also concluded that the visual system implicitly extracts variance and range of various lower-level (e.g. Cant & Xu, 2020; Khayat & Hochstein, 2018; Maule & Franklin, 2020; Morgan, Chubb & Solomon, 2008; Tokita, Ueda & Ishiguchi, 2016; Tong, Ji, Chen & Fu, 2015; Ward, Bear & Scholl, 2016) and higher-level features (e.g. Haberman, Lee & Whitney, 2015; Phillips, Slepian & Hugher, 2018). These findings have led to the question of whether separate systems are responsible for extracting mean and variance information. To date, most studies have shown independence of errors in mean and variance judgments, arguing that mean and variance are extracted via separate systems (e.g. Khvostoz & Utochkin, 2019; Utochkin & Vostrikov, 2017; Yang, Tokita & Ishiguchi, 2018). Others have supported this claim by demonstrating that there were no dual-task costs associated with concurrently extracting mean and variance (Khvostov & Utochkin, 2019) and that the variance aftereffect was not affected by mean changes in color or orientation (e.g. Maule & Franklin, 2020; Norman, Heywood & Kentridge, 2015).

While the individual differences findings supporting independent systems for mean and variance extraction have not been empirically challenged, there have been a number of studies arguing that mean and variance judgments may nevertheless interact.

For instance, Corbett and colleagues demonstrated that increased variance in the adapting set reduced the magnitude of the aftereffect in mean estimates systematically across low, medium and high variance conditions (Corbett, Wurnitsch, Schwartz & Whitney, 2012). Im & Halberda (2013) reported that in a forced choice task in which viewers had to determine the larger of two sequentially presented sets, the discrimination threshold increased as the item size variability increased. Michael, de Gardelle & Summerfield (2014) demonstrated that variance match between a prime and a target display resulted in the facilitation of average color or shape-based classifications. Recently, Kanaya, Hayashi & Whitney (2018) reported of an amplification effect-referring to an increase in perceived subjective equality as a function of increased variance for temporal frequency and size summaries. Finally, Jeong & Chong (2020) demonstrated a bidirectional relationship between mean and variance, showing that both perceived orientation variance and perceived mean orientation impact one another.

These findings on the interactions between mean and variance judgments do not necessarily challenge the claim of independent mean and variance extraction mechanisms. Even if these summaries are independently extracted, they may still interact in later stages of information processing, impacting judgments (Jeong & Chong, 2020, Utochkin & Vostrikov, 2017). Alternatively, variance in displays directly impact extraction of an ensemble summary by directly influencing which items are subsampled. These possibilities highlight the need to further understand how mean and variance judgments interact. In this study, we specifically wanted to determine whether variance of task-relevant and task-irrelevant features systematically impact mean judgments.

Chong & Treisman (2003) was one of the earliest studies to indirectly address the issue of how variance and mean information interact during ensemble perception. They were actually interested in determining whether the discrimination thresholds for heterogeneously sized and homogeneously sized sets were similar. Their third experiment which was designed to rule out an area-based strategy, compared discrimination thresholds across uniform, normal, twin-peaked and homogeneous distributions. Relevant for our purposes, while the threshold difference for differentiating mean size of homogeneously-sized displays was the lowest, it was highest for comparisons involving two-peaked (highest variance) and homogeneous sets (for a similar finding see Utochkin & Tiurina, 2014). This and subsequent findings suggest that increased variance results in greater imprecision in mean extraction. This may have to do with how in high variable displays, on average, each item is further away from the mean compared to that in lower variance displays. Thus, in higher variance displays, subsampling any subset of the items may to lead to greater distortion during mean extraction than subsampling in lower variance displays (Myczek & Simons, 2008). This idea was more directly tested by Kanaya et al. (2018) who presented participants with single item, homogeneous, low variance, and high variance displays. For both temporal frequency and size summaries, they found that higher variance led to greater point of subjective equality (PSE); this amplification effect was not due to increases in set size and there was no difference in PSEs between single and homogeneous conditions. However, greater error in mean estimates of higher variance displays does not necessitate a subsampling based explanation of mean extraction. It is possible that the mean may be extracted concurrently with the processing of a particular item; in a higher variance display, that item is more likely to have a featural value more distant from the mean. Consequently, any pull from this one item may lead to greater distortion of the mean at the final decision stage.

In this study, our goal was twofold. First, we wanted to determine whether we would be able to replicate the finding of greater error in higher variance as opposed to lower variance displays. More critically, we were interested in whether the imprecision reported in mean summaries in higher as opposed to lower variance displays was due to any systematic bias to overestimate the mean. To determine the nature of errors in mean estimation under high variance conditions, we carried out three experiments in which we presented participants with single item sets, homogeneously sized sets as well as lower and higher variance sets consisting of circles (Experiment 1) and lines (Experiment 2A and 2B). Specifically, in both Experiment 2A and 2B, we probed mean length; but, in Experiment 2B we also randomly varied the orientation of all the lines. While we expect task-relevant variance in all three experiments to be associated with increased error, based on Kanaya et al. (2018), it is possible that variance would lead to a systematic bias to overestimate mean size/length. However, if variance in a task-irrelevant dimension (e.g. orientation in Experiment 2B) impacts summarizing of the task-relevant dimension, it is unlikely that observed biases are only due to which items are sampled during summarization. This brings the possibility of concurrently summarized features interacting with one another shaping responses.

Because we were particularly interested in over estimation tendencies in mean extraction, we primarily focused on the ratio of the reported to the actual target value. If participants were overestimating the mean, then we would expect this value to be greater than 1. On the other hand, underestimation of mean should lead to values smaller than 1. We also carried out a complementary set of analyses, to ensure that we replicated earlier findings of increased error in greater variance displays. To hint at our results, we found that for the mean estimation under increased variance conditions were more erroneous. More critically, in higher variance conditions, there was a tendency to overestimate the mean. Variance in a task-irrelevant feature further degraded performance, however this did not further add to the overestimation tendency.

Experiment 1

Method

Participants

Thirty-six undergraduates taking the introductory psychology courses at Bogazici University participated in Experiment 1 in exchange for course credit. All had normal or corrected-to-normal vision. This study and subsequent studies were approved by the ethics board at Bogazici University. Sample size for this and subsequent studies were determined based on similar studies on ensemble perception (e.g. Brady & Alvarez, 2011)[1]. Post-hoc power analyses for variance and mean size effects revealed a statistical power of .94 or greater in this and for all subsequent experiments.

Apparatus

The stimuli were presented on a 17-in. CRT computer monitor with a screen resolution set to 640 x 480 pixels and a viewing distance of approximately 57 cm. One degree of visual angle was approximately 20 pixels. Stimulus generation, timing operations, and data collection were controlled by a PC running E-prime 1.2 software (Psychology Research Tools, Inc).

[1] While Brady & Alvarez (2011) had approximately 20 participants in each one of their experiment, in Experiment 1, we chose to have a slightly larger sample to allow us to explore the link between WM capacity and ensemble perception. Since those data are not directly linked to the main question of interest addressed in this paper, we do not discuss them further.

Stimuli

Stimuli were white circles with various sizes displayed on a gray background. We created displays with a set of nine circles and manipulated the mean and the variance of the sets to create four different conditions: *smaller mean / lower variance*, *smaller mean / higher variance*, *larger mean / lower variance*, and *larger mean / higher variance*. In addition, we created displays with nine same-size circles (*homogeneous* displays) and displays with a single circle (*single-circle* displays). For both *homogeneous* and *single-circle* conditions, the mean sizes of the circles were identical to those in the *lower* and *higher* variance conditions.

The diameter of each individual circle was randomly selected from a uniform distribution. To equate the effects of possible perceptual outliers on mean size estimations among sets with different mean sizes, we used a large range to draw higher variance sets (0.75°, 4.75°). Because of the lower probability of generating sets with lower variance in large ranges, which could meet the specified constraints described next, we used smaller intervals while generating lower variance sets for smaller (0.75°, 2.75°) and larger (2.75°, 4.75°) mean conditions. Each randomly generated set has to satisfy a set of criteria to be included in the experimental stimuli: None of the individual circles in the lower and higher variance conditions were the same size as the mean of the set. Additionally, there was no repetition of size within a set of stimuli for a given trial. Only sets that met the specified mean and variance criteria were selected.

The location of circles within a set was randomly determined in an invisible 5 x 4 grid; with a jitter of +/- 10 pixels to minimize regularities. As a consequence, the center of the screen remained empty. No more than two circles were placed adjacent to each other and there were no more than three circles either in a row or a column. In order to avoid the edge of the screen occluding stimuli on a given trial, no circles were presented within 50 pixels (2.5º) of the edge of the display.

Procedure

Each trial started with a fixation cross presented at the center of the screen for 500 ms (see Figure 1). Following fixation, a stimulus display (either an array of nine circles, or a single circle) was shown for 1000 ms followed by a gray screen for 1000 ms, after which a single randomly sized response circle appeared at the center of the screen[2]. Participants were required to resize the response circle to the mean size of the previous set via key press. Each key press changed the radius of the circle by 1 pixel (0.05°). Then, they pressed another key to record their response.

To determine the size of the response probe, we first chose two random numbers between 13 and 17 (in pixels) and these two numbers were subtracted from and added to the mean size of the display, yielding upper and lower bounds for the response probe radius. From this range, the radius of the response circle was randomly determined. This ensured that the response probe varied across trials and conditions such that there was no set difference between the probe radius and the target response.

Conditions were presented in a mixed order. There were 20 trials in each condition, resulting in a total of 120 experimental trials. In addition, participants completed 12 practice trials at the start of the experiment to familiarize themselves with the task. During the practice trials, participants were given feedback after responding as to the accuracy of their response; they were shown a circle with the correct mean size

[2] We decided on these parameters based on Brady & Alvarez (2011) that also provided a systematic exploration of biases in responses. We specifically discuss possible implications of the delay interval on the overall pattern of results in the general discussion.

displayed above the response circle. No feedback was provided during the experimental trials. The experiment took approximately 1-hour to complete.

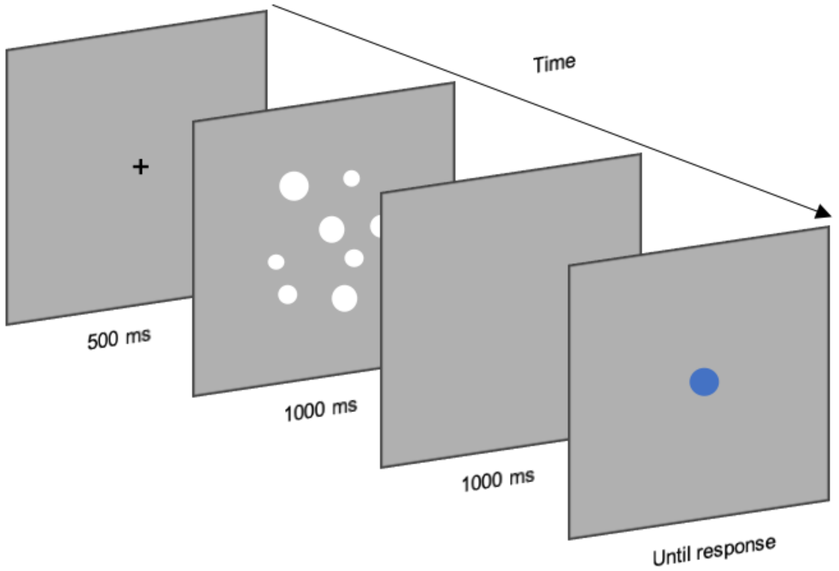

Figure 1. Timeline for the mean size estimation task in Experiment 1. A fixation cross was presented at the center of the screen for 500 ms. Following fixation, a stimulus display (either an array of nine circles, or a single circle) was shown for 1000 ms followed by a gray screen for 1000 ms, after which a single randomly sized response circle appeared at the center of the screen. Participants were required to resize the response circle to the mean size of the previous set via key press.

Results & Discussion

In all analyses, the primary measure was the estimated radius of the circles in terms of pixels. For each trial, we calculated both the absolute error by computing the absolute value of the difference between the estimated size and the correct size and the ratio of the estimated size over the correct size.

Absolute Error Analyses

To examine whether variance affected mean size estimation, we conducted a 2 (variance type: lower or higher) X 2 (mean size: smaller or larger) within-subjects ANOVA. As expected, participants made more errors in the higher variance condition (*M* = 4.68, *SD* = 1.40) compared to the lower variance condition (*M* = 3.37, *SD* = .92), *F* (1, 35) = 49.91, $p < .001$, $\eta^2 = .59$. Error in the smaller mean size condition (*M* = 3.87, *SD* = 1.05) was not different from the error in the larger mean size condition (*M* = 4.19, *SD* = 1.27), *F* (1, 35) = 2.79, $p = .10$, $\eta^2 = .07$. The interaction effect did not reach significance, *F* (1, 35) = .98, $p = .33$, $\eta^2 = .03$ (see Figure 2b). Our comparison of the error in the single-circle and the homogeneous display conditions replicated general findings in the ensemble literature (e.g. Chong & Treisman, 2003). A 2 (display type: single-circle or homogeneous) X 2 (mean size: smaller or larger) within-subjects ANOVA revealed that participants were equally accurate in the single-circle condition (*M* = 2.42, *SD* = .85) and the homogeneous condition (*M* = 2.63, *SD* = .93), *F* (1, 35) = 2.55, $p = .12$, $\eta^2 = .07$. There was a main effect of mean size, *F* (1, 35) = 42.71, $p< .001$, $\eta^2 = .55$, with participants making more error in the larger mean size (*M* = 3.00, *SD* = .99) compared to the smaller mean size condition (*M* = 2.05, *SD* = .79). There was no interaction, *F* (1, 35) = .87, $p = .36$, $\eta^2 = .02$ (see Figure 2a).

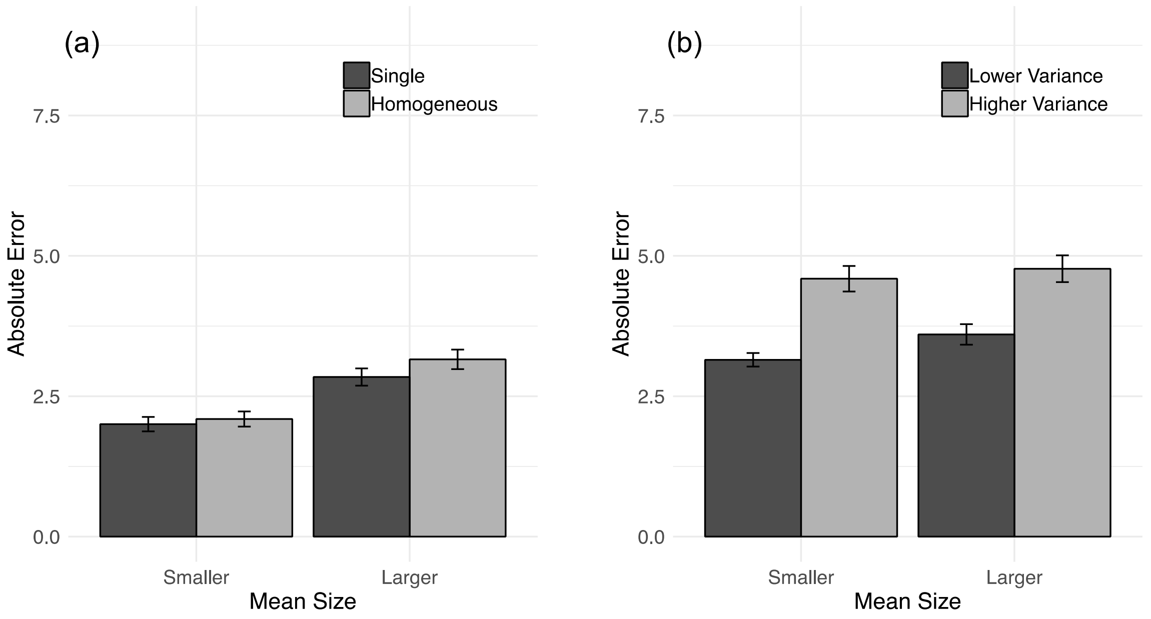

Figure 2. Absolute error in pixels for each condition in Experiment 1. The left-hand side (Figure 2A) presents the comparison between single and homogeneous conditions and the right-hand side (Figure 2B) presents the data for the lower and higher variance conditions. Error bars indicate standard error.

Estimation Bias Analyses

To test our primary question of interest as to whether there was a systematic bias in these estimation errors, we calculated estimation bias scores by computing the ratio of the estimated size over the correct size. An estimation bias score greater than 1 would indicate an overestimation tendency while an estimation bias score less than 1 would indicate an underestimation tendency. Because there was no difference in error across smaller and larger mean conditions, we collapsed across mean size. When compared to 1 by a series of single *t*-tests, only the higher variance condition (*M* = 1.055, *SD* = .14) was significantly different than the cut-off, suggesting a tendency to overestimate of mean size, *t* (35) = 3.62, *p* = .001, *d* = .37 (see Figure 3). For all the other conditions, there was not a significant bias in either direction, all *d*s <.16.[3]

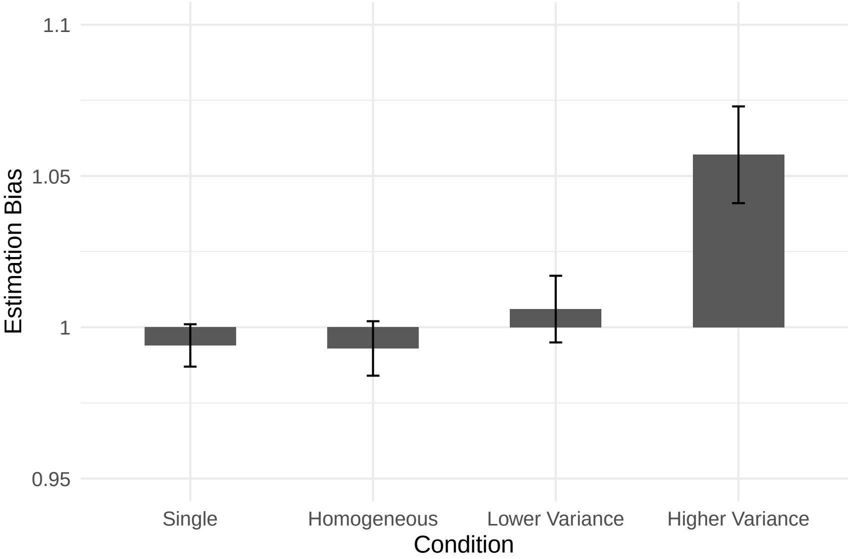

Figure 3. Estimation bias for each condition in Experiment 1. Error bars indicate standard error.

We wanted to ensure that this pattern of overestimation in the higher variance displays was not driven by certain display characteristics that may have uniquely impacted performance in this condition. In the lower and higher variance displays, out of the 9 circles, naturally some circles were larger and some were smaller in size than the set mean. If there were more trials in which the sets were made up of circles with sizes larger than the mean and if participants were inadvertently reporting the size of any one of these circles, then the above described pattern of overestimation could have been

[3] We also calculated real error in each condition by subtracting participants' response from the size for each trial, taking into account directionality of errors. Comparison of the average error against 0 (criteria of no error) revealed an overestimation of mean size only in the higher variance condition, *t* (35) = 1.22, *p* = .005. For the remaining experiments, analyses based on real error were similar to results based on estimation bias scores. For purposes of brevity, we only report the estimation bias scores.

observed for reasons other than biases in mean extraction. However, this was unlikely because the proportion of trials in which there were more circles with sizes larger than the set mean was similar across conditions (18/40 and 19/40 for lower and higher variance trials, respectively).

Experiment 2

In Experiment 1, we demonstrated that people made larger errors in estimating the mean of displays that had higher variability. Furthermore, in the higher variability condition, participants were more likely to overestimate the mean size. In Experiment 2 we wanted to replicate this pattern and eliminate the possibility that the pattern observed in Experiment 1 was an artifact of the circular nature of the stimuli. Therefore, in Experiment 2 we replicated the same procedure with one exception: participants were to extract the mean length of line stimuli, instead of the mean size of circles. These lines were presented either horizontally (Experiment 2A) or at random orientations (Experiment 2B); the latter manipulation served two functions. One, it prevented participants from relying on a strategy where they could more easily compute the total length of lines and divide that by the number of lines to obtain the mean. While this is a possibility especially given the longer durations of stimuli presentation in our study, there are a number of studies arguing that this is not how participants estimate mean size/length (e.g. Raidvee, Toom, Averin & Allik, 2020; Utochkin & Vostrikov, 2017). More critically, this manipulation of presenting lines at random orientations allowed us to see the impact of how variance in a task-irrelevant dimension impacts mean precision. Many items in the natural world are multidimensional and it is no surprise that the visual system can concurrently summarize different dimensions (e.g. Attarha & Moore, 2015;

Poltoratski & Xu, 2013; Utochkin & Vostrikov, 2017; Yörük & Boduroglu, 2020), even creating integrated summaries (e.g. Boduroglu & Yıldırım, 2020; Rodriguez-Cintron, Wright, Chubb & Sperling, 2019). It is likely that variance in task-irrelevant dimensions are likely to be coded by the system. We specifically ask whether this additional information impacts precision of mean estimates. Kanaya et al. (2018) argues that variance in the task-relevant dimension influence how the visual system samples the items in the array, ensuring items in the higher edges of the distribution to be disproportionately represented in the summary, leading to an amplification of the summary. If separate mechanisms concurrently summarize different featural properties like length and orientation, it would seem that the characteristics of the task-irrelevant dimension—i.e. orientation in this case, should have no impact on what is sampled as mean length is extracted. Thus, if these featural summaries are independently carried out, we would expect Experiment 2B to replicate Experiment 2A. On the other hand, if participants extract the mean and variance of the orientation of the lines even though it is task-irrelevant, this could possibly incur costs not during ensemble coding but subsequently, at response generation, by increasing interference between available summary representations. This would result in an increase in error, without creating a systematic bias to over or under estimate.

Method

Participants

Bogazici University undergraduates taking introductory psychology courses participated in Experiment 2A (*N* = 20) and 2B (*N* = 21) in exchange for course credit. All had normal or corrected-to-normal vision. All provided informed consent.

Apparatus

Apparatus were the same as in Experiment 1.

Stimuli

Stimuli were displays consisting of lines. These lines were the diameters of the circles presented in Experiment 1 and they were presented either in a horizontally aligned fashion (Experiment 2A) or at random orientations (Experiment 2B). Figure 4 shows example displays.

In experiment 2B, randomly oriented stimulus sets were created by holding one edge of the lines constant while rotating each to one of 18 possible locations, each 10° apart. To prevent perceptual grouping, no more than two lines with the same orientation were presented within the same set. Also, lines within a set were never presented horizontally.

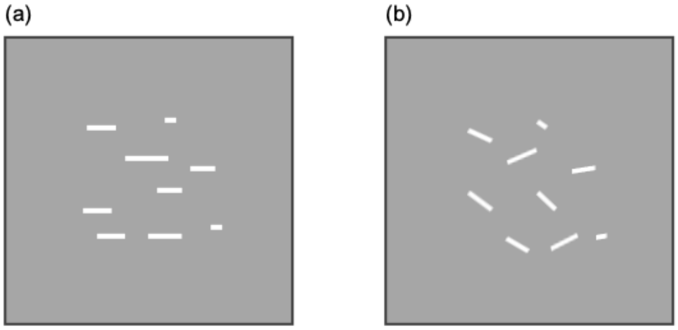

Figure 4. Example displays from Experiment 2A (on the left) and 2B (on the right). Displays consisted of only the diameters of the circles, either in horizontal (a) or in random (b) orientation.

Procedure

Procedure was the same as in Experiment 1.

Results & Discussion

Experiment 2A

Absolute error analysis. To examine whether variance affected mean length estimation, we carried out a 2 (variance type: lower or higher) X 2 (mean length: smaller or larger) within-subjects ANOVA. As in Experiment 1, there was a main effect of variance, F (1, 19) = 17.91, $p < .001$, $\eta^2 = .49$. Participants made more errors in the higher variance condition (M = 4.76, SD = 1.42) compared to the lower variance condition (M = 3.72, SD = .98). There was no main effect of mean length, F (1, 19) = 2.07, $p = .17$, $\eta^2 = .10$, and there was no interaction between variance type and mean length, F (1, 19) = .80, $p = .33$, $\eta^2 = .04$ (see Figure 5b). We also ran a within-subjects ANOVA to compare error in single-line and homogeneous-line displays in each mean length condition. As in Experiment 1, there was no main effect of display type, F (1, 19) = 2.61, $p = .12$, $\eta^2 = .12$, suggesting that errors in the single-line condition (M = 2.68, SD = .73) and the homogeneous condition (M = 3.02, SD = 1.12) were not statistically different. As in Experiment 1, participants made more errors in the larger mean length (M = 3.36, SD = 1.03) compared to the smaller mean length condition (M = 2.33, SD = 1.64), F (1, 19) = 27.98, $p < .001$, $\eta^2 = .60$. The interaction effects did not reach significance, F (1, 19) = 1.58, $p = .23$, $\eta^2 = .08$ (see Figure 5a).

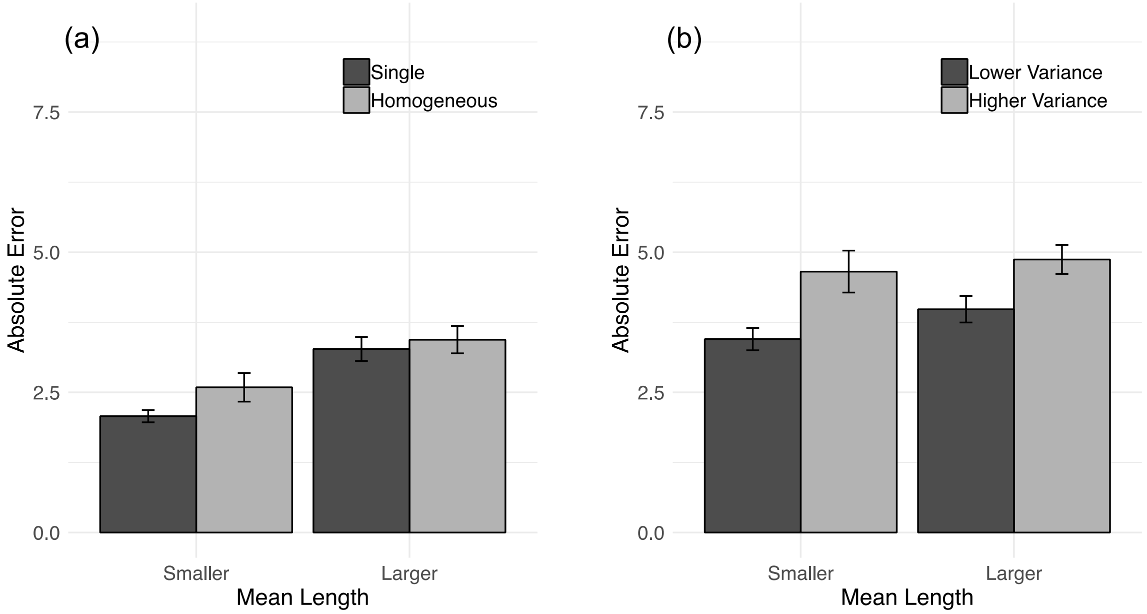

Figure 5. Absolute error in pixels in each condition in Experiment 2A. The left-hand side (Figure 5a) presents the comparison between single and homogeneous conditions and the right-hand side (Figure 5b) presents the data for the lower and higher variance conditions. Error bars indicate standard error.

Estimation bias analysis. A series of single *t*-tests revealed that only in the higher variance condition was the estimation bias score significantly different than 1, *t (*19) = 2.80, *p* = .012, *d*=.54. Similar to Experiment 1, there was an overestimation of the mean length in the higher variance condition (see Figure 6).

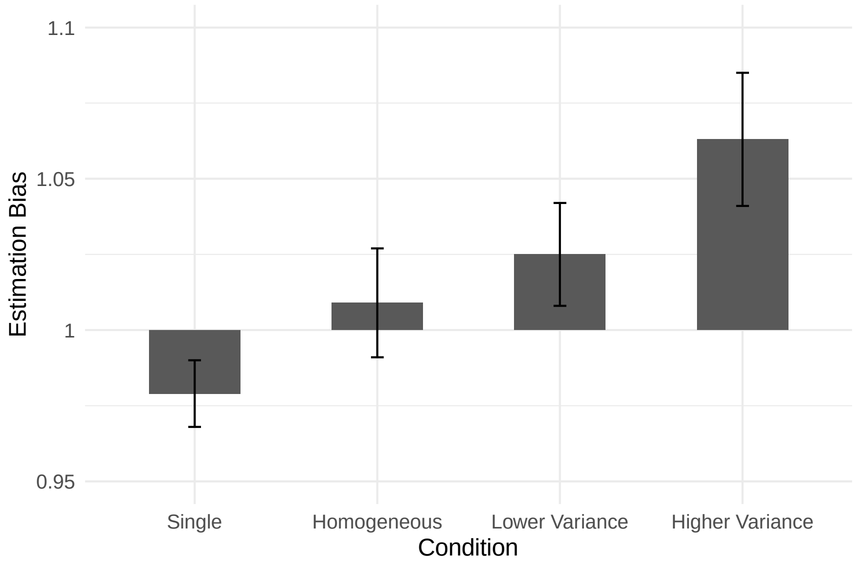

Figure 6. Estimation bias in each condition in Experiment 2a. Error bars indicate standard error.

Experiment 2B

Absolute error analysis. Analysis on the effects of variance on mean estimations revealed similar results to those of the previous experiments. Participants were more erroneous in the higher variance condition (*M* = 5.12, *SD* = 1.29) compared to the lower variance condition (*M* = 4.39, *SD* = .97), *F* (1, 21) = 13.35, *p* = .002, η^2= .40. There was no difference between the smaller (*M* = 4.54, *SD* = 1.22) and larger mean conditions (*M* = 4.96, *SD* = 1.05), *F* (1, 21) = 2.35, *p* = .14, η^2= .11. There was no interaction, *F* (1, 21) = .61, *p* = .45, η^2= .03 (see Figure 7b). The comparison between the single and the homogeneous sets, for the first time revealed a main effect of trial type, *F (*1,21) = 8.04, *p* = .01, η^2= .29; with participants making more error in the homogeneous condition (*M* = 3.89, *SD* = 1.20) compared to the single-line condition (*M* = 3.35, *SD* = .81), suggesting that variance in the task-irrelevant dimension (i.e. orientation) was interfering with mean extraction. Similar to previous experiments, participants made more errors in the larger mean length condition (*M* = 4.05, *SD* = 1.15) compared to the smaller mean length condition (*M* = 3.18, *SD*= .85), *F* (1, 21) = 26.41, *p*< .001, η^2= .57. There was no interaction between size and condition, *F* (1, 21) = 4.06, *p* = .06, η^2= .17 (see Figure 7a).

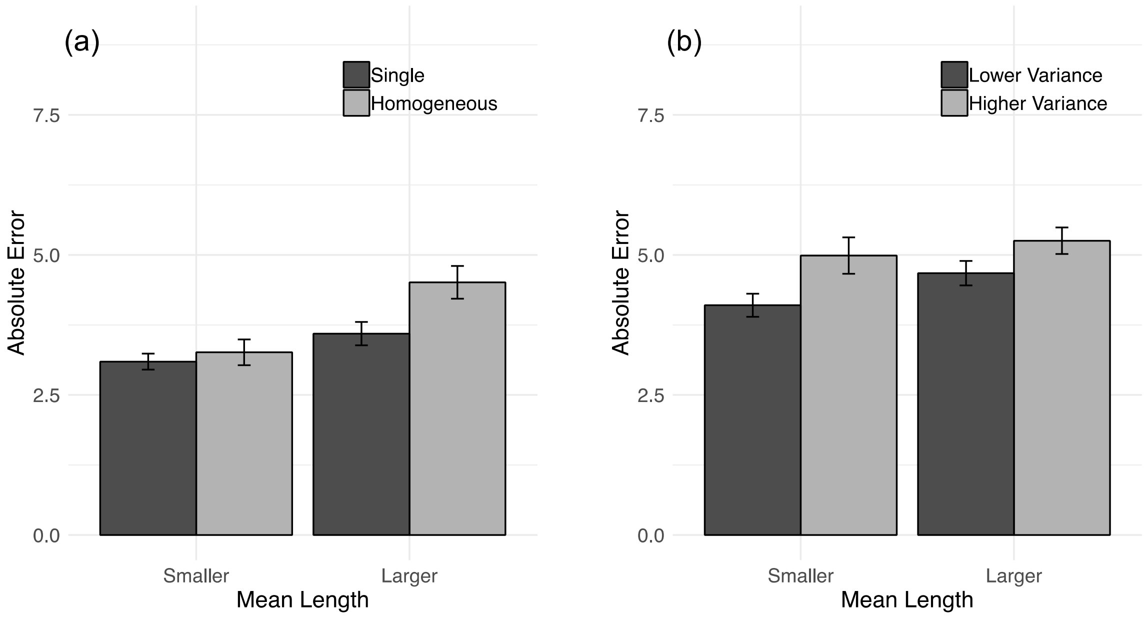

Figure 7. Absolute error in pixels in each condition in Experiment 2B. The left-hand side (Figure 5a) presents the comparison between single and homogeneous conditions and the right-hand side (Figure 5b) presents the data for the lower and higher variance conditions. Error bars indicate standard error.

Estimation bias analysis. As in the previous experiments, we compared estimation bias scores against 1 (see Figure 8). A series of single *t*-tests revealed that in all conditions estimation bias scores were significantly higher than 1, indicating that participants were overestimating the mean in all conditions (for single-line condition: *t*(20) = 2.10, *p* = .049, *d* = .43, homogeneous condition: *t*(20) = 3.45, *p* = .003, *d* = .50, lower variance condition: $t(20) = 2.96$, $p = .008$, $d = .71$, and the higher variance condition, $t(20) = 2.28$, $p = .03$, $d = .45$.

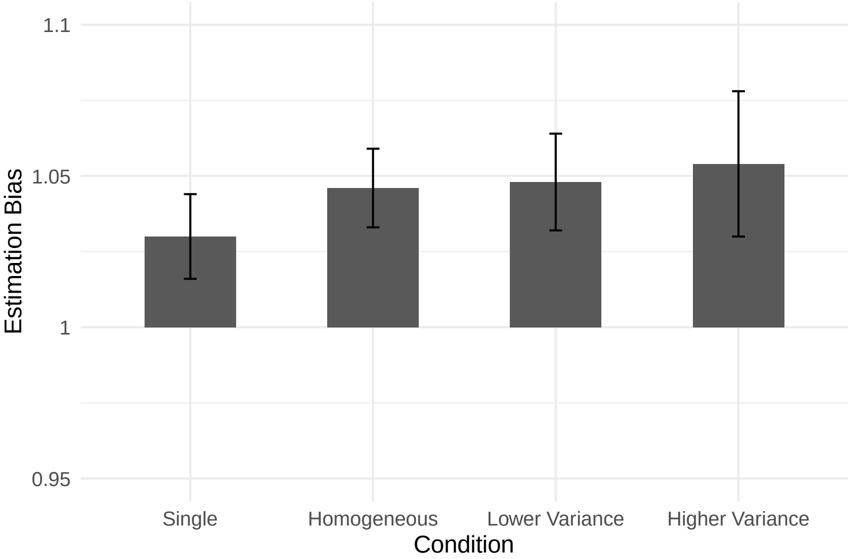

Figure 8. Estimation bias in each condition in Experiment 2b. Error bars indicate standard error.

Comparison of Experiment 2A and 2B

Absolute error analysis. To further examine the effect of variability in the task irrelevant feature on mean length judgments in variance displays, we conducted a 2 (experiment: 2A or 2B) X 2 (variance type: lower or higher) X 2 (mean length: smaller or larger) mixed design ANOVA, with one between-subjects variable (experiment) and 2 within-subjects variables (variance type and mean length). The ANOVA revealed main effects of experiment, $F(1, 39) = 4.42$, $p = .042$, $\eta^2 = .10$; of variance, $F(1, 39) = 31.26$, $p < .001$, $\eta^2 = .45$; and of mean length, $F(1, 39) = 4.42$, $p = .042$, $\eta^2 = .10$. In particular, participants made more error in Experiment 2B ($M = 4.75$, $SD = 1.21$) than in Experiment 2A ($M = 4.24$, $SD = 1.33$), suggesting that the variability in task irrelevant features degrades the performance. Additionally, participants made more error in the higher variance displays

(*M* = 4.95, *SD* = 1.36) compared to the lower variance displays (*M* = 4.06, *SD* = 1.05), and in the larger mean displays (*M* = 4.70, *SD* = 1.15) compared to the smaller mean displays (*M* = 4.31, *SD* = 1.40). Critically, none of the two-way interactions or the three-way interaction were significant (all *F*s < 1.38, all *p*s > .05).

Similarly, for homogenous and single sets, the ANOVA revealed main effects of experiment, *F(*1, 39) = 12.83, *p* < .001, η^2= .25; of variance, *F*(1, 39) = 9.76, *p* = .003, η^2= .20; and of mean length, *F*(1, 39) = 54.36, *p* < .001, η^2= .58. Specifically, participants made more error in Experiment 2B (*M* = 3.62, *SD* = 1.15) than in Experiment 2A (*M* = 2.85, *SD* = 1.09), suggesting that the variability in task irrelevant features impairs the performance. Moreover, participants made more error in the higher variance displays (*M* = 3.46, *SD* = 1.34) compared to the lower variance displays (*M* = 3.02, *SD* = 0.96) and in the larger mean displays (*M* = 3.71, *SD* = 1.18) compared to the smaller mean displays (*M* = 2.77, *SD* = 0.98). None of the two-way interactions were significant (all *F*s < .82, all *p*s > .05). However, the analysis revealed a three-way interaction effect, *F*(1, 39) = 5.50, *p* = .024, η^2= .12, suggesting that compared to Experiment 2A, Experiment 2B resulted in greater performance differences between single and homogenous displays when the mean was larger.

Estimation bias analysis. To compare the estimation bias between Experiment 2A and 2B in variance displays, we conducted a 2 (experiment: 2A or 2B) X 2 (variance type: lower or higher) mixed design ANOVA, with one between-subjects variable (experiment) and one within-subjects variables (variance type). The ANOVA did not reveal any significant effects (all *F*s < 2.44, all *p*s > .05). The comparison between single and homogenous displays across two experiments revealed a main effect of experiment, *F*(1, 39) = 6.18, *p*

= .017, η^2 =.14 ; with participants having larger bias in Experiment 2B (*M* = 1.04, *SD* = .06) than in Experiment 2A (*M* = .99, *SD* = .07), and a main effect of display type, *F*(1, 39) = 4.94, *p* = .032, η^2 = .11; with participants having larger bias in the homogenous line displays (*M* = 1.03, *SD* = .07) than in single line displays (*M* = 1.01, *SD* = .06). The interaction effects did not reach significance, *F* (1, 39) = .47, *p* = .50, η^2= .01).

In sum, comparing the pattern across Experiment 2A and 2B suggests that as the irregularity in displays increased, so did errors as revealed by the main effect of Experiment on absolute errors. The 3-way interaction for error in single and homogeneous conditions across experiments and mean size, along with findings of greater error in the larger mean condition in the previous analyses suggest that the additional variance in the orientation dimension Experiment 2B resulted in a greater noise in responses. Interestingly, there was not a significant interaction across the experiments for the lower and higher variance conditions, suggesting that the added task-irrelevant variance did not differentially contribute to error. The bias analyses also suggest that the task-irrelevant orientation variance impaired performance but this did not interact with the effect of variance. Overall, these analyses suggest that variance of task-relevant and task-irrelevant dimensions may be operating independently on responses.

General Discussion

In this study, across three experiments, we demonstrated that participants were more erroneous in reporting the mean size of circles and mean length of lines in displays that had higher as opposed to lower variance. This finding is in line with claims that argue for the interdependence of mean and variance representations, showing that variance impacts mean judgments (Chong & Treisman, 2003; Corbett et al., 2012; Im & Halberda, 2013; Jeong & Chong, 2020). Critically, we go beyond these studies to demonstrate that there is a systematic tendency to overestimate the mean in highly variable displays, replicating Kanaya et al. (2018), using a continuous report measure (for similar results increased variance on mean emotion extraction from faces, see Ji & Pourtois, 2018). We argue that this pattern cannot be simply explained by a tendency to respond based on a single item or a subset of items that are larger than the mean. In Experiment 1, the proportion of trials in which there were more circles sized larger than the set mean was approximately similar in the smaller and higher variance conditions; if participants were responding based on larger items, then we should have seen a significant overestimation tendency in both conditions. But we only see an overestimation tendency for high variance displays; in this regard our findings differ from Kanaya et al's findings in Experiment 2. However, these differences may be partly due to specific task parameters involving smaller variance conditions. The second reason we argue against an explanation based on a tendency to respond based on a single item or a subset of items that are larger than the mean, is the Experiment 2B results. Specifically, in Experiment 2B, even when reporting the mean of a set of lines that were identical in length (i.e. the homogeneous condition), but that varied in orientation, participants still overestimated the mean. This pattern cannot be explained by what gets subsampled from displays. One possibility is that task-relevant and task-irrelevant variance may impact different stages of ensemble response generation. Comparison of Experiment 2A and 2B results on absolute error and bias and the null experiment x condition interactions hint at this possibility. Even if task-relevant variance impacts subsampling in a manner akin to what Kanaya et al. suggests, variance in a task-irrelevant dimension does not seem to add to the overestimation tendency. Rather, the consistent main effect of Experiment when comparing Experiment 2A and 2B suggest that variance in task-irrelevant dimensions may independently incur costs on responses. While costs associated with interference from concurrent summarization of distributions could be deemed plausible given a system constrained in capacity, what remains unclear is why there would be a systematic tendency to overestimate when there is additional task-irrelevant variance. While our analyses on bias scores did not reveal a significant experiment X condition interaction on bias scores, we would like to note the shift apparent bias shift only in the single item condition. While in Experiment 2A, there was a tendency to underestimate the mean, in Experiment 2B, there was a clear tendency to overestimate, *t*(39)=2.83, *p*=.007. These pattern of results suggests that task-irrelevant variance-that is tiltation in this case, does not increase overestimation tendency overall, but nevertheless impact ensemble responses. Further research is necessary to directly identify mechanisms through which task relevant and task-irrelevant features impact judgments. We discuss some possibilities below

It is no surprise that various visual features and different summary statistics are concurrently summarized (Attarha & Moore, 2015; Boduroglu & Yildirim, 2020; Khvostov & Utochkin, 2019; Poltoratski & Xu, 2013; Utochkin & Vostrikov, 2017; Yörük & Boduroglu, 2020). This ability is useful given the multidimensional nature of natural objects. Furthermore, recent studies have argued that variance and mean information can be extracted in parallel by independent systems (Khayat & Hochstein, 2018; Khvostov & Utochkin, 2019; Yang et al., 2018; but also see Jeong & Chong, 2020). However, less is known about how such statistical summaries, often extracted effortlessly, interact. One possibility is that there may be attentional limits to ensemble perception leading to costs associated with concurrent processing. That might be why we see greater error in mean estimates in the high variance condition; participants may be concurrently summarizing both the mean and variance of these displays, even though the latter is not required. It is known that participants can automatically summarize range (Khayat & Hochstein, 2018) and tag outliers (Cant & Xu, 2020) during mean extraction. However, such an explanation would still be insufficient to explain why there is a systematic tendency to overestimate. Furthermore, such costs have been more typically reported for tasks in which participants had to concurrently summarize several ensembles rather than when they had to extract several summaries (e.g. Attarha, Moore & Vecera, 2014; Emmanoil & Treisman, 2008; Utochkin & Vostrikov, 2017). In our case, in the higher variability condition, participants may have been extracting two different types of summaries from a single ensemble, rendering this explanation unlikely.

We believe that our findings along with other reports of increased error in higher variability displays need to be considered in relation to the question of what makes a visual ensemble. It is possible that as variability further increases in displays, the visual system's ability to utilize ensemble perception mechanisms to summarize them may decrease, because displays may lose their "gestalt" quality (e.g Utochkin & Tiurina, 2014). Our earlier work has shown that when displays consisted of two spatially segregated sets of circles, participants were able to extract the global mean in these displays as efficiently as extracting the mean in displays that had the same number of circles presented as part of a single spatial group (Yildirim, Öğreten & Boduroglu, 2018).

Future research has to identify what type of and how much variability may interfere with a display's perceived gestalt quality. In a related vein, future research has to more directly investigate whether this tendency to overestimate the mean when features have higher variability has a functional value or whether it reflects a limitation of the ensemble perception mechanisms.One possibility is that in higher variance displays, overestimating the mean may reduce the likelihood of tagging items that are on the higher edge of the feature range as outliers. This in turn would may reduce the likelihood of bigger, longer features from being discounted as outliers. Rather, when the mean is overestimated, these items may be more likely to be considered a part of the ensemble, yet they may still have greater saliency and be included in the summaries, contributed to the overestimation as Kanaya et al. (2018) suggests. Otherwise, if such features are tagged as perceptual outliers, this might result in them being represented with high precision items at the expense of reduced precision of summary statistics of the remaining set of features (Avcı & Boduroglu, accepted; Cant & Xu, 2020). One other consequence of the tendency to overestimate the mean may result in the items on the lower set-range to be tagged as outliers (or outlier-like). Discounting these latter group of items on the lower range of the feature distribution may be relatively inconsequential for ensemble perception (e.g. for outlier discounting see Haberman & Whitney, 2010) and for the subsequent scaffolding processes supporting visual short-term memory (e.g. Brady & Alvarez, 2011; Brady & Tenenbaum, 2013). A similar pattern was also reported during averaging of Arabic numerals: participants discounted smaller digits and over-weighed of larger digits (van Opstal, de Lange, & Dehaene, 2011). In a related vein, participants overestimated the sum when items were more irregular. For instance, Charras and Lupiáñez (2009) demonstrated that the sum of two asymmetrically bisected parts of a line were perceived as greater than the sum of two symmetrically bisected parts. Charras, Bord and Lupiáñez (2012) reporting that the sum of a set including repeated numbers were underestimated while the sum of a set with unique numbers were overestimated. More research is necessary to delineate how summation and averaging are linked. Within ensemble perception, it is typically accepted that there is no devoted mechanism supporting summation of variables as length or area (e.g. Raidvee et al., 2020). Also, numerous studies have shown that mean size is not extracted by dividing the total area covered by the number of items in a display (e.g. Cain & Cain, 2018; Utochkin & Vostrikov, 2017). Nevertheless, it may be that summation judgments may depend on mean and numerosity judgments; the multiplication of these two statistical summaries may lead to similar biases in mean and summation judgments. Future studies have to more directly investigate this possible link.

We do acknowledge there are some limitations of the current work. In all experiments, we chose to use a rather long delay interval, mimicking the procedural timeline in Brady & Alvarez (2011). This delay period might have caused working memory processes to further amplify or reduce biases in responses. While we cannot rule out this possibility, we found no evidence of a link between error and span measures[4]. The conceptual replication of the second experiment in Kanaya et al. (2018) further makes us think that these findings are not merely artifacts of the delay period. A second issue pertains to the difference in the characteristics of the distributions from which the

[4] For exploratory purposes not central to the main questions discussed in this paper, we also had all participants complete the automated operation span task (Unsworth, Heitz, Schrock, & Engle, 2005). For each experiment, for all conditions we separately calculated the correlation between error and ospan scores (Experiment 1, all r<.29; Experiment 2: all rs<.44, and Experiment 3: all rs<.38; all ps>.05)

lower and higher variance sets were generated. Specifically, the larger and smaller mean sets used in the higher variance displays had a larger range of possible values to draw from compared to those in the lower variance displays. Our aim in choosing a large range was to equate the effects of possible “perceptual” outliers on mean size estimations among sets with different mean sizes. In other words, observers would be presented with the same individual item, regardless of the mean size of the set. However, because of the lower probability of generating sets with low variance in larger ranges, which could meet the specified constraints, we used non-overlapping intervals for smaller and larger mean sets while generating lower variance displays. This means that the sets in the higher variance displays might have inherently had more extreme values. However, it must be noted that probabilistically speaking these extreme values are equally likely to be drawn from smaller or larger extremes. If the larger subset of extreme values were subsampled, this might have exaggerated the overestimation bias we observed. This might have been true especially when participants were presented with circles (Expt. 1) as opposed to lines. In other words, while attempting to equate the mathematical variance across conditions, we might have created stimuli sets that differed on perceived variance. Future research is needed to establish the correspondence between mathematical and perceived variance and whether these have different impact on summarization. This possible confound though partially allows a saliency- based interpretation of the overestimation effect.

In sum, we argue that our data add to a growing body of work that suggests that the visual system may be concurrently extracting different types of summary statistics (e.g. mean, variance, or possibly range) for different visual features (e.g. length and orientation) in a somewhat interdependent fashion. The exact interplay between these may depend on the featural properties and the capacity limitations, resulting in systematic biases to overestimate a given characteristic.

Open Practices Statement

The data and materials for all experiments are available upon request. None of the experiments were preregistered.

References

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

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science*, *12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Attarha, M., & Moore, C. M. (2015). The perceptual processing capacity of summary statistics between and within feature dimensions. *Journal of Vision*, *15*(4), 9–9. https://doi.org/10.1167/15.4.9

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. https://doi.org/10.1037/a0036206.

Avcı, B. & Boduroglu, A. (under review). Contributions of ensemble perception to outlier representation precision..

Boduroglu, A., & Yildirim, I. (2020). Statistical summary representations of bound features. *Attention, Perception, & Psychophysics*, *82*(2), 840–851. https://doi.org/10.3758/s13414-019-01944-9

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109. https://doi.org/10.1037/a0030779

Cain, S., & Cain, M. (2018). A Texture Representation Account of Ensemble Perception. *Journal of Vision*, *18*(10), 618–618. https://doi.org/10.1167/18.10.618

Cant, J. S., & Xu, Y. (2020). One bad apple spoils the whole bushel: The neural basis of outlier processing. *NeuroImage*, *211*, 116629. https://doi.org/10.1016/j.neuroimage.2020.116629

Charras, P., Brod, G., & Lupiáñez, J. (2012). Is 26 + 26 smaller than 24 + 28? Estimating the approximate magnitude of repeated versus different numbers. *Attention, Perception, & Psychophysics*, *74*(1), 163–173. https://doi.org/10.3758/s13414-011-0217-4

Charras, P., & Lupiáñez, J. (2009). The Relevance of Symmetry in Line Length Perception. *Perception*, *38*(10), 1428–1438. https://doi.org/10.1068/p6287

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404. https://doi.org/10.1016/S0042-6989(02)00596-5

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

Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across feature dimensions in statistical processing of perceptual groups. *Perception & Psychophysics*, *70*(6), 946–954. https://doi.org/10.3758/PP.70.6.946

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432–446. https://doi.org/10.1037/xge0000053

Haberman, J., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision*, *15*(4), 16–16. https://doi.org/10.1167/15.4.16

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

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. https://doi.org/10.3758/s13414-012-0399-4

Jeong, J., & Chong, S. C. (2020). Adaptation to mean and variance: Interrelationships between mean and variance representations in orientation perception. *Vision Research*, *167*, 46–53. https://doi.org/10.1016/j.visres.2020.01.002

Ji, L., & Pourtois, G. (2018). Capacity limitations to extract the mean emotion from multiple facial expressions depend on emotion variance. *Vision Research, 145*, 39-48. https://doi.org/10.1016/j.visres.2018.03.007

Kanaya, S., Hayashi, M. J., & Whitney, D. (2018). Exaggerated groups: Amplification in ensemble coding of temporal and spatial features. *Proceedings of the Royal Society B: Biological Sciences, 285*(1879), 2017-2770. https://doi.org/10.1098/rspb.2017.2770

Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, *18*(9), 23–23. https://doi.org/10.1167/18.9.23

Khvostov, V. A., & Utochkin, I. S. (2019). Independent and parallel visual processing of ensemble statistics: Evidence from dual tasks. *Journal of Vision*, *19*(9), 3–3. https://doi.org/10.1167/19.9.3

Maule, J., & Franklin, A. (2020). Adaptation to variance generalizes across visual domains. *Journal of Experimental Psychology: General*, *149*(4), 662–675. APA PsycArticles. https://doi.org/10.1037/xge0000678

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

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A 'dipper' function for texture discrimination based on orientation variance. *Journal of Vision*, *8*(11), 9–9. https://doi.org/10.1167/8.11.9

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. https://doi.org/10.3758/PP.70.5.772

Norman, L. J., Heywood, C. A., & Kentridge, R. W. (2015). Direct encoding of orientation variance in the visual system. *Journal of Vision*, *15*(4), 3–3. https://doi.org/10.1167/15.4.3

Phillips, L. T., Slepian, M. L., & Hughes, B. L. (2018). Perceiving groups: The people perception of diversity and hierarchy. *Journal of Personality and Social Psychology*, *114*(5), 766. https://doi.org/10.1037/pspi0000120

Poltoratski, S., & Xu, Y. (2013). The association of color memory and the enumeration of multiple spatially overlapping sets. *Journal of Vision*, *13*(8), 6–6. https://doi.org/10.1167/13.8.6

Raidvee, A., Toom, M., Averin, K., & Allik, J. (2020). Perception of means, sums, and areas. *Attention, Perception, & Psychophysics*, *82*(2), 865–876. https://doi.org/10.3758/s13414-019-01938-7

Rodriguez-Cintron, L. M., Wright, C. E., Chubb, C., & Sperling, G. (2019). How can observers use perceived size? Centroid versus mean-size judgments. *Journal of Vision*, *19*(3), 3–3. https://doi.org/10.1167/19.3.3

Tokita, M., Ueda, S., & Ishiguchi, A. (2016). Evidence for a Global Sampling Process in Extraction of Summary Statistics of Item Sizes in a Set. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00711

Tong, K., Ji, L., Chen, W., & Fu, X. (2015). Unstable mean context causes sensitivity loss and biased estimation of variability. *Journal of Vision*, *15*(4), 15–15. https://doi.org/10.1167/15.4.15

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498-505.

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

Utochkin, I. S., & Vostrikov, K. O. (2017). The numerosity and mean size of multiple objects are perceived independently and in parallel. *PLoS ONE*, *12*(9). https://doi.org/10.1371/journal.pone.0185452

Van Opstal, F., de Lange, F. P., & Dehaene, S. (2011). Rapid parallel semantic processing of numbers without awareness. *Cognition*, *120*(1), 136–147. https://doi.org/10.1016/j.cognition.2011.03.005

Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access. *Cognition*, *152*, 78–86. https://doi.org/10.1016/j.cognition.2016.01.010

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology*, *69*(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Yang, Y., Tokita, M., & Ishiguchi, A. (2018). Is There a Common Summary Statistical Process for Representing the Mean and Variance? A Study Using Illustrations of Familiar Items. *I-Perception*, *9*(1), 2041669517747297. https://doi.org/10.1177/2041669517747297

Yildirim, I., Öğreden, O., & Boduroglu, A. (2018). Impact of spatial grouping on mean size estimation. *Attention, Perception, & Psychophysics*, *80*(7), 1847–1862. https://doi.org/10.3758/s13414-018-1560-5

Yörük, H., & Boduroglu, A. (2020). Feature-specificity in visual statistical summary processing. *Attention, Perception, & Psychophysics*, *82*(2), 852–864. https://doi.org/10.3758/s13414-019-01942-x