Impact of spatial grouping on mean size estimation

Irem Yildirim¹ · Oğuzhan Öğreden² · Aysecan Boduroglu^{1,3}

Published online: 3 July 2018
© The Psychonomic Society, Inc. 2018

Abstract

People represent summary statistics of visual scenes, but it is not fully clear whether such summary statistics are extracted automatically. To determine whether different levels of summary representation (i.e., at the perceptual-group or the entire-display level) may be formed differently, in two experiments we investigated how people extracted summary statistics for displays consisting of spatially segregated groups. Participants were asked to report the mean sizes of either entire sets or perceptual groups in precue and postcue conditions. There was no precueing advantage in the mean size estimations of entire sets. However, when these precues identified target perceptual groups, participants reported the perceptual-group means more accurately than when postcues were used. In the postcue condition, participants were biased toward the entire-set mean even when they were probed to report the perceptual-group mean. There was also greater bias toward the entire-set mean for more erroneous perceptual-group summaries. These findings suggest that ensemble representations are extracted more efficiently for the whole than for the perceptual parts and that ensemble perception is not a uniform process across perceptual groups and entire sets.

Keywords Automaticity · Perceptual organization · Visual perception

It is known that people can extract summary statistics of visual scenes (for a review, see Alvarez, 2011). Statistical summary representations, known also as *ensemble representations*, include compressed information about a set of similar objects in a given display, albeit less veridical representations of individual items (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2014; Ariely, 2001; Corbett & Oriet, 2011; Ward, Bear, & Scholl, 2016). This higher-level information can be the gist of a natural scene (Oliva & Torralba, 2006) or the statistical properties of simple features such as their average (Ariely, 2001; Chong & Treisman, 2003) or their variance (Morgan, Chubb, & Solomon, 2008; Peterson & Beach, 1967; Semizer & Boduroglu, 2016). Research revealed that people can form accurate ensemble representations for size of constant (Ariely, 2001; Chong & Treisman, 2003) and dynamic objects (Albrecht & Scholl, 2010), length (Semizer & Boduroglu, 2016), orientation (Parkes, Lund, Angelucci, Solomon, &

Morgan, 2001), brightness (Bauer, 2009), speed (Emmanouil & Treisman, 2008), centroid of object locations (Alvarez & Oliva, 2008; Boduroglu & Shah, 2014; Mutlutürk & Boduroglu, 2014), spatial patterns of orientation (Alvarez & Oliva, 2009), and auditory tones (Albrecht, Scholl, & Chun, 2012). The ability to establish statistical summaries is also evident with higher-level stimuli. For instance, people can average emotional expression and gender of faces (Haberman & Whitney, 2007, 2009), face identities (de Fockert & Wolfenstein, 2009; Kramer, Ritchie, & Burton, 2015; Neumann, Schweinberger, & Burton, 2013), and biological motion of crowds (Sweeny, Haroz, & Whitney, 2013). Moreover, ensemble representations were shown to influence motor behavior and visual experience (Corbett & Song, 2014; Dubé & Sekuler, 2015; Joo, Shin, Chong, & Blake, 2009) pointing to its critical role for our daily lives. Nevertheless, the exact mechanisms of ensemble perception is not fully known. Although some evidence has suggested that statistical summaries may be extracted automatically (e.g., Chong & Treisman, 2003), other researchers have argued that these representations are formed via the averaging of certain regions or elements in a display (i.e., subsampling; e.g., Myczek & Simons, 2008). A third possibility is that independent summary mechanisms may operate at different levels of the visual hierarchy, differentially impacted by processing demands of particular stimuli. For instance, Haberman, Brady, and

✉ Aysecan Boduroglu
aysecan.boduroglu@boun.edu.tr

¹ Bogazici University, Bebek-Istanbul, Turkey

² Utrecht University, Utrecht, The Netherlands

³ Bogazici University, Bebek-Istanbul, Turkey

Alvarez (2015) has shown that performance on tasks that require summarizing facial information (identify, emotion etc.) and more basic features (orientation and color) are independent and they rely on different pooling mechanisms. Similarly, it is also possible that modality-specific, even feature-specific, pooling mechanisms may enable summarizing of different types of information. Previous work from our lab has shown that errors in summarizing visual and spatial displays were independent (Uner, Mutlutürk, & Boduroglu, 2014); more recently, we have demonstrated that mean length and orientation estimates were not related (Yoruk & Boduroglu, 2018). These findings suggests that ensemble perception may not be a uniform process.

In light of these, we wanted to investigate how displays consisting of spatially segregated perceptual groups are summarized and determine whether different levels of summary representations (e.g., at the perceptual-group level or at the entire display level) are all formed automatically.

Automaticity is typically attributed to a very rapid process that is not impacted by attentional manipulations (Brown, Gore, & Carr, 2002; Gmeindl, Nelson, Wiggin, & Reuter-Lorenz, 2011; Naveh-Benjamin, 1987; Posner, Sandson, Dhawan, & Shulman, 1989; Posner & Snyder, 2004). In other words, this suggests that there should be neither an ameliorative effect of selective attention nor a detrimental effect of divided attention on automatic processes. Furthermore, if a process is automatic, the outcome of this process could impact and bias other ongoing or subsequent processes. To date, research on ensemble representations has investigated whether statistical summary accuracy is influenced by manipulating factors like set size, density, attentional demands and encoding duration (e.g., Chong, Joo, Emmanouil, & Treisman, 2008; Chong & Treisman, 2003, 2005a, b); some studies have also investigated whether item representations are biased toward available ensemble summaries (e.g., Brady & Alvarez, 2011; Mutlutürk & Boduroglu, 2014). Below we first review the contradictory findings on whether ensemble representations are formed automatically, and then present our present set of experiments.

Display characteristics and ensemble perception

Typically, most studies on ensemble representations ask people to indicate the average of simple features in a briefly presented display. There are some conflicting results about how set size and encoding manipulations affect the accuracy of ensemble representations. One of the most important pieces of evidence for the automaticity claim is that set size, the number of items that are averaged together, does not influence the accuracy of that average (Ariely, 2001; Chong & Treisman, 2003). These findings suggest that ensemble

perception is based on distributed attention (Maule & Franklin, 2015) and is driven by the perception of a whole or a pool of items. However, in the early studies the set-size manipulation was confounded such that the displays consisted of multiple identical items (Ariely, 2001). This led others to argue that people may strategically extract ensemble information by focusing on just a few objects of the display—that is, *subsampling* (de Fockert & Marchant, 2008; Myczek & Simons, 2008). When repetitions were not allowed, set size impacted performance; mean size estimations were worse for larger than smaller sets (Marchant, Simons, & de Fockert, 2013), and this decrement in performance was attributed to limitations in visual working memory capacity. More recently, Maule and Franklin (2015) showed that the number of items did not impact hue averaging performance; the pattern of results remained the same despite variations in the number of repetitions, but instead was impacted by within-set variability. Thus, they concluded that their findings supported a distributed-attention-based mechanism for ensemble perception. However, they also noted that a distributed-attention-based account was not fully incompatible with subsampling; viewers might attend holistically, but then they might exclude certain items and focus on others while determining the summary. Such a flexible use of subsampling while a display is being holistically attended might be the basis of outlier exclusion (Haberman & Whitney, 2010; Yildirim & Boduroglu, 2018). Indeed, it is known that viewers can flexibly adopt a subsampling strategy during ensemble perception based on task demands. For instance, Dakin, Bex, Cass, and Watt (2009) demonstrated that observers make adjustments in the number of items that they sample on the basis of the number of items in the ensemble. Thus, under certain conditions, subsampling-based ensemble perception may be an efficient and viable strategy (Sweeny, Wurnitsch, Gopnik, & Whitney, 2015).

Time course of ensemble perception

There has also been some debate regarding how rapidly ensemble representations are formed. Those who claim that ensemble perception happens in the early perceptual processing stage argue that object identification is not necessary for the extraction of summaries (Allik et al., 2014; Ariely, 2001; Corbett & Oriet, 2011; Ward et al., 2016). Chong and Treisman (2003) demonstrated that statistical summary representations could be formed within 50 ms. In a similar vein, Choo and Franconeri (2010) showed that mean size was affected even by the presence of extra circles that were masked after 30 ms, to reduce their visibility. These findings are more in favor of the automaticity view. Nevertheless, some evidence has suggested that ensemble representations are impacted by later object processing. For instance, Whiting and Oriet

(2011) reported that a 200-ms encoding duration was needed to form summary representations when objects were masked after the presentation. In a similar vein, Jacoby, Kamke, and Mattingley (2013) found that object substitution masking degraded both mean size and mean orientation estimates. Thus, the evidence is at best mixed and does not conclusively support the automaticity claim.

Perceptual groups and ensemble perception

Studies investigating whether statistical summaries of perceptual groups in a given display are processed simultaneously have also contributed to the debate on how ensemble representations are formed. Specifically, the presence of multiple perceptual groups in displays allows researchers to determine whether concurrent summary mechanisms operate at the perceptual-group as well as the entire-display level. Furthermore, these types of displays make it possible to manipulate attention toward a subset of items. Two studies directly investigating the role of perceptual grouping on ensemble representations have yielded conflicting results (Brand, Oriet, & Tottenham, 2012; Chong & Treisman, 2005b). In both studies, circle displays were composed of two interspersed color groups. Participants were randomly asked to determine the mean size of a particular color-coded subset¹ via a two-alternative forced choice test. There were three conditions: precueing the to-be-tested color subset, postcueing, and a control condition consisting of a single set of circles lacking any subsets. In the precue condition, participants could selectively attend to a particular perceptual subset. In the postcue condition, since participants did not know which subset they would be tested on, they needed to summarize both color groups concurrently and to selectively attend to a particular subset representation. Both studies confirmed that participants were able to concurrently average two perceptual groups, indicated by better-than-chance postcue performance. To determine whether there was any cost of concurrent extraction of averages, the precue and postcue conditions were compared. In Chong and Treisman (2005b), similar performance across the precue and postcue conditions demonstrated that concurrent averaging occurs without any cost. On the other hand, Brand et al. (2012) criticized Chong and Treisman (2005b), arguing their alternatives did not prevent participants from utilizing the entire-set mean to determine the correct answer. Indeed, once Brand et al. (2012) eliminated this confound and provided performance-based feedback to participants, a precue advantage emerged in the data, suggesting a cost of concurrent averaging.

¹ “Subset” and “perceptual group” are used interchangeably throughout the article.

More recently, Attarha and colleagues (Attarha & Moore, 2015; Attarha, Moore, & Vecera, 2014) demonstrated that viewers may not be able to concurrently average a high number of subsets. They used the simultaneous–sequential paradigm, in which four spatially determined subsets were presented either simultaneously in one screen or sequentially in two consecutive screens, each containing two subsets. The authors reported that simultaneous presentation of four groups led to worse performance in identifying the subset than did the “odd” average size (Attarha et al., 2014) and orientation (Attarha & Moore, 2015), pointing to a limit of averaging ability. This result is also similar to findings in which people have enumerated only up to three color groups in parallel (Halberda, Sires, & Feigenson, 2006; but see Poltoratski & Xu, 2013). Attarha and colleagues also found that mean size (Attarha et al., 2014) and mean orientation (Attarha & Moore, 2015) performance for a single, spatially nonsegregated ensemble was the same across simultaneous and sequential presentations. Thus, a single ensemble was processed automatically. However, it should be noted that in their experiment, the single ensemble lacked any obvious perceptual groups. When an ensemble contained color subsets, Brady and Alvarez (2011) showed that observers biased their reports of individual item sizes toward the entire-set mean instead of the color-group mean that the item belonged to. This suggests that viewers may be more likely to automatically encode the mean of the entire set than the means of two subsets. Also, Oriet and Brand (2013) demonstrated that people could not ignore task-irrelevant subsets, since reports of the relevant subset’s mean size were biased toward the irrelevant subset mean, even if there was unlimited time to study the display. Therefore, it appears that the summary of an entire ensemble may be extracted in a rather automatic and unintentional fashion, but that the nature of the perceptual groups in displays and the specific task demands may reduce the efficiency with which the perceptual groups are summarized.

The present study

In two experiments, we investigated how people extracted summary statistics for displays consisting of spatially segregated perceptual groups. In a cueing paradigm (Brand et al., 2012; Chong & Treisman, 2005b), we asked participants to report either the mean size of one of the two spatially segregated groups, in the upper and lower visual fields, or the entire set, by adjusting the size of a response circle. The present study differs from previous studies that have used cueing in an ensemble perception context in three major ways. First, we set spatial segregation as the grouping factor instead of color. Spatial segregation taps on the Gestalt principle of proximity, which leads to perceptual grouping of the items closer to each other (Brooks, 2015; Wertheimer, 1923). In fact, proximity is

a powerful principle that acts very quickly (e.g., quicker than similarity; Ben-Av & Sagi, 1995) and leads to greater perceived segregation between groups and less perceived distance between items within a group (Coren & Girgus, 1980). Second, in our study the perceptual groups were not interspersed, thus eliminating any possible location-based interference. Finally, it was possible that performance on the recognition-based tasks used in both previous studies (Brand et al., 2012; Chong & Treisman, 2005b) may have been differentially impacted by the forced choice alternatives. Therefore, to exclude those influences, in the present study we required participants to generate the mean size instead of recognizing it from the alternatives. Although it is difficult to characterize open-ended continuous responses normatively (as being below or above chance), interpreting the findings in comparison to random observer simulations can circumvent this caveat. Furthermore, this open-ended, continuous response also allowed us to compute a bias score, in order to determine whether the perceptual-group responses were pulled toward the more readily available entire-set mean. We expected this bias—if it occurred at all—to be stronger whenever the perceptual summaries were more erroneous.

The novelty of our study lies in the fact that we asked viewers to process two perceptually segregated groups and report either one of their summaries or the summary of the entire set. This allowed us to examine whether the perceptual groups were less efficiently summarized than entire sets and whether holistic processing of the entire set occurred above and beyond that of perceptual sets. Also, spatially segregated groups allowed us to explore whether perceptual groups in the upper (UVF) and lower (LVF) visual fields were summarized equally efficiently and specifically test the prediction that a more erroneous subset summary would be more biased toward the efficient, entire-set summary. Because performance on various visual tasks are impacted by vertical meridian asymmetry (for reviews, see Karim & Kojima, 2010; Previc, 1990), we explored whether such a pattern would emerge during ensemble perception. It is widely proposed that attentional selection might favor the LVF over the UVF (Cameron, Tai, & Carrasco, 2002; Fuller, Rodriguez, & Carrasco, 2008; He, Cavanagh, & Intriligator, 1996; Talgar & Carrasco, 2002) and that attentional resolution might be better in the LVF than in the UVF. When attentional resolution is higher, visual analysis may operate at a finer-scale to give rise to more detailed representations. In contrast, lower attentional resolution would result in coarser analysis and more holistic/global processing and consequently more pooling of input. In support of this attentional mechanism, detection of an item crowded by nearby items was found to be better in the LVF than that in the UVF (Bulakowski, Post, & Whitney, 2011; He et al., 1996). That is, item detection was impacted more by the presence of flanker items in the UVF, consistent with claims suggesting that the UVF is processed more globally. Because both crowding and

ensemble perception has been suggested to rely on similar pooling mechanisms (Fischer & Whitney, 2011; Greenwood, Bex, & Dakin, 2009; Parkes et al., 2001), we expected that there might be more global processing and pooling in the UVF, leading to more accurate summary representations. Consequently, if there was greater error in the LVF, than we would expect the summaries to be biased more toward the entire-set mean than the subsets in the UVF would be.

Experiment 1

Participants reported the mean size of either a subset or the entire set under both precue and postcue conditions. In addition, there was a control condition in which participants reported the mean size of a single spatially nonsegregated set. On the basis of the literature, we expected that viewers would extract the mean size of the entire set automatically. Therefore, we expected entire-set performance to be similar across the precue, postcue, and single-set conditions. If people average perceptual groups concurrently without any cost, there should be no difference between the precue and postcue conditions. If concurrent averaging comes with a cost, on the other hand, then postcuing should result in less accurate subset mean estimations than the precueing and single-set conditions. Also, if there is an asymmetry along the vertical meridian, a UVF advantage is more likely to be revealed, since attentional resolution is lower and pooling mechanisms work better in the UVF than in the LVF. Thus, mean size accuracy may be better for groups in the upper than in the lower half of the display. Finally, we predicted that subset reports would be biased toward entire-set mean whenever the representations are more erroneous.

Method

Participants Twenty-five (18 female, seven male; $M_{\text{age}} = 20$ years, age range: 18–23 years) Bogazici University undergraduates were recruited for the study in exchange for course credit. All reported normal or corrected to normal vision. When exploring the data, we looked at the Shapiro–Wilk normality results for average error in precue, postcue, and single-set conditions. These were significant for the precue and postcue conditions, $ps = .024$ and $.032$, respectively, indicating the nonnormality of the accuracy (i.e., error in the reported size) distribution. Inspection of the boxplots identified one participant as an outlier in both the precue and postcue conditions. When this person's data were excluded, the Shapiro–Wilk normality test was no longer significant for any of the three main conditions, all $ps \geq .7$. Thus, the reported analyses are based on 24 participants (17 female, seven male).

Materials and stimuli We used the mean size estimation task under different attentional cueing conditions to determine whether people can modulate their attention to average two vertically segregated groups as well as the entire set. The task had three conditions, two attentional cueing conditions with precue and postcue and one control condition containing non-segregated single sets. The conditions were blocked, and the order was counterbalanced in a Latin square design across participants. Each block consisted of 160 trials. In half of the trials, participants generated perceptual-group means (or the sets of 8) and in the other half entire-set means (sets of 16). We asked for the mean sizes of the subsets equally often in the UVF and the LVF across all cueing conditions. No more than four consecutive trials asked for the mean size of the subset in the UVF, the subset in the LVF, or the entire set in either cueing condition. Similarly, in the single-set condition no more than four consecutive trials had 8-circle or 16-circle sets.

In the precue and postcue conditions, the displays of circles were vertically segregated into two groups such that one spatial group of eight circles were located in the upper half and the other group of eight circles were in the lower half of the screen. In the precue condition, two arrows were presented that indicated the to-be-tested set. These arrows were on both sides of the fixation cross and remained visible for another 500 ms after the cross was removed. Both arrows were either up or down arrows, for the subsets, or pointed to opposite sides from the fixation point, referring to the entire set. In the postcue condition, the position of the test circle—at the center of the upper half, the lower half, or the whole screen—indicated the relevant set. Thus, in the precue condition observers could have utilized the arrows to covertly focus their attention on the relevant perceptual set, whereas in the postcue condition they had to choose the right perceptual set representation as their target response using their inner representational space.

In the single-set condition, the displays were composed of either 8 or 16 circles, which matched the set size of the group (for sets of 8) or entire-set (for sets of 16) trials in the cueing conditions. Because there were no groups in the single-set trials, the test circle was always shown at the center of the whole screen.

Overall, a trial (see Fig. 1a) began with a fixation cross that appeared at the center of the screen for 500 ms. Then, circle sets were displayed for 200 ms. After the study display, a test circle was immediately presented. Participants reported the target's mean size by adjusting the test circle with keypresses (D to make the response circle smaller, and K to make it larger). The response range was not restricted and was very large. Technically, the given test circle could be made as small as a dot (i.e., a circle with a 2-pixel/0.1° diameter) and as large as to cover the whole screen (i.e., a circle with a 480-pixel/24° diameter). One keypress changed the circle diameter by 2 pixels/0.1°. The initial test item diameter was at least 10 pixels/0.5° and at most 20 pixels/1° smaller or larger than

the correct response diameter. Thus, participants had to press the key at least five at most ten times in order to achieve the correct response. That is to say, the test item differed from the correct size by at least 15% and at most 71%. During the 1,000-ms intertrial interval, a blank gray screen was presented.

In each trial, white outlined circles whose diameters ranged from 1° to 4.2° were presented on a gray background. Circles were chosen randomly from this range with 0.2° increments. This least difference for our size range was calculated on the basis of a power function between perceived and physical circle sizes with an exponent of 0.76 (Teghtsoonian, 1965). That is, when there was only a 0.2° difference in their physical sizes, they were still perceived as different (i.e., above the just noticeable difference). In each perceptual group and for the 8-circle single sets, a maximum of two circles were of the same size; in the 16-circle single sets, four circles could be the same size. The mean size of a perceptual group in the cueing conditions differed at least by 40% from the other group mean, and by 20% from the entire-set mean. This created a smaller and a larger mean in a given display. These smaller and larger means occupied the UVF and the LVF evenly across the different cueing conditions. Also, the larger or the smaller subset means were asked for equally in the UVF and LVF. Thus, when the smaller mean in the display was required, it was located equally often in the upper and in the lower visual field.² We matched the possible means of the grouped displays to those of the nongrouped displays in the single-set condition (i.e., the means of the 8-circle single sets equaled the subset means, and the means of the 16-circle single sets equaled the entire-set means).

In the study displays, we left empty 1.45° of visual angle from the left and right and 0.95° from the top and bottom edges of the screen. In the precue and postcue trials, the two subsets were spatially segregated by 1.5° of horizontal space at the center. Eight circles within each perceptual group were pseudorandomly placed in 18 unevenly sized cells of an imaginary 3×6 grid, with some restrictions to minimize collinearity. There were at most three circles in a row and two circles in a column of each perceptual group. Moreover, a maximum of three circles were adjacent to each other in each subset. In the single-set condition there was not any horizontal gap. Either eight or 16 circles were positioned on an imaginary 5×6 grid, leading to 30 equally sized cells. There were at most three circles in a row for both the 8-circle and 16-circle sets. A maximum of two circles and four circles in a column were presented for the sets of eight and 16, respectively. Also, there

² We carried out a pilot study in which we did not control the distribution of the larger and smaller means in the UVF and LVF. The general pattern of results replicated the main findings of the study presented here, in the sense that the greatest amount of error was in the postcue and the least was in the single-set conditions. Similar patterns of bias emerged. For brevity's sake we do not include those results, but instead focus on the results from the study that did not have this methodological confound.

Fig. 1 Representative illustration of the task: The precue, postcue, and single-set conditions, from top to bottom. (a) Experiment 1. (b) Experiment 2. In the original task the background was gray, and in Experiment 2 an additional set of catch trials were presented that are not depicted above

were no more than two circles in the surrounding cells of an occupied cell. The circle locations in all conditions were jittered in such a way that the minimum interitem distance of 0.2° was maintained.

The task was prepared and run via the E-Prime 2.0 software (Psychology Software Tools, 2012) on a 17-in. computer screen measuring 640×480 pixels. Participants sat at a distance of 57 cm from the screen, and thus 1° of visual angle corresponded to approximately 20 pixels.

Results and discussion

Entire-set and subset averaging For each participant, we calculated the absolute error in each trial as the index of accuracy by subtracting the reported radius from the correct radius in pixels. On the basis of the absolute error data of the whole sample, we excluded trials exceeding three standard deviations from the mean of the particular level. Then we averaged the absolute errors in the trials of each level for each observer. We conducted a 3 (Condition: precue, postcue, single set) $\times$ 2 (Set Type: subsets/8-circle sets, entire sets/16-circle sets) repeated measures analysis of variance (ANOVA) on absolute errors (Fig. 2a). As expected, we found no effect of set type on errors, $F < 1$, n.s., replicating earlier research that set size did not impact accuracy

levels (Ariely, 2001; Maule & Franklin, 2015). The main effect of condition was significant, $F(2, 46) = 11.92$, $MSE = .58$, $p < .001$, $\eta_p^2 = .34$. Pairwise comparisons³ showed that participants made more errors in the postcue ($M = 4.22$, $SD = 0.89$) than in the precue ($M = 3.43$, $SD = 0.66$), $p < .001$, $d = 1.19$, and single-set ($M = 3.48$, $SD = 0.86$), $p = .002$, $d = 0.82$, conditions. Critically, there was a significant interaction, $F(2, 46) = 8.53$, $MSE = .26$, $p = .001$, $\eta_p^2 = .27$. Entire-set averaging was similar across conditions, $F(2, 46) = 1.77$, $MSE = .405$, $p = .183$, $\eta_p^2 = .07$, whereas subset mean reports were more erroneous in the postcue ($M = 4.37$, $SD = 1.12$) than in the precue ($M = 3.32$, $SD = 0.65$), $p < .001$, $d = 1.43$, and single-set ($M = 3.38$, $SD = 0.83$) conditions, $p = .001$, $d = 0.86$. Thus, worse performance in the postcue condition for perceptual-group mean estimations indicated that concurrent averaging of subsets came at a cost. On the other hand, entire-set mean representations were not affected by the cuing manipulation.

We predicted that if entire-set averaging is automatic, people might rely on the entire-set summary whenever the perceptual subset representations were less precise. To explore whether there was any bias toward the entire-set mean when

³ All reported pairwise and one-sample tests were Bonferroni-corrected for multiple comparisons throughout the study.

Fig. 2 Graph depicting accuracy for the conditions of (a) Experiment 1 ($N = 24$); (b) Experiment 2 ($N = 26$); (c) Simulation 1, which picked either one random item from the target set or averaged two random items from

two sets in the single condition ($N = 24$); and (d) Simulation 2, which picked one random size from the whole set ($N = 24$). Error bars indicate .95 confidence intervals (CIs). * $p = .04$, ** $p \leq .003$

a subset mean was asked for, we computed a bias score. For each trial for a participant, the difference between the reported and correct subset means was divided by the difference between the reported subset mean and entire-set mean, and the absolute value of the result was taken. A bias toward the entire-set mean instead of the correct subset mean would result in this value being greater than 1.0. Values less than 1.0, would mean that the reported value was closer to the correct value than to the entire-set mean. For each participant, we averaged the bias scores of each subset trial across all cueing conditions. A one-sample t test with a test value of 1.0 showed that the subset mean estimations were significantly biased toward the entire-set mean in the postcue ($M = 1.59$, $SD = 0.66$), $t(23) = 4.36$, $p < .001$, $d = 0.89$, but not in the precue ($M = 1.08$, $SD = 0.4$) condition (Fig. 3a). The bias difference between the precue and postcue conditions was also significant, $t(23) = 3.74$, $p = .001$, $d = 0.8$. Therefore, participants' responses were influenced significantly more by the available entire-set means in the more erroneous postcue than in the precue condition.

Subset averaging in the UVF and the LVF To determine whether statistical summaries were better in the UVF than in the LVF, a 2 (Cueing Condition: precue, postcue) $\times$ 2 (Subset VF: UVF, LVF) repeated measures ANOVA was conducted on the absolute errors of subset trials (Fig. 4a). We observed a main effect of cueing condition, $F(1, 23) = 36.7$, $MSE = .72$, $p < .001$, $\eta_p^2 = .61$. As we reported in the previous analysis, more errors occurred in the postcue than in the precue condition. Critically, we found a main effect of subset VF, $F(1, 23) = 7.05$, $MSE = .17$, $p = .014$, $\eta_p^2 = .23$. Contrary to our predictions, we found greater errors in the UVF ($M = 3.96$, $SD = 0.9$) than in the LVF ($M = 3.73$, $SD = 0.76$). There was no significant interaction, $F(1, 23) = 1.12$, n.s.

To determine whether there was greater bias toward the entire-set mean when people were performing less accurately—that is, in the UVF trials—we ran a 2 (Cueing Condition) $\times$ 2 (Subset VF) within-subjects ANOVA on the bias scores (Fig. 5a). Indeed, stronger bias was apparent in the UVF ($M = 1.43$, $SD = 0.47$) than in the LVF ($M = 1.24$, $SD = 0.44$) trials, $F(1, 23) = 12.55$, $MSE = .07$, $p = .002$, $\eta_p^2 = .35$. A main effect

Fig. 3 Graphs depicting bias toward the entire-set mean in the subset trials of (a) Experiment 1 ($N = 24$), (b) Experiment 2 ($N = 26$), and (c) the random-observer simulations. Error bars indicate .95 CIs. * $p \leq .003$

of cueing condition, $F(1, 23) = 14.46$, $MSE = .45$, $p = .001$, $\eta_p^2 = .39$, reflected the previously found bias only in the postcue condition. The interaction did not reach significance, $F < 1$, n.s.

Random observer simulations We carried out two separate simulations with two related goals in mind. First, we used the simulations to determine a baseline-level performance

for the open-ended continuous-response task we used. Second, we wanted to ensure that the responses reflected ensemble perception rather than random reproduction of a single item from the circle display. In the first simulation (Simulation 1, from here onward), each observer was assumed to attend to the two perceptual sets in their respective visual fields and to randomly choose one of the circles from each set. Then,

Fig. 4 Graph depicting accuracy for the subsets in the upper visual field (UVF) and the lower visual field (LVF) of (a) Experiment 1 ($N = 24$) and (b) Experiment 2 ($N = 26$). Error bars indicate .95 CIs. * $p = .014$

Fig. 5 Graphs depicting bias toward the entire-set mean for subsets in the upper visual field (UVF) and the lower visual field (LVF) of (a) Experiment 1 ($N = 24$) and (b) Experiment 2 ($N = 26$). Error bars indicate .95 CIs. * $p = .012$, ** $p = .002$

simulated participants reported the randomly chosen individual size from the cued subset. If the trial cued to give the entire-set mean, the simulated participants responded with the average size of the two randomly chosen circles from each field. Hence, the simulation was not sensitive to the different attentional demands across the precue and postcue conditions. In the single-set condition, simulated participants responded by selecting one random circle from the whole display. In the second simulation (Simulation 2), across all three conditions, each simulated participant was assumed to randomly choose a circle size from all possible ones presented in a given display.

Both simulations used the exact same circle sets as in Experiment 1. As in Experiment 1, there were 24 simulated participants. To achieve enough approximations to normal distributions, 500 such simulated experiments were run for each simulation. We believe the presented values were valid for this sample size and were independent of the number of experiments, given that the number of samples, 500, was sufficiently large. These simulations allowed us to achieve empirically defined null sampling distributions to be used as a baseline/floor performance, in which participants would engage in a simple strategy of randomly selecting an individual circle. It should be noted that these simulations are simplistic in that they neither incorporate representational noise nor model a noisy averaging or response selection process. Therefore, the simulated errors are likely to yield a conservative level of error, possibly representing the lower bound of error.

We calculated the means and 95% confidence intervals (CIs) for error based on the simulations. There was much greater error in both Simulations 1 (Fig. 2c) and 2 (Fig. 2d) than in the Experiment 1 data (Fig. 2a); the actual errors were lower than and outside the 95% CIs of the simulated data. For instance, the error for the 8-circle-set trials was higher in Simulation 1 than in the actual data ($M = 5.89$, 95% CIs [5.76, 6.05] vs. $M = 3.38$, 95% CIs [3.03, 3.73], respectively). The difference between the simulated and the actual data was

even larger for the 16-circle sets ($M = 7.65$, 95% CIs [7.47, 7.81] vs. $M = 3.58$, 95% CIs [3.15, 4.0], respectively). A similar finding emerged in the comparison of Simulation 2 and the actual data. Overall, the simulated data did deviate from the actual data. Despite the simplistic yet conservative nature of the simulations, the much higher error levels in the simulations suggested that our viewers were not randomly reporting individual items, but rather were integrating multiple sizes in their reports. Also, for both sets of simulated data, the patterns of findings were different from the observed one, independent of the error levels. This further shows that ensemble coding occurred at different levels.

In the simulated data, we also calculated bias toward the entire-set mean scores. As can be seen in Fig. 3c, both simulations resulted in significant biases toward the entire-set means in subset trials, $M = 1.54$, 95% CIs [1.43, 1.66], for Simulation 1, and $M = 1.91$, 95% CIs [1.81, 2.02], for Simulation 2. This suggests that whenever a response was within the presented size range, a bias toward the entire-set mean was possible. However, the actual data suggested that viewers did not always respond in this fashion. Particularly, in the precue condition (Fig. 3a), there was no bias toward the entire-set mean, suggesting that these responses reflected pooled rather than randomly chosen sizes. Also, the fact that the bias was stronger in Simulation 2 than in Simulation 1 suggests that the possible inclusion of an item from the uncued set is likely to increase the bias.

In sum, Experiment 1 demonstrated that people can efficiently summarize an entire set and that this process is not impacted by attentional modulation. Thus, entire-set averaging seems to operate in a rather automatic fashion. However, we demonstrated that averaging of perceptual subsets was associated with costs: Postcue performance was significantly worse than single-set and precue performance, suggesting that people were not concurrently averaging the perceptual subsets. The fact that the empirical results differed from the

simulations suggests that viewers were pooling sizes in order to generate a response rather than picking a particular size from the studied display. This suggests that different levels of automaticity might operate in the ensemble perception of single and multigroup sets. Although the comparison of performance across UVF and LVF revealed a pattern contrary to our expectations based on similar phenomena (e.g., crowding), this difference allowed us to test our prediction that people would be pulled toward the entire-set summary when perceptual summaries were less accurate. Indeed, we confirmed this prediction through the greater bias in the postcue than in the precue condition, and in the upper than in the lower circle set.

Experiment 2

Experiment 1 demonstrated that although summarizing entire displays can be rather automatic, there are costs associated with the concurrent processing of perceptual sets. Furthermore, these costs were eliminated by attentional precues, suggesting that viewers could covertly attend to the target set. However, in Experiment 1 we did not control whether viewers were overtly shifting their attention to the target perceptual set in the precue condition. Since the precue was 100% valid and was presented for 500 ms, it was possible that viewers selectively attended to one region in the precue trials and summarized those items only. In Experiment 2, we wanted to reduce the likelihood that participants would rely on overt shifts of attention in the precue condition and that they were instead covertly attending to the target regions while maintaining central fixation. To ensure this, we reduced the duration of the precue to 150 ms, incorporated central catch trials, added controlled periods for eye movements between the trials, and instructed participants on the importance of maintaining fixation. We also told them we were monitoring their eyes with the cameras fixed on top of the monitors, although we actually were not. The catch trials were intermixed with the original set of trials (20% of the trials) and required participants to respond to the sudden and brief appearance of a central blue square (2,000 ms) as quickly as possibly by hitting the B key. These catch trials served as a manipulation check and helped us identify participants who were not engaging in central fixation. If overt attention was driving the results in Experiment 1, then the difference between the precue/single-set and postcue conditions was expected to become smaller and even to disappear with these modifications. However, if the results of Experiment 1 were due to covert shifts of attention, then the precue benefit for perceptual subsets should be replicated despite these experimental modifications.

Method

Participants The participants were 35 Bogazici University undergraduates (18 female, 17 male; $M_{\text{age}} = 21.09$ years, age range: 19–29 years) who took course credit in exchange for their participation. All reported normal or corrected-to-normal vision. In the initial Shapiro–Wilk normality tests, the precue condition turned out not to be normally distributed, $p = .04$. Thus, stem-and-leaf and boxplots were generated for the conditions. Three participants were detected as outliers. When they were discarded, all conditions were normally distributed, all $ps \geq .18$. Thus, the data of these three participants were not included in the remaining analyses, and the reported results reflect the data of 32 participants (16 female, 16 male; $M_{\text{age}} = 20.97$ years).

Materials and stimuli In Experiment 2, we used the same stimuli as in Experiment 1. However, to minimize eye movements, we introduced the following changes to the cued mean size estimation task (see Fig. 1b). Critically, we reduced the duration of the precues to 150 from 500 ms. The precues were presented along with the red plus sign at fixation. To make the conditions more comparable, we also matched the timeline of trials across the precue, postcue, and single-set conditions; in all three conditions, a red fixation plus sign was visible for 650 ms. To allow participants a chance to rest their eyes in a controlled fashion, during the 1,000-ms intertrial interval, a green fixation plus sign was shown on a blank screen. This green fixation plus sign changed to red and remained on the screen for 650 ms in order to alert the participant that the trial was about to begin. To further encourage central fixation, we also intermixed a set of catch trials, in which observers were asked to respond to the sudden appearance of a blue square instead of a circle set. On 20% of the trials, participants were briefly presented with a blue square at the center of the screen (200 ms). They were instructed to make a keypress as quickly as possible. This keypress directly changed the display to the intertrial interval screen. If participants made no response, the display was changed to the intertrial interval screen after a 5-s delay. The catch trials were evenly substituted for the subset/8-circle-set and entire-set/16-circle-set trials and the UVF and LVF trials in each condition. That means that equal numbers of trials were replaced with catch trials from each level.

Procedure To emphasize central fixation to the participants, the instructions were revised. Specifically, we explicitly instructed participants not to take their eyes off of the screen center until they saw the green fixation cross at the end of each trial. Participants were further told that fixating their eyes on the center would increase their success in this task. Moreover, we warned them that we were monitoring their eyes with the cameras on the computer desk, although we were not.

Results and discussion

Catch trials Catch trials were added in order to check whether or not people were focusing on the screen center. We measured the reaction time (RT) to the blue square appearing at the screen center in catch trials. Six of the observers had missing data. Five of these did not respond to the blue square in their first block, and the other participant did not respond during any of the catch trials. We took this as an indicator that these participants may not have been centrally fixating. Therefore, we excluded their data from the remaining analyses, and the results thus reflect the data from 26 participants. No RT differences were found across the precue, postcue, and single-set conditions, $F = 1.7$, n.s., suggesting that observers were equally fast at detecting a central target and equally likely to maintain central focus across the three conditions.

We also looked at the precue condition more closely. In the precue condition, prior to the appearance of the blue square, participants had already been cued to attend to either the entire set or a particular visual field. Comparison of the RTs across these two conditions revealed that observers were quicker to detect the blue square when they were precued for the entire set ($M = 405.84$, $SD = 175.64$) than when they were precued for one of the subsets ($M = 449.51$, $SD = 182.78$), $t(25) = 3.1$, $p = .005$, $d = 0.61$. This RT difference, along with the uniform performance across the three conditions, suggests that in the precued trials there may have been some covert orienting to the cued region following the 150-ms precue. Such anticipatory orienting of attention following valid cues is a well-documented phenomenon (e.g., Posner, 1980).

Entire-set and subset averaging All dependent measures were the same as in Experiment 1. As can be seen in Fig. 2, comparison of the mean errors for all conditions (Fig. 2b) with those in Simulations 1 (Fig. 2c) and 2 (Fig. 2d) confirmed that, as in Experiment 1, the participants engaged in ensemble perception and did not randomly select an item from either the cued or the entire region. The amount of error observed in this experiment was outside the 95% CI of the simulated results. A 3 (Condition: precue, postcue, single set) $\times$ 2 (Set Type: subsets/8-circle sets, entire sets/16-circle sets) repeated measures ANOVA on absolute errors (Fig. 2b) showed that there was no main effect of set, $F < 1$, n.s. We did observe a main effect of condition, $F(2, 50) = 6.81$, $MSE = .75$, $p = .002$, $\eta_p^2 = .21$. Pairwise comparisons showed that the postcue condition ($M = 4.27$, $SD = 0.82$) resulted in more erroneous mean estimates than did the single-set condition ($M = 3.65$, $SD = 0.94$), $p = .006$, $d = 0.68$. The error in the precue condition ($M = 4$, $SD = 0.94$) was not significantly different from the errors in the postcue, $p = .28$, or the single-set, $p = .17$, conditions. More critically, as in Experiment 1, the condition by set interaction was significant, $F(2, 50) = 4.5$, $MSE = .18$, $p = .016$, $\eta_p^2 = .15$. The entire-set mean estimates were equally accurate

across all three conditions, $F(2, 50) = 2.67$, $MSE = .45$, $p = .079$, $\eta_p^2 = .1$. This pattern replicated the Experiment 1 results. Entire-set averaging was not influenced by the presence of perceptual subsets (i.e., no difference between single set and any of the cueing conditions) or by different attentional demands (i.e., no difference between the precue and postcue conditions). In the subset/8-circle-set trials, however, there were more errors in the postcue ($M = 4.39$, $SD = 1.04$) than in both the single-set ($M = 3.54$, $SD = 0.92$), $p = .003$, $d = 0.76$, and precue ($M = 3.91$, $SD = 0.89$), $p = .04$, $d = 0.52$, conditions. As in Experiment 1, the difference between the precued subset mean estimates and the 8-circle single-set mean estimates was not significant, $p = .125$, suggesting that summarizing perceptual groups still benefited from the availability of precues, even when central fixation was strongly encouraged. More importantly, the least accurate mean estimations were again in the postcue condition, suggesting a cost for simultaneous subset averaging. Thus, promoting central fixation in Experiment 2 did not alter the pattern of results, indicating that viewers were covertly attending to the displays.

We analyzed bias scores to determine the tendency of the subset reports to be pulled toward the entire-set mean (see Fig. 3b). One-sample t tests with the criterion set to 1 revealed significant bias in both the precue ($M = 1.36$, $SD = 0.51$), $t(25) = 3.63$, $p = .003$, $d = 0.71$, and postcue ($M = 1.69$, $SD = 0.55$), $t(25) = 6.49$, $p < .001$, $d = 1.27$, conditions. However, this bias was stronger in the more erroneous postcue condition than in the precue condition, $t(25) = 3.48$, $p = .002$, $d = 0.68$. The emergence of a significant and yet smaller bias toward the entire set in the precue condition in Experiment 2 may have been partly driven by an attempt to minimize overt shifts of attention to the cued perceptual group. Indeed, in the second experiment, we observed more error in the precue perceptual-subset trials ($M = 3.91$ [95% CI: 3.55–4.27]) than in the first experiment ($M = 3.32$ [95% CI: 3.05–3.59]).

Subset averaging in the UVF and the LVF To explore the accuracy of the mean representations along the vertical meridian, we ran a 2 (Cueing Condition: precue, postcue) $\times$ 2 (Subset VF: LVF, UVF) repeated measures ANOVA (Fig. 4b). The main effect of cueing condition was significant, $F(1, 25) = 6.96$, $MSE = .87$, $p = .014$, $\eta_p^2 = .22$, reflecting worse subset averaging performance in the postcue than in the precue condition. The subset VF and interaction effects were not significant, both F s < 1 , n.s. That is, unlike in Experiment 1, the subset mean estimations were not affected by the location of the subset when central focus was encouraged. Nevertheless, a 2 (Cueing Condition) $\times$ 2 (Subset VF) within-subjects ANOVA (Fig. 5b) on bias scores still revealed a main effect of subset VF, $F(1, 25) = 7.26$, $MSE = .14$, $p = .012$, $\eta_p^2 = .23$. As in Experiment 1, subsets in the UVF ($M = 1.63$, $SD = 0.45$) resulted in more biased estimations than did those in the LVF ($M = 1.43$, $SD = 0.56$). Cueing condition was

also significant, $F(1, 25) = 12.09$, $MSE = .23$, $p = .002$, $\eta_p^2 = .33$, again reflecting the previously reported stronger bias in the postcue than in the precue condition. There was no interaction, $F < 1$, n.s. When central fixation was emphasized, covert attentional orienting may not have been sufficient to yield an accuracy advantage, although it still showed more of a bias in the UVF. It should be noted that these experiments could not strongly explain the visual field results, and a more controlled study with an eyetracker should be done to find out the mechanism behind these visual field differences.

General discussion

In two experiments, we investigated how spatial grouping affects mean size estimations for perceptual groups and entire sets. We replicated the general pattern of findings across two experiments, demonstrating that people can automatically extract entire-set means (Ariely 2001; Attarha et al., 2014), even when the sets are composed of perceptual groups (Brady & Alvarez, 2011; Oriet & Brand, 2013). Similar performance for the entire-set mean across cueing conditions suggested that entire-set averaging was not influenced by concurrent subset averaging. A secondary task had previously been used in the literature to detect the automaticity level of ensemble perception. These studies had concluded that ensemble coding is not influenced by a concurrent task when the secondary task required distributed or global attention (Alvarez & Oliva, 2009; Chong & Treisman, 2005a). The secondary subset-averaging task in our study required distributed attention, and consistent with the literature, this secondary task had no impact on the summarization of the entire display. However, we further showed that a cost was associated with the extraction of spatial group means simultaneously. Specifically, the error for subsets was greater in the postcue than in the precue and single-set conditions, demonstrating that averaging perceptual groups was affected by the increased attentional demands in the postcue condition and was not as efficient as averaging entire sets. The significant bias toward the entire-set mean when the perceptual-group average was asked for suggested that the former is a robust, easily available representation. Furthermore, this bias was most apparent for the more erroneous perceptual set representations. In both experiments, the bias was significantly stronger in the postcue than in the precue condition. Also, the bias was stronger for more erroneous UVF responses in Experiment 1. We also demonstrated that effective precueing reduced errors for perceptual-group summary and mitigated this bias (e.g., via overt attention in Exp. 1). The bias toward the entire-set mean especially when responses were erroneous complements the finding of no difference between entire-set summaries across all conditions in both experiments. Both pieces of evidence together support the argument that entire-set means are extracted automatically

and are easily available to guide or interfere with other ongoing processes, especially when these processes are operating at lower levels of efficiency. Moreover, encouraging central fixation in Experiment 2 did not change the main findings suggesting that viewers were overtly attending to cued regions.

Our bias results, suggesting that ensemble summaries bias other ongoing processes, are also consistent with previous findings from the literature. Brady and Alvarez (2011) had also shown that people utilized statistical summaries when they had imprecise item representations. However, the fact that bias was more pronounced in the UVF, even though the mean representations at the upper and lower visual fields were equally precise in Experiment 2, pointed out that bias may not be directly related to the precision of the representations. Instead, one might speculate that the stronger bias may reflect viewers' uncertainty about the fidelity of their representations. In fact, this interpretation would also be in line with the finding that precued subset averages were significantly biased toward the entire-set mean in Experiment 2, whereas there was no significant bias in such trials in Experiment 1. That is to say, keeping central fixation in Experiment 2 might have increased the uncertainty of the subset representations in precued trials and thus resulted in a significant bias. Likewise, Dubé, Zhou, Kahana, and Sekuler (2014) previously found that averaging impacts spatial frequency discriminations especially when people were uncertain about the precision of their memory representations. Also, Oriet and Brand (2013) showed that mean length reports were biased toward the mean of the task irrelevant set. These findings from the literature together with the bias findings reported here points to an automatic formation of ensemble representations for the whole without intention and its greater impact on performance when uncertainty is most.

Alternative interpretations of these bias results are possible, yet unlikely. First, observers might regress toward the center of the possible response range; in these experiments, such a strategy would have resulted in responses that were closer to the entire-set mean. However, in both experiments, the response range was very large and the size in the middle of the response range (240-pixel/12° diameter), was much larger than the sizes of any of the circles used in our displays (20-pixel/1° to 84-pixel/4.2° diameter). Second and in a related vein, participants might bias to the mean of the initially given test circle's range. The test circle was 15% to 71% different from the correct answer. If participants become sensitive to this range, then their reports should be biased towards the size that was 43% larger or smaller than the correct answer. To make sure that participants were not reporting the average of the test-circle range, we calculated a bias score toward the entire-set mean over the test circle range (i.e., absolute value of difference to the test circle's range mean divided by difference to the entire-set mean) in Experiment 2. In fact, in all

conditions, viewers were pulled disproportionately toward the entire-set mean instead of the average range of the test circle. In addition, the pattern of results was similar with the previously reported results of bias toward entire-set mean instead of the correct response. Therefore, we believe that even if it was possible that people might have utilized from the range of possible test circle sizes, their reports were still pulling toward entire-set mean.

A third alternative regarding the pattern of bias of results might have been driven by participants selecting a subset to average and inferring the other subset on the basis of the selected one and their knowledge of the entire-set mean. The difference between the entire-set and subset means were at least 20% (and the difference between the two subsets at least 40%). Even though this was not explicitly stated to the participants, they may have nevertheless figured this out implicitly. If they relied on the “chosen” subset to generate their response, their responses should be highly accurate for the matching trials (in which the chosen one was asked for) and inaccurate for the nonchosen and yet cued subsets. The overall performance, then, would be expected to bias toward the entire-set mean. However, we believe that this was unlikely. Regression analyses with accuracy as the predictor of the postcue bias in Experiment 1 revealed that accuracy was not a significant predictor, $R^2 < .001$. Furthermore, this strategy would also indicate that participants would bias toward the chosen subset’s mean in their entire-set representations. The question is which subset participants were likely to choose to average. The possible candidate seemed the subset in the LVF in Experiment 1 since their representations were more precise for that than subsets in the UVF. When we analyzed entire-set trials in the precue and postcue conditions, there was not significant bias toward the mean of the subset in the LVF, both p s $\geq .78$. Thus, we concluded that participants did not attend only to the subset in the LVF as a general strategy.

One final explanation for the bias findings is that the observed bias results were driven by the specific stimuli characteristics. In this study, the mean of the entire-set mean was 20% different than each of the subset means. Because the entire-set mean had to be the average of the subset means, it was always closer to the middle of the range of all the sizes in a given display. Thus, any response within this range would naturally be pulled toward the entire-set mean. This reasoning is also in line with the random observer simulations in which simulated responses were chosen from the sizes in the screen and necessarily within the size range. Both simulations resulted in significant biases toward the entire-set mean in subset trials. Thus, it is possible that the significant bias may just be reflecting an awareness of the size range in a given display. Indeed, the range is also a statistical summary that can be readily extracted from an ensemble. Ariely (2001) reported that participants successfully rejected a test item that was outside the size range as not belonging to the studied set. Chetverikov, Campana, and

Kristjánsson (2016, 2017) showed that participants’ search performances were influenced by the distributional characteristics of the distractors, including the range, as well as the mean of the distractors. Furthermore, studies on outlier items in an ensemble demonstrated that outliers are excluded from the summary representations (Haberman & Whitney, 2010) and they were better represented than nonoutlier items (Yildirim & Boduroglu, 2018). Thus, viewers are sensitive to the range of values in a given ensemble. However, the simulation results also suggested that viewers were also specifically sensitive to the presence of perceptual sets. For instance, in Simulation 1, in which the random observer was constrained to randomly pick from a group, the bias scores were significantly less than that in Simulation 2, which relied on fully random picks from the entire set (see Fig. 3c). This difference demonstrates that the inclusion of a size from the other group may be sufficient to increase the bias toward the entire-set mean. Therefore, we argue that the bias results do not merely reflect an awareness of the range of the stimuli but rather the prioritization of certain perceptual groups over others and relatively more reliance on the entire set.

Our study was important for clarifying the literature on the simultaneous processing of perceptual-group means. Two studies in the literature utilizing a similar cueing paradigm had yielded conflicting results. In both studies, participants were able to average subsets concurrently above the chance level (Brand et al., 2012; Chong & Treisman, 2005b). Chong and Treisman (2005b) demonstrated that participants’ performances for color-coded subsets were not influenced from pre- or postcueing. However, Brand et al. showed that performance decreased with postcues; thus, there was a cost of concurrent averaging. Our study was different in that we explored the performance for averaging the entire set in addition to the subsets constituting the entire set. We showed that these two averaging processes could be separable. Mean size of entire sets were established with similar precision across conditions. Thus, mean size of a set was automatically formed regardless of whether the set was composed of groups or not. Critically, in the postcue condition, entire-set averaging was not impacted by additional attentional demands. On the other hand, mean size estimations of subsets in the postcue condition were less accurate than in the precue and single set conditions. This suggested that observers disregarded the spatially segregated groups while averaging the entire set and may not have been processing those groups separately. This conclusion is in line with the findings of Brand et al. Additionally, we furthered Brand et al. by demonstrating this cost with spatially segregated groups rather than color-coded groups. Comparison of our data with the random observer simulations suggested that people were in fact integrating multiple sizes in their mean responses in all of the conditions. Hence, it is possible that people were averaging parts of separate perceptual groups simultaneously, but with less efficiency.

Whereas Experiment 1 revealed that performance was better in the LVF, in Experiment 2 no difference was found between the LVF and UVF. Interestingly, in both experiments there was a stronger bias toward the entire-set mean in the UVF than in the LVF, suggesting that factors others than crowding driven by vertical meridian asymmetry may have been at play here. Actually, our overall findings replicated Bulakowski et al. (2011), who failed to show any performance difference between the UVF and the LVF when they asked observers to report mean orientation of tilted lines. However, one must approach the similarities in conclusion with caution given the task differences between our study and theirs. First, whereas we asked for mean size reports, Bulakowski et al. used mean orientation reports. Second, they tested performance on a two-alternative forced choice task, whereas we collected data using a continuous report measure. Most importantly, Bulakowski et al. presented a set in either the UVF or the LVF. We, however, presented two separate sets, one of which was in the UVF and the other in the LVF. But from all of these conditions, our precue condition in Experiment 1 was most similar to the conditions in the Bulakowski et al. study, because viewers were allowed to make eye movements and fixate on the precued subset in Experiment 1. Therefore, in the precue condition of Experiment 1, viewers might have selected one visual field and ignored the other subset more easily. Indeed, in Experiment 1 there was no difference in the accuracy of the precued perceptual-group estimates presented in the UVF ($M = 3.37$, $SD = 0.75$) and LVF ($M = 3.27$, $SD = 0.65$), $t(23) = 0.86$, $p = .4$, $d = 0.18$. These findings provide preliminary data suggesting that pooling mechanisms and crowding may not operate by the same rules, and that there may be no effect of vertical meridian asymmetry. Further research, targeted specifically at understanding the crowding–pooling relationship, will be necessary.

All together, these findings tell us that statistical summaries are formed with different levels of efficiency across entire-set and perceptual-group levels. Whereas processing perceptual groups as efficient ensembles requires attention, processing the whole does not. Once again, these point to different levels of efficiency at different levels of ensembles. The impact of attention on ensemble processing has been studied in the literature. Whereas some have found that formation of the ensemble representations could be formed while engaging in another task requiring distributed attention (Alvarez & Oliva, 2009; Bronfman, Brezis, Jacobson, & Usher, 2014; Chong & Treisman, 2005a), others have found that changes in ensemble characteristics went unnoticed in an inattentional blindness paradigm (Jackson-Nielsen, Cohen, & Pitts, 2017). We believe that these results could be explained by different levels of efficiency such that some summary information might be more efficiently drawn out than others. In a related vein, emerging literature suggests that there may be more than one mechanism for summarizing visual displays and that these

mechanisms may operate independently, and may be impacted by different factors (Haberman et al., 2015). Finally, the more pronounced bias in summary representations under less efficient conditions suggested that people may rely on available summary representations—that is, the entire-set mean or the range—to alleviate any uncertainty arising from concurrent averaging or lower attentional resolution.

Although we believe that our study has increased the understanding of how perceptual groups are summarized along with entire sets, our study may be limited in that we only tested summarization in spatially segregated perceptual groups. It is possible that the perceptual groups in our study were not sufficiently separable perceptually. Because we wanted to include a larger number of items of varying sizes, we were not able to maximally segregate the two perceptual groups spatially. However, we also ensured that the means of the subsets differed by at least 40%, which resulted in overall size differences across perceptual sets. This overall size similarity along with the spatial segregation could be argued to reinforce perceptual grouping (Brooks, 2015). Another important point is that we examined perceptual-group averaging in the same domain—that is, size, and with the same objects. Nevertheless, the natural world has a lot of variability in textures of objects, their identities, and their placement. Thus, perceptual groups may be formed in more complex ways in the real world and groups may be summarized on the basis of unique salient features. Future research is necessary to determine whether greater perceptual set segregation would result in greater efficiency in the processing of perceptual sets.

Author note This project is supported by a grant from The Scientific and Technological Research Council of Turkey (TÜBİTAK, 113K217) and a BAGEP award granted to A.B. We would like to thank Harun Yörük for his assistance in data collection. Also, we thank anonymous reviewers for their constructive comments.

References

- Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world extracting statistical summary representations over time. *Psychological Science*, 21, 560–567. <https://doi.org/10.1177/0956797610363543>
- Albrecht, A. R., Scholl, B. J., & Chun, M. M. (2012). Perceptual averaging by eye and ear: Computing summary statistics from multimodal stimuli. *Attention, Perception, & Psychophysics*, 74, 810–815. <https://doi.org/10.3758/s13414-012-0293-0>
- Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2014). Obligatory averaging in mean size perception. *Vision Research*, 101, 34–40.
- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15, 122–131. <https://doi.org/10.1016/j.tics.2011.01.003>
- Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological*

- Science*, 19, 392–398. <https://doi.org/10.1111/j.1467-9280.2008.02098.x>
- Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, 106, 7345–7350. <https://doi.org/10.1073/pnas.0808981106>
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12, 157–162.
- Attarha, M., & Moore, C. M. (2015). The capacity limitations of orientation summary statistics. *Attention, Perception, & Psychophysics*, 77, 1116–1131.
- 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, 1440–1449. <https://doi.org/10.1037/a0036206>
- Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *Psychological Record*, 59, 171–186.
- Ben-Av, M. B., & Sagi, D. (1995). Perceptual grouping by similarity and proximity: Experimental results can be predicted by intensity autocorrelations. *Vision Research*, 35, 853–866.
- Boduroglu, A., & Shah, P. (2014). Configural representations in spatial working memory. *Visual Cognition*, 1, 997–1007.
- Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22, 384–392. <https://doi.org/10.1177/0956797610397956>
- Brand, J., Oriet, C., & Tottenham, L. (2012). Size and emotion averaging: Costs of dividing attention after all. *Canadian Journal of Experimental Psychology*, 66, 63–69.
- Bronfman, Z. Z., Brezis, N., Jacobson, H., & Usher, M. (2014). We see more than we can report: “Cost free” color phenomenality outside focal attention. *Psychological Science*, 25, 1394–1403.
- Brooks, J. L. (2015). Traditional and new principles of perceptual grouping. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization* (pp. 57–87). Oxford, UK: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199686858.013.060>
- Brown, T. L., Gore, C. L., & Carr, T. H. (2002). Visual attention and word recognition in Stroop color naming: Is word recognition “automatic”? *Journal of Experimental Psychology: General*, 131, 220–240. <https://doi.org/10.1037/0096-3445.131.2.220>
- Bulakowski, P. F., Post, R. B., & Whitney, D. (2011). Reexamining the possible benefits of visual crowding: Dissociating crowding from ensemble percepts. *Attention, Perception, & Psychophysics*, 73, 1003–1009.
- Cameron, E. L., Tai, J. C., & Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision Research*, 42, 949–967. [https://doi.org/10.1016/S0042-6989\(02\)00039-1](https://doi.org/10.1016/S0042-6989(02)00039-1)
- Chetverikov, A., Campana, G., & Kristjánsson, Á. (2016). Building ensemble representations: How the shape of preceding distractor distributions affects visual search. *Cognition*, 153, 196–210.
- Chetverikov, A., Campana, G., & Kristjánsson, Á. (2017). Representing color ensembles. *Psychological Science*, 28, 1510–1517.
- Chong, S. C., Joo, S. J., Emmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics*, 70, 1327–1334. <https://doi.org/10.3758/PP.70.7.1327>
- Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, 43, 393–404. [https://doi.org/10.1016/S0042-6989\(02\)00596-5](https://doi.org/10.1016/S0042-6989(02)00596-5)
- Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, 67, 1–13. <https://doi.org/10.3758/BF03195009>
- Chong, S. C., & Treisman, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, 45, 891–900.
- Choo, H., & Franconeri, S. L. (2010). Objects with reduced visibility still contribute to size averaging. *Attention, Perception, & Psychophysics*, 72, 86–99. <https://doi.org/10.3758/APP.72.1.86>
- Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica*, 138, 289–301. <https://doi.org/10.1016/j.actpsy.2011.08.002>
- Corbett, J. E., & Song, J. H. (2014). Statistical extraction affects visually guided action. *Visual Cognition*, 22, 881–895.
- Coren, S., & Girgus, J. S. (1980). Principles of perceptual organization and spatial distortion: The gestalt illusions. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 404–412.
- Dakin, S. C., Bex, P. J., Cass, J. R., & Watt, R. J. (2009). Dissociable effects of attention and crowding on orientation averaging. *Journal of Vision*, 9(11), 28:1–16. <https://doi.org/10.1167/9.11.28>
- de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, 70, 789–794.
- de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology*, 62, 1716–1722. <https://doi.org/10.1080/17470210902811249>
- Dubé, C., & Sekuler, R. (2015). Obligatory and adaptive averaging in visual short term memory. *Journal of Vision*, 15(4), 13:1–13. <https://doi.org/10.1167/15.4.13>
- Dubé, C., Zhou, F., Kahana, M. J., & Sekuler, R. (2014). Similarity-based distortion of visual short-term memory is due to perceptual averaging. *Vision Research*, 96, 8–16.
- Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across future dimensions in statistical processing of perceptual groups. *Perception & Psychophysics*, 70, 946–954. <https://doi.org/10.3758/PP.70.6.946>
- Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology*, 106, 1389–1398.
- Fuller, S., Rodríguez, R. Z., & Carrasco, M. (2008). Apparent contrast differs across the vertical meridian: Visual and attentional factors. *Journal of Vision*, 8(1), 16:1–16. <https://doi.org/10.1167/8.1.16>
- Gmeindl, L., Nelson, J. K., Wiggin, T., & Reuter-Lorenz, P. A. (2011). Configural representations in spatial working memory: Modulation by perceptual segregation and voluntary attention. *Attention, Perception, & Psychophysics*, 73, 2130–2142.
- Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2009). Positional averaging explains crowding with letter-like stimuli. *Proceedings of the National Academy of Sciences*, 106, 13130–13135.
- 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, 432–446.
- Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17, 751–753.
- Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 718–734. <https://doi.org/10.1037/a0013899>
- Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, 72, 1825–1838. <https://doi.org/10.3758/APP.72.7.1825>
- Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, 17, 572–576.
- He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, 383, 334–337. <https://doi.org/10.1038/383334a0>

- Jackson-Nielsen, M., Cohen, M. A., & Pitts, M. A. (2017). Perception of ensemble statistics requires attention. *Consciousness and Cognition*, 48, 149–160.
- Jacoby, O., Kamke, M. R., & Mattingley, J. B. (2013). Is the whole really more than sum of its parts? Estimates of average size and orientation are susceptible to object substitution masking. *Journal of Experimental Psychology: Human Perception and Performance*, 39, 233–244.
- Joo, S. J., Shin, K., Chong, S. C., & Blake, R. (2009). On the nature of the stimulus information necessary for estimating mean size of visual arrays. *Journal of Vision*, 9(9), 7:1–12. <https://doi.org/10.1167/9.9.7>
- Karim, A. K. M. R., & Kojima, H. (2010). The what and why of perceptual asymmetries in the visual domain. *Advances in Cognitive Psychology*, 6, 103–115.
- Kramer, R. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision*, 15(4), 1:1–9. <https://doi.org/10.1167/15.4.1>
- Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica*, 142, 245–250.
- Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision*, 15(4), 6:1–18. <https://doi.org/10.1167/15.4.6>
- Morgan, M., Chubb, C., & Solomon, J. A. (2008). A “dipper” function for texture discrimination based on orientation variance. *Journal of Vision*, 11(12), 13:1–11. <https://doi.org/10.1167/11.12.13>
- Mutlutürk, A., & Boduroglu, A. (2014). Effects of spatial configurations on the resolution of spatial working memory representations. *Attention, Perception, & Psychophysics*, 76, 2276–2285.
- Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, 70, 772–788. <https://doi.org/10.3758/PP.70.5.772>
- Naveh-Benjamin, M. (1987). Coding of spatial location information: An automatic process? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 595–605. <https://doi.org/10.1037/0278-7393.13.4.595>
- Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, 128, 56–63.
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155, 23–36. [https://doi.org/10.1016/S0079-6123\(06\)55002-2](https://doi.org/10.1016/S0079-6123(06)55002-2)
- Oriet, C., & Brand, J. (2013). Size averaging of irrelevant stimuli cannot be prevented. *Vision Research*, 79, 8–16. <https://doi.org/10.1016/j.visres.2012.12.004>
- Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 4, 739–744.
- Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68, 29–46.
- Poltoratski, S., & Xu, Y. (2013). The association of color memory and the enumeration of multiple spatially overlapping sets. *Journal of Vision*, 13(8), 6:1–11. <https://doi.org/10.1167/13.8.6>
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32, 3–25. <https://doi.org/10.1080/00335558008248231>
- Posner, M. I., Sandson, J., Dhawan, M., & Shulman, G. L. (1989). Is word recognition automatic? A cognitive–anatomical approach. *Journal of Cognitive Neuroscience*, 1, 50–60.
- Posner, M. I., & Snyder, C. R. (2004). Attention and cognitive control. In D. A. Balota & E. J. Marsh (Eds.), *Cognitive psychology: Key readings* (pp. 205–223). New York, NY: Psychology Press.
- Previc, F. H. (1990). Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. *Behavioral and Brain Sciences*, 13, 519–575.
- Psychology Software Tools, Inc. (2012). E-Prime (Version 2.0). Pittsburgh, PA: Author. Retrieved from www.pstnet.com
- Semizer, Y., & Boduroglu, A. (2016). *Representing visual variance*. Unpublished manuscript.
- Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, 39, 329–337.
- Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental Science*, 18, 556–568.
- Talgar, C. P., & Carrasco, M. (2002). Vertical meridian asymmetry in spatial resolution: Visual and attentional factors. *Psychonomic Bulletin & Review*, 9, 714–722.
- Teghtsoonian, M. (1965). The judgment of size. *American Journal of Psychology*, 78, 392–402. <https://doi.org/10.2307/1420573>
- Uner, O., Mutlutürk, A., & Boduroglu, A. (2014, November). *The impact of videogame playing on the extraction of statistical summaries in visual displays*. Poster presented at the Annual Meeting of the Psychonomic Society, Long Beach, CA.
- 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>
- Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt: II. *Psychologische Forschung*, 4, 301–350. <https://doi.org/10.1007/BF00410640>
- Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review*, 18, 484–489. <https://doi.org/10.3758/s13423-011-0071-3>
- Yildirim, I., & Boduroglu, A. (2018). *Outlier and ensemble representations of visual displays*. Manuscript under review.
- Yoruk, H., & Boduroglu, A. (2018). *Feature-specificity in visual statistical summary processing*. Manuscript in preparation.