



# An explicit investigation of the roles that feature distributions play in rapid visual categorization

Hee Yeon Im<sup>1,2</sup> · Natalia A. Tiurina<sup>3</sup> · Igor S. Utochkin<sup>3</sup>

© The Psychonomic Society, Inc. 2020

## Abstract

Ensemble representations are often described as efficient tools when summarizing features of multiple similar objects as a group. However, it can sometimes be more useful *not* to compute a single summary description for all of the objects if they are substantially different, for example when they belong to entirely different categories. It was proposed that the visual system can efficiently use the distributional information of ensembles to decide whether simultaneously displayed items belong to single or several different categories. Here we directly tested how the feature distribution of items in a visual array affects an ability to discriminate individual items (Experiment 1) and sets (Experiments 2–3) when participants were instructed explicitly to categorize individual objects based on the median of size distribution. We varied the width (narrow or fat) as well as the shape (smooth or two-peaked) of distributions in order to manipulate the ease of ensemble extraction from the items. We found that observers unintentionally relied on the grand mean as a natural categorical boundary and that their categorization accuracy increased as a function of the size differences among individual items and a function of their separation from the grand mean. For ensembles drawn from two-peaked size distributions, participants showed better categorization performance. They were more accurate at judging within-category ensemble properties in other dimensions (centroid and orientation) and less biased by superset statistics. This finding corroborates the idea that the two-peaked feature distributions support the “segmentability” of spatially intermixed sets of objects. Our results emphasize important roles of ensemble statistics (mean, range, distribution shape) in explicit visual categorization.

**Keywords** Ensemble statistics · Categorization · Segmentation · Mean size · Mean orientation · Centroid

## Introduction

The visual system organizes complex scenes using strategies for forming coherent and concise representations, rather than passively receiving all (millions of) bits of information hitting our retinas at any given moment. One powerful heuristic is

representing sets of similar objects as an ensemble using summary statistics. Ensemble coding provides global information about a group of items in the entire image such as average across multiple dimensions (Alvarez & Oliva, 2008; Ariely, 2001; Bauer, 2009; Chong & Treisman, 2003; Dakin & Watt, 1997; Haberman & Whitney, 2007, 2009), variance (Morgan, Chubb, & Solomon, 2008; Solomon, 2010), or approximate number (Chong & Evans, 2011; Feigenson, Dehaene, & Spelke, 2004; Halberda, Sires, & Feigenson, 2006). Such global information can be extracted through a pooling process across multiple objects. It provides a quick and precise description about the image as a whole, even when the number of objects to be averaged exceeds the cognitive capacity, which is severely constrained by selective attention and working memory systems (e.g., Cowan, 2001; Luck & Vogel, 1997; Pylyshyn & Storm, 1988) and when little conscious access or selective attention to the image is available (e.g., Alvarez, 2011; Alvarez & Oliva, 2008; Alvarez & Oliva, 2009; Ariely, 2001; Corbett & Oriet, 2011; Im & Halberda, 2013; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001).

✉ Igor S. Utochkin  
isutochkin@inbox.ru

- <sup>1</sup> Fetal-Neonatal Neuroimaging and Developmental Science Center, Boston Children’s Hospital, Harvard Medical School, Boston, MA, USA
- <sup>2</sup> Division of Newborn Medicine, Department of Medicine, Boston Children’s Hospital, Harvard Medical School, Boston, MA, USA
- <sup>3</sup> Psychology Department, National Research University Higher School of Economics, 4-2, Armyansky per, Moscow, Russia 101000

Ensemble coding is of great value in our daily perception and cognition of visual scenes. If we look around, we always find some redundancy and regularity in real-world images: Buildings in a city, trees in a forest, and fruit in a bush, for example, are often seen as groups of similar but not identical objects. For most everyday needs, we may not need to store all individuating information from these scenes. We can instead only extract summary statistics of a scene to make sense of the overall layout, pattern, and gist concisely and compactly. Such perceptual ability to extract ensemble representations allows our brain to “do more with less” and better interact with the complex, dynamic visual world. For example, representing and storing an ensemble (e.g., average) of multiple objects helps the visual system to maintain and recall an image better. At the object level, only a few items (up to three or four) can be remembered at a time; the rest may be missed entirely due to the limited memory capacity. When attempting to recall missed objects, one would have to make random guesses. However, a higher-level representation of average extracted from all the objects in the image can guide one to recall the missed object to some extent by retrieving values biased toward the average and reducing the overall expected errors (Brady & Alvarez, 2011; Im & Chong, 2014). Say, you try to remember different colors of six disks in an image and report the colors you remember. If you remember that the disks were in “cool” colors on average (even if you cannot remember the exact colors for each disk), you will likely reduce the overall error by choosing six colors from only the continuum of “cool” colors and avoid “warm” colors. However, if you remember only the colors of three disks and completely missed the others (without remembering average information), you cannot avoid making extreme errors when you recall the color of one of the three forgotten colors (e.g., by randomly choosing “red” for a disk that is turquoise).

Previous studies have demonstrated that multiple sets of objects, up to three or four, can be extracted in parallel, as higher-order units for perception and memory (e.g., Halberda et al., 2006; Im & Chong, 2014; Im, Park, & Chong, 2015). The limit on the number of ensembles that can be extracted and remembered at any given time also converges with the well-documented three-or-four-object limits of visual attention (e.g., Pylyshyn & Storm, 1988) and working memory (e.g., Luck & Vogel, 1997; Zhang & Luck, 2008). Such convergence illustrates how items in an image can be represented hierarchically (e.g., as individuals or ensembles). Results of different units of visual processing can allow complementary information about the image (e.g., local vs. global) to be available to an observer at the same time. In a display of 20 dots in four different colors (five red, five blue, five yellow, and five green dots), for example, an observer can represent 20 individual dots, ensemble features of four colorsets, and even those of a *superset*. Previous work has empirically investigated such a notion of hierarchical coding in

visual perception and memory (Brady & Alvarez, 2011; Corbett, 2017; Halberda et al., 2006; Im & Chong, 2014; Im, Zhong, & Halberda, 2016). This work collectively suggests that the nature of visual representations in perception and memory is constructive, hierarchical, and interactive across multiple levels of abstraction. Hierarchical and constructive visual representation in perception and memory can be made possible when the extraction of ensemble features is as rapid as those of individuals so that both levels of representation are available to interact with each other. Indeed, previous studies have shown that ensembles can be extracted from groups of objects very rapidly (e.g., Im et al., 2016; Leib, Kosovicheva, & Whitney, 2016). The question that remains to be addressed is how the visual system utilizes ensemble features that are rapidly extracted to facilitate perceptual and cognitive processes underlying hierarchical coding of complex, cluttered visual scenes. In the current study, we report three case studies that empirically tested the roles global representation of ensembles created from multiple items in a visual image plays in rapid visual segmentation, categorization, and perceptual grouping of visual arrays. New findings from the current study will provide an insight into how ensemble representations can be extracted from multiple sets of similar objects in a visual array and serve as perceptual bases for hierarchical coding of the scene to make sense of it.

In the context of texture perception and visual search, it has been shown that the visual system can split multiple textures or individual items into clearly distinguishable (e.g., pre-attentive; Julesz, 1981) global subsets; and the roles of various spatial factors are widely discussed as principal determinants of subset formation: local proximity and local contrasts (Bacon & Egeth, 1991; Bravo & Nakayama, 1992; Itti & Koch, 2001; Treisman, 1988; Wolfe, 1994) as well as more global factors, such as abrupt violation of spatial statistics over a region (Nothdurft, 1992, 1993). As soon as spatial interactions take an important part in these models, the explanations for grouping and segmentation strongly rely on well-established mechanisms of space-based, retinotopic interactions akin to lateral inhibition (e.g., Knierim & van Essen, 1992). However, not much of the prior work was done to examine how individual objects can be categorized into discrete, higher-level sets when spatial layout does not support strong organization of similar elements into compact patches lying apart from dissimilar elements. Perceiving a set of apples among leaves and branches is an example showing how common the categorization of spatially intermixed sets can be in real-world perception.

Recently, Utochkin (2015) has suggested that ensemble summary statistics can be a candidate representation that supports the rapid categorization of multiple objects into sets in a visual image. Ensemble representation can be “spatially abstract (or spatially blind in an extreme case),” in a sense that extracted summary statistics do not have to retain an exact

knowledge of how individual elements are located, once created in another feature domain such as size, orientation, and so on. Thus, it can be well suitable for the rapid categorization of spatially intermixed items of different kinds. Human observers appear to be very sensitive to how features of individual items in a visual array are distributed, such that their perceptual ability to segment and discriminate groups of items is systematically influenced by the shape of the distribution of features that are tested (e.g., Chetverikov, Campana, & Kristjánsson, 2016, 2017; Chong & Treisman, 2003; Corbett, Wurnitsch, Schwartz, & Whitney, 2012; Im & Halberda, 2013; Oriet & Hozempa, 2016; Rosenholtz, 2000; Utochkin, Khvostov, & Stakina, 2018; Utochkin & Yurevich, 2016).

According to Utochkin (2015; also see Utochkin & Yurevich, 2016; Utochkin et al., 2018), the central concept that is related to categorization is *segmentability*. Segmentability is derived from the shape of a feature distribution – in particular, from its peaks and pits. If individual features of all presented items smoothly cover the entire range of displayed features, thus forming either no sharp peak (as in the uniform distribution) or a single peak (as in the Gaussian distribution), such an ensemble is non-segmentable and is likely to be categorized as consisting of items of one kind. In contrast, if individual features are distributed unevenly, forming dense clusters (peaks) separated by relatively large gaps within the range, then such an ensemble is segmentable and is likely perceived as consisting of several categorical groups. A similar idea of the distributional difference as the determinant of efficient segmentation was suggested for explaining pop-out visual search (Hochstein, Pavlovskaya, Bonne, & Soroker, 2018; Rosenholtz, 1999, 2001).

In their previous work, Utochkin and colleagues (Utochkin & Yurevich, 2016; Utochkin et al., 2018) manipulated the shape of ensemble distributions to empirically test whether it supported categorical grouping in the manner predicted by the segmentability hypothesis. To test the effects of distribution shape, they used indirect measurements such as proportion correct or response time (RT) using the visual search task or texture discrimination task paradigms. In these task paradigms, higher accuracy or faster RTs were considered to reflect greater segmentability between a target and distractors or between two different texture patches. Yet, they have not explicitly asked or instructed human observers to report whether they perceive objects in a visual array as belonging to the same or different categories based on the perceived feature distributions. Therefore, the first aim of the current study was to examine how human observers use such feature distributions for rapid categorization of individual objects into subsets in a visual image (Experiment 1). The second aim is to examine further how the rapid categorization based on one feature dimension (e.g., size) serves as the basis of segmenting subsets of individual objects to mediate extraction of ensemble summary statistics in another feature dimension (e.g., location).

## Experiment 1

### Rapid categorization by size: Assigning individual objects into subsets in a visual image

In Experiment 1, we first examined how human observers segment objects in an image into categorical subsets relying on feature distributions of the objects. Although previous work has shown that human observers are sensitive to feature distributions in a visual image, none of them has directly tested how feature distribution is utilized when observers perceive “subsets” in the image and categorize individual items into the subsets. Here we tested our hypothesis that human observers can categorize individual objects into two subsets in an image very rapidly, relying on summary statistics about the whole image that describe how individual features of the objects are distributed (e.g., mean, median, variability, and the shape of distribution). We first conducted a simple study using a straightforward and explicit approach by asking participants to categorize an individual item into one of two subgroups in a visual array based on a feature dimension of size (e.g., “does this circle belong to a larger set or a smaller set in this image?”). Although perceptual categorization by size (e.g., small vs. large sets) has not been tested yet in the context of ensemble coding, it is a testable and viable category. It is easy to imagine you are sorting out a pile of apples into two piles, so that you can use the pile of small piles to make an apple pie and the pile of large apples to eat them raw. We examined how participants’ categorization performance was affected by the shapes of feature distributions of individual objects in the image. We predicted that participants’ ability to categorize individual objects into subsets would be systematically varied by the global properties of the feature distribution (e.g., mean, variance, and the smoothness of the distribution presumably affecting ensemble segmentability) and by the individual objects’ relative dispositions in the distribution.

## Method

**Participants** Twenty undergraduate students (13 females; age range: 18–27 years) of the Higher School of Economics took part in the experiment for extra course credits. All reported having normal or corrected-to-normal vision, normal color vision, and no neurological problems. All were naïve as to the purpose of the experiment. Written informed consent was obtained for the experiment from the participants in accordance with the Declaration of Helsinki.

**Apparatus and stimuli** The stimulation was developed and presented via PsychoPy v1.82 for Linux Ubuntu (Peirce et al., 2019) on a standard VGA-monitor (screen diagonal 19 in., 75 Hz refresh rate, resolution of 1,200 × 800 pixels, which was 30.65° × 20.43° in visual angle). Observers responded by

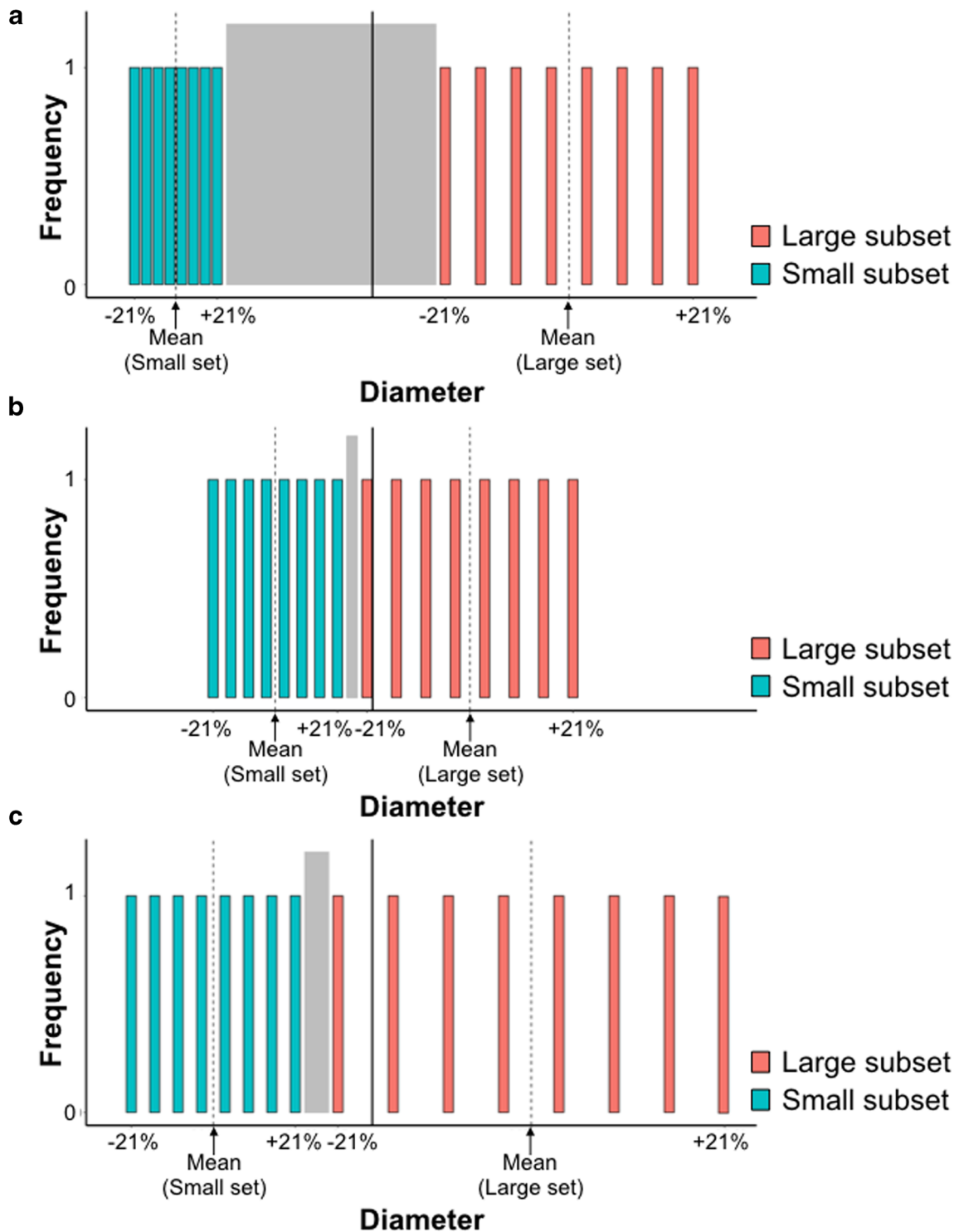
pressing keys on a computer keyboard with their dominant hand.

Each visual stimulus contained a set of 16 white circles (superset) randomly positioned in a gray screen. The distribution of individual sizes of the white circles was divided into two categories (subsets) containing eight items each: “small” (ranging from  $\sim 0.3^\circ$  to  $\sim 1.3^\circ$  in visual angle across all trials) and “large” (ranging from  $\sim 0.7^\circ$  to  $\sim 2.3^\circ$  across the entire experiment). The overlap between the entire ranges of “small” and “large” provided that the categorical belongingness of an item with a particular size could change from trial to trial, therefore encouraging participants to “calibrate” their impression of an item’s category only based on a current trial. Item sizes in one-half of the circles were always smaller than the median of the whole ensemble, and item sizes in the other half of the circles were always greater than the median. The grand mean diameter varied in a broad enough range (from  $0.8^\circ$  to  $1.5^\circ$ ) across trials to ensure that the “categorical boundary” varied unpredictably and observers relied on an impression from the current trial only, rather than the memory of the grand mean diameter of the whole ensemble from previous trials. Individual sizes of the circles could be drawn from one of three distributions varying in range and shape, detailed in the following:

- 1) *Two-peaked distribution*: To generate a size distribution that contains two thin distributions of highly separable subsets, we used the large set:small set mean ratio of 3:1, where the gap between the smallest circle of set 1 and the largest circle of set 2 was as big as 100% of the least mean. Within each subset, individual sizes differed from the set mean by -21%, -15%, -9%, -3%, 3%, 9%, 15%, and 21%. It can be seen in Fig. 1A that size distributions for each of the sets were relatively narrow, but together they formed two clusters of sizes separated by a substantial gap. We predicted that such size distributions would make segmentation and categorization of two subsets within a superset relatively easy.
- 2) *Smooth narrow distribution*: We generated a smooth narrow distribution to provide the same relative range within each category as in the two-peaked distribution condition (from -21% to 21% of the set mean) but with a smaller separation between the sets, not much exceeding the separation within each set. To generate size distributions of the two sets, we used the large set:small set mean ratio of 1.7:1. In this case, the transition between sets was just slightly bigger than the transition within each set (Fig. 1B). We predicted that this condition would make the segmentation of two subsets more difficult based on the size distributions than the two-peaked distribution condition since there was no discontinuity in the overall distribution that could support peak separation.
- 3) *Smooth fat distribution*: When we generated a smooth distribution with a greater bandwidth from which individual sizes that belonged to both of the subsets were drawn, we ensured that the overall range of the smallest and biggest items of the superset was about the same as in the two-peaked distribution. Here, the large set/small set mean ratio was 2.4:1 and individual sizes within each set differed from the set mean by -35%, -25%, -15%, -5%, 5%, 15%, 25%, and 35% (Fig. 1C). We predicted that segmenting and categorizing two subsets in this condition would be difficult since two distributions of the subsets were not easily segmentable.

Figure 1 describes how individual sizes in the small and large sets were generated and distributed based on our size-generation algorithm. Because the actual mean sizes of the small and large sets were varied in every trial, we plotted the distributions of sizes as proportions to the mean sizes of small and large subsets. As shown in Fig. 1, individual sizes were spaced equally (by 6% from their adjacent neighbors) within their categories (as they were scaled relative to categorical mean sizes). On the other hand, the individual sizes ended up being spaced unequally in terms of the entire ensemble so that the absolute step size between items within the large set was always bigger than that within the small set according to the large set/small set mean ratio. Such asymmetry resulting from our size generation algorithm, in fact, complied with Weber’s law that perceived difference between two sizes is approximately proportional to their sizes in the domains of both individual size and mean size (e.g., Allik et al., 2013). This way of size generation for individual circles in the small and large sets was implemented in previous work and has been shown to ensure that perceived variability of individual members (e.g., variance or range) is roughly the same across the categories (e.g., Khvostov & Utochkin, 2019).

Moreover, this algorithm also made the whole feature distribution inherently skewed, resulting in asymmetric probability density. As a result, the grand mean, suggested to be one of the robust ensemble representatives (Alvarez, 2011; Ariely, 2001; Chong & Treisman, 2003; Khayat & Hochstein, 2018, 2019), was shifted to the right compared to the median (which is defined as a categorical boundary in our task). In other words, the smallest items from the large category were always closer to the grand mean than the largest items of the small category. As can be seen in Fig. 1B and C, the smallest items of the large category were even smaller than the grand mean in the smooth narrow and smooth fat conditions. As will be seen later, this mismatch between the task-defined categorical boundary (median) and the grand mean would provide a sensitive case to query the type of the internal rule (e.g., whether observers rely more on the median or the grand mean) for establishing that perceptual boundary.



**Fig. 1** Histograms of size distributions used in Experiment 1: (A) two-peaked distribution; (B) smooth narrow distribution; (C) smooth fat distribution. The shaded regions show the space between the large and small subsets where the category changes. The solid vertical line depicts the

superset grand mean and the dashed lines depict subset mean sizes. Note that, although in each subset the sizes cover the same relative range (percentage of the subset mean size), the absolute ranges are different, as they are scaled to fit Weber's law

**Procedure** Experimental sessions were run in a darkened room. Participants were seated approximately 50 cm from a monitor. On each trial, they were instructed to categorize a probed item as small or large, depending on its relative size within the whole

set, including all the items presented briefly in the visual array. The categorization rule was explicitly stated as median-based: Participants were told to answer whether the probed item belonged to the smallest or the largest half of the set.

A sample trial of Experiment 1 is shown in Fig. 2. After a ready signal, a stimulus image that contained all the 16 circles was presented for 100 ms. After the stimulus image, only one of the circles from the stimulus image remained, and the rest of the circles disappeared to instruct participants to indicate which of the subsets – either large set or small set – the remaining circle (a test circle) seemed to belong to. The participant responded whether this object belongs to the “small” or to the “large” category by pressing the “left” or the “right” button, respectively. After their response was made, the feedback was provided. Participants completed a total of 576 trials (16 relative sizes  $\times$  3 size distributions  $\times$  12 repetitions per condition). Twelve trials were added at the beginning of the experimental session for practice but were excluded for data analyses.

## Results and discussion

Trials with excessively fast responses ( $< 200$  ms) were excluded from the analysis. Data from one participant who made such excessively fast responses  $\sim 56\%$  of the time were also excluded from the analysis. Therefore, the data from 19 participants were analyzed. Overall, less than 0.3% of trials were excluded from the data analysis of these participants due to excessively fast responses.

Participants’ response accuracy for categorizing an item into one of the two subsets showed the clear “v-shaped” curves for all the three different types of the size distribution, two-peaked, smooth narrow, and smooth fat distributions. Figure 3A plots the participants’ percentage correct responses as a function of the distance between the size of the single

circle to be categorized and the categorical boundary (median) of the entire set of the circles shown in the visual image. Each of the 16 items shown in the visual image had its unique size, and the absolute distance from the categorical border strongly depended on the type of the size distribution. To resolve and control for such variations, we merged each of two neighbor sizes starting with the smallest and ranked them: We assigned ranks -4 to -1 to the items of the “small” category (with -4 being the smallest item) and ranks 1 to 4 to the “large” category (with 4 being the largest item). Therefore, the ranks define the relative position of a probed circle, both within a category (e.g., either a “small” or “large” category) and away from the categorical boundary.

Not surprisingly, participants’ accuracy for the item categorization was the worst when the test circle to be categorized was close to the categorical boundary (e.g.,  $\pm 1$ ), but systematically improved as the size of the circle became more deviated from the categorical boundary. This observation was confirmed by the significant main effect of the size distance between the test circle and the superset ( $F(7,126) = 125.96$ ,  $p < 0.001$ ,  $\eta^2_G = 0.68$ ) from the statistical test using the two-way repeated-measures ANOVA with the two factors of the size distance rank (eight levels: -4, -3, -2, -1, +1, +2, +3, and +4) and the distribution types (three conditions: two-peaked, smooth narrow, and smooth fat distributions).

From the same ANOVA test, we also found a significant main effect of the distribution types ( $F(2,36) = 157.14$ ,  $p < 0.001$ ,  $\eta^2_G = 0.40$ ). Specifically, the overall accuracy for categorization was better for the two-peaked distribution than the other two conditions (smooth narrow and smooth fat), which was further confirmed by the post hoc contrast analyses (two-

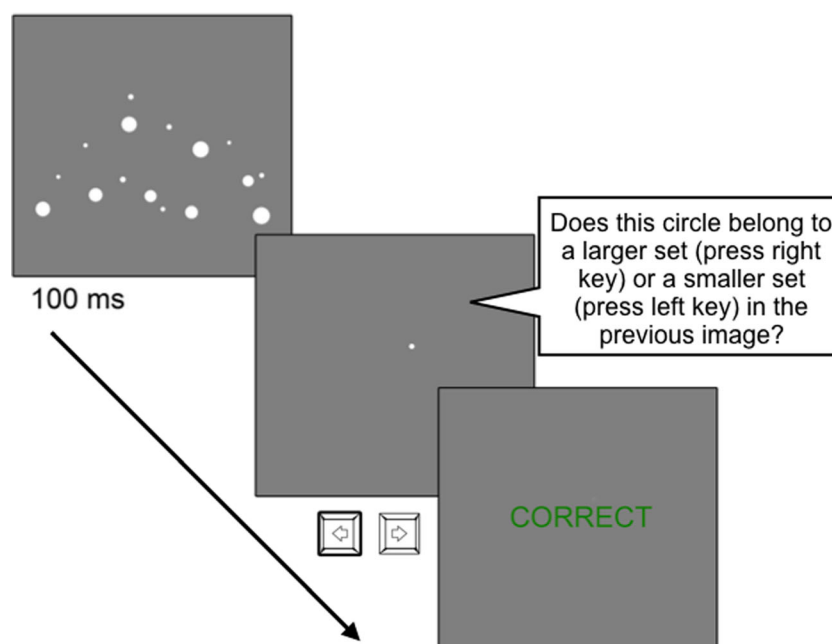
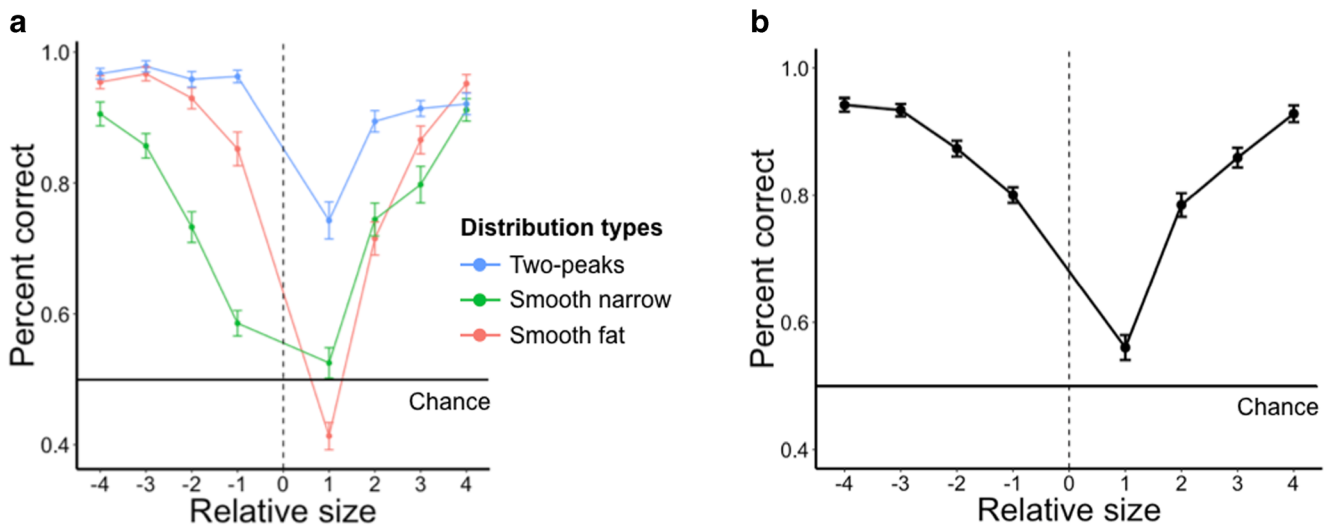


Fig. 2 An example of a trial in Experiment 1



**Fig. 3** Percent correct of the categorization task as a function of the relative item size in Experiment 1: (A) for different types of the size distribution and (B) overall performance (averaged across the

distribution types). The vertical dashed line shows the categorical boundary (relative position 0) that has not been actually presented. Error bars denote the standard error of the mean

peaked vs. smooth narrow:  $t(18) = 14.39$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen's  $d = 3.3$ ; two-peaked vs. smooth fat:  $t(18) = 10.55$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen's  $d = 2.4$ ). For the two-peaked distribution, the categorization accuracy mostly reached near the plateau except for the point at size rank +1. Conversely, the other two conditions showed categorization accuracy that strongly depended on the size difference between the test circle and the mean of the superset. In turn, in the smooth fat distribution participants were overall more accurate than in the smooth narrow one ( $t(18) = 9.99$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen's  $d = 2.3$ ), presumably because the former one included some exemplars more distinct from the categorical boundary (at least two extreme values at both sides of the “tails” of the size distribution). This pattern of results suggests that the ease of categorical parsing was systematically varied across the distribution types of the superset. The participants appeared to be sensitive to size distributions of circles and capable of categorizing individual members into one of the subsets, based on the size distributions shown in visual stimuli.

Moreover, both the range and smoothness of distributions seemed to affect the accuracy of categorization systematically, as supported by the significant interaction ( $F(14,252) = 33.06$ ,  $p < 0.001$ ,  $\eta^2_G = 0.39$ ) between the two factors – size distances and the distribution types. This significant interaction suggests increasing profoundness (e.g., depth) of the “dipper parts” of the V-shapes across the distribution types (Fig. 3A). That is, participants showed quite similar levels of accuracy for the extreme items at the “tails” (e.g., ranks -4 and +4), whereas decrement in the accuracy for the center of the distribution (especially ranks -1 and +1) was much greater in the two smooth distributions than in the two-peaked distribution.

This pattern is presumably because middle items of the smooth distributions were closer to the boundary and between-categorical gap was smaller than in the two-peak distribution.

Finally, we found an interesting asymmetry in our V-shaped functions. For all the three types of size distributions, the worse categorization response was observed at +1, suggesting that the participants made more errors when the probed test circle was the smallest circle in the large category (but slightly larger than the categorical boundary), by erroneously categorizing them into a small subset. As a result, the V-shaped functions were all shifted to the right relative to the task-defined categorical boundary. When collapsed across the distribution types (Fig. 3B), this shift was most clear, with rank -1 = rank +2, rank -2 = rank +3, and rank -3 = rank +4 ( $ps > 0.33$ , Cohen's  $ds < 0.23$ ) in the observed categorization accuracies, whereas symmetrical ranks yielded strongly asymmetrical results, with a systematic prevalence of the small category ( $ps < 0.001$ , Bonferroni-corrected  $\alpha = 0.002$ ,  $ds > 1.1$ ; except rank -4 = rank +4,  $p = 0.3$ ,  $d = 0.25$ ). These results suggest that participants made more error responses when they had to categorize items from the large category compared to the small category, especially when these items were relatively close to the categorical boundary. To recap, the way we generated individual sizes made the whole distribution skewed, such that the grand mean was shifted towards smallest items of the large category more, compared to towards largest items of the small category. Importantly, in the smooth narrow and smooth fat distributions, rank +1 items were greater than the median but smaller than the grand mean. For these particular points, we observed that the accuracy dropped to chance (smooth narrow: accuracy = 0.52, smooth fat: accuracy = 0.41, Fig. 3A). At the same time, in the two-peak distribution, where even the smallest

item of the “large” category was fairly greater the grand mean, the drop in the accuracy at the rank of +1 was not that dramatic (accuracy = 0.74, Fig. 3A). Therefore, the whole pattern of asymmetry, with the “rank +1 effect” remarkably correlated with an actual item position relative to the grand mean, suggests that the internal categorical boundary was shifted in the direction of the grand mean of the superset. The finding that this shift occurs despite the instruction to use the median criterion of categorization can also suggest that people tend to rely on a representation of the mean as a categorical boundary automatically.

## Experiment 2

### Categorization in the domain of size and ensemble extraction in the domain of location

In Experiment 1, we have shown that participants could categorize an individual object very rapidly into one of the two subsets based on the average size of the superset. The ease of such segmentation and categorization was systematically varied both by the size difference between the individual object to be categorized and the mean of the superset and by the shape of the feature distribution of all the items shown in a visual array. In Experiment 2, we extended the findings from Experiment 1 and further tested how participants use the distributional properties of an ensemble in one feature dimension (e.g., size) to parse items into categories and make task-relevant, category-specific judgments in another feature dimension. Returning to the farmer’s market example, you might also want to make sure you choose apples that are ripe enough as well as big enough. You will first compare the overall size of each pile of apples to check which pile is relatively larger, but then also compare other qualities such as the overall color or hue of the two piles (smaller and larger sets) for the final decision. Experiment 2 sought to characterize such a process: ensemble-based segmentation in one feature dimension (e.g., size) for extraction of ensembles in the other feature dimension (e.g., centroid).

In many previous studies, ensemble extraction in a particular feature domain (e.g., average size, numerosity, and so on) has been tested independently from other features. For example, estimation of the average size or numerosity of multiple subsets was tested by using different color cues that are discrete and separable enough for each of the subsets (e.g., Chong & Treisman, 2005; Halberda et al., 2006; Im & Chong, 2014, Utochkin & Vostrikov, 2017, etc.). In other studies, location (e.g., spatial separation between subsets of circles) was utilized for segmentation of subsets (e.g., left vs. right sets; Chong & Treisman, 2003; Corbett, Wurnitsch, Schwartz, & Whitney, 2012; Epstein & Emmanouil, 2017). Many of these studies assume that multiple subsets can be

(almost) perfectly segmented based on the color or spatial cues prior to ensemble extraction. Although the results collectively suggest that human observers are capable enough of segmenting subsets by color cues or spatial separation, this process is not necessarily “cost-free,” given that combining color and location cues for segmentation can significantly improve participants’ performance on ensemble extraction from multiple subsets, compared to when only one feature is provided as a segmentation cue (e.g., Im, Park, & Chong, 2015).

It has been shown that within-subset ensemble judgments can be penetrable for the influence of another, irrelevant subset (Inverso, Sun, Chubb, Wright, & Sperling, 2016; Oriet & Brand, 2013; Utochkin et al., 2018). Here, we hypothesize that the degree of such penetrability should depend on the ease and robustness of segmentation of subsets based on the shape of the distribution in the feature domain of categorization. That is, if the estimated summary differs between two subsets and these subsets can be segmented into two different categories quite easily (e.g., as in the two-peaked distribution in Experiment 1), then the extracted ensemble representation by participants would be closer to the genuine summary of the subsets. In contrast, if subsets are hardly distinguishable and less separable as two categories, then participants’ ensemble estimation of subsets should be reported with a greater error, biased towards the common summary of all subsets (e.g., grand mean), possibly because some elements of the irrelevant subsets are confused with relevant elements. Furthermore, we predict that the ease of subset segmentation determined by the shape of the distribution in one feature domain (e.g., size) would systematically modulate the precision of ensemble extraction in another visual feature domain (e.g., location).

**Participants** Twenty-one students (16 females; age range: 19–24 years) of the Higher School of Economics took part in Experiment 2. All reported having normal or corrected-to-normal vision, normal color vision, and no neurological problems. All were naïve as to the purpose of the experiment. Written informed consent was obtained for the experiment from the participants in accordance with the Declaration of Helsinki.

**Apparatus and stimuli** In Experiment 2, participants responded by using a mouse cursor connected to the computer to indicate the position in the display with their dominant hand. As in Experiment 1, each stimulus included white circles with different sizes, presented on a gray background. As in Experiment 1, a superset contained 16 circles that were divided into two categories (subsets; large and small) based on their sizes, such that the “large” subset contained eight larger circles in the superset, whereas the “small” subset contained eight smaller circles in the superset. Grand mean sizes and size range for the “small” and “large” subsets were



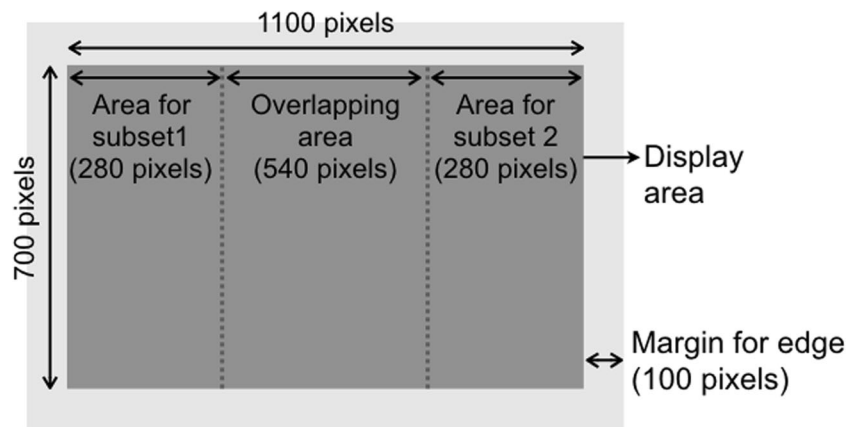
generated the same way as in Experiment 1. The detailed procedures and parameters for creating the three different size distributions (two-peaked, smooth narrow, and smooth fat) was identical to those described in Experiment 1.

The locations of individual circles were pre-generated for all the stimuli using the following algorithm (also see Fig. 4 for visual illustration). First of all, all the (x, y) coordinates for the center locations of circles were randomly chosen within the central display area (1,100 × 700 in pixels, and 28.05° × 17.85° in visual angle), which was surrounded by the margin of 2.55°. Next, the 16 randomly chosen locations were then assigned to one of the three imaginary parts (left, middle, and right) based on their x-coordinates. Specifically, if the x-coordinates were within the range of [1–280], the (x, y) coordinates were assigned to circles of subset 1 (which could be either large or small subset), so that the centroid of subset 1 was slightly shifted towards the left. If the x coordinates were within the range of [821–1,100], on the other hand, the locations were assigned to circles of subset 2, so that its centroid was slightly shifted towards the right. Because we wanted to ensure that circles of both subsets were spatially intermixed in the center of the display, the half of the (x, y) coordinates within the center area (with the range of [301–900] for the x coordinates) were randomly chosen and allocated to the subset 1 and the rest was allocated to the subset 2. This spatial arrangement allowed us to ensure that some of the circles of the two subsets were spatially intermixed so that simply clicking somewhere in the left visual field for one subset and the right visual field for the other subset would not systematically improve participants’ centroid extraction. This spatial arrangement also ensured that the stimulus always appeared to occupy locations randomly chosen from the same designated area across conditions and over the trials.

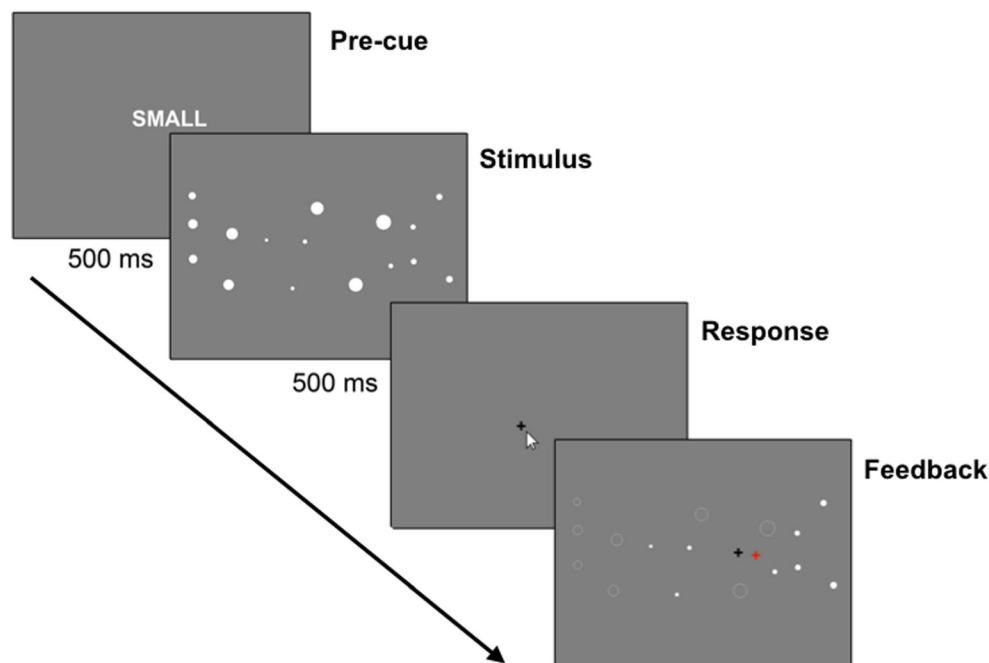
**Procedure** Experimental sessions were run in a darkened room. Participants were seated approximately 50 cm from a monitor. Prior to the experiment, participants received six

demo trials that were intended to provide them with a sense of “centroid.” In these trials, we showed them groups of circles for an unlimited time, giving an opportunity to estimate the centroid of the circles in a completely visible stimulus. To report the estimated centroid, the participants moved a mouse cursor (small black cross) around the screen and clicked a left mouse button to make their response. Immediately after the click, the cursor stayed on the screen, and a red cross appeared to mark the position of the correct answer, so the distance between the locations of two crosses served as feedback about response precision. Each experimental trial (shown in Fig. 5) began with a pre-cue that informed participants of which set (small subset, large subset, or all items) they had to attend to. The stimulus image was presented for 500 ms. Immediately after the stimulus, an empty screen was presented with a mouse cursor to adjust the centroid of the pre-cued set. The initial location of the mouse cursor was randomly determined so that it would not systematically bias participants’ responses. After a response was made, a feedback display was presented with the original image returned and a red cross indicating the correct centroid location. The pre-cued set of circles were shown solid on the feedback screen, whereas the irrelevant circles (if any) were shown as outlines.

There were three conditions in Experiment 2: attended-subset, half-set-only, and superset conditions. In the attended-subset condition, participants were instructed by a pre-cue to attend to either a large set or a small set (defined by a median, as in Experiment 1) in the following stimulus image, then report an average location (centroid) of the attended set only. In the half-set-only condition, only one subset (either small or large subset) was presented such that participants were not required to segment or parse any subsets from the display. Finally, the superset condition did not require participants to segment any subsets, either, even though both small and large subsets were presented in the stimulus image. Thus, the participants’ performance on the attended-subset condition would reflect their perceptual ability to categorize individuals into



**Fig. 4** Visual illustration of the areas (left, middle, and right) for assigning locations of circles. The (x,y) coordinates that were first randomly generated were assigned to one of these areas based on the x coordinates



**Fig. 5** An example trial of Experiment 2 demonstrating the “attend-subset” task

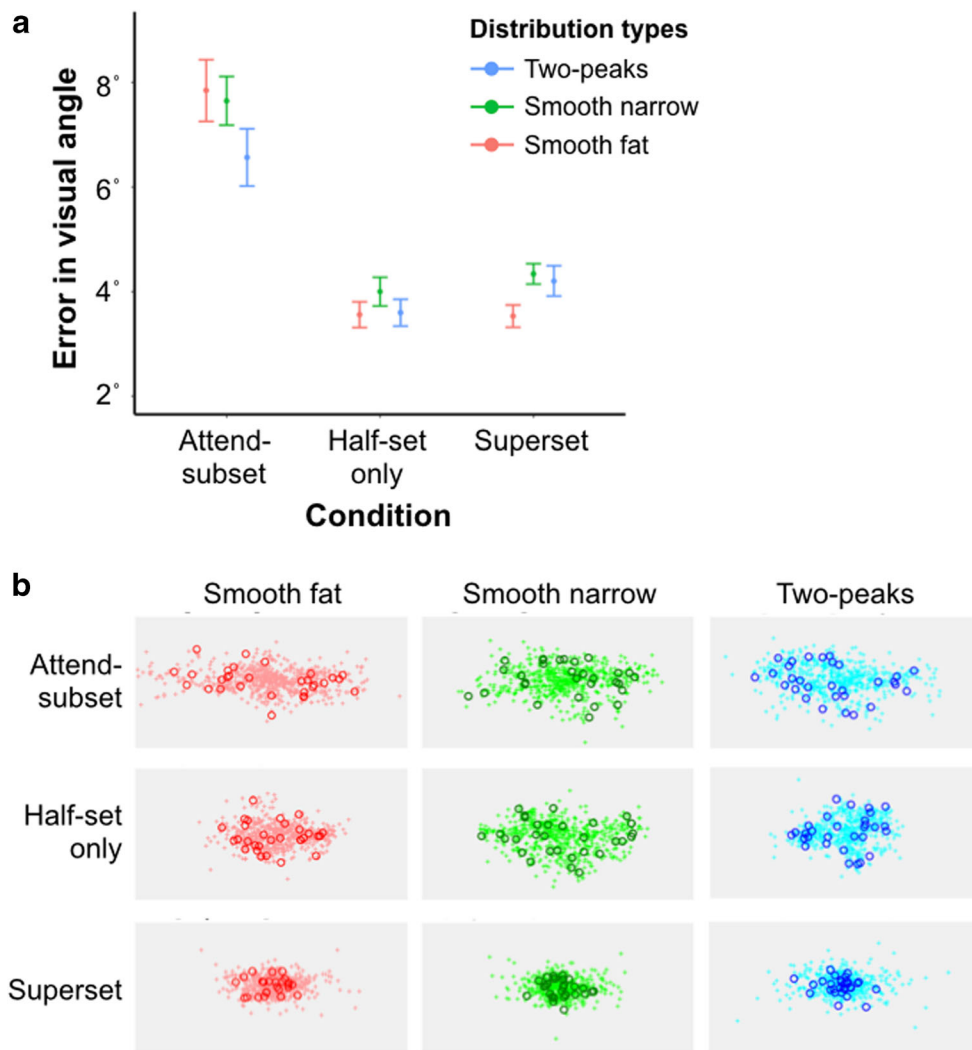
two subsets and extract ensemble representation from the subsets. Their performance on the half-set-only condition would reflect their ability to extract centroid from a subset when there is no need to segment subsets at all. Finally, the performance on the superset condition would tell us whether the shape of the size distribution of all the individual objects still influences participants’ perceptual ability to extract centroid from the superset, even when they did not need to segment subsets. In both the half-set and the superset conditions, participants were pre-cued that they would have to attend all the circles, then report the centroid of them. The trials of all the conditions (attended-subset, half-set-only, and superset conditions) were randomly interleaved, rather than being blocked, to ensure that participants would not consistently employ any different strategies across the conditions. In each condition, we used the three different types of size distributions for individual circles as in Experiment 1: two-peaked, smooth narrow, and smooth fat distributions. Thus, we had a  $3$  (three conditions: attended-subset, half-set-only, and superset)  $\times$   $3$  (three types of size distributions: two-peaked, smooth narrow, and smooth fat distributions) experimental design.

## Results and discussion

For each trial, the correct answer for centroid was calculated as an average Euclidean distance between the locations of all the set members. As a measurement for error, we calculated the distance between the actual centroid of the set to be extracted and the location at which the participants pointed by using the mouse cursor. Figure 6A summarizes the mean error for each condition (attended-subset, half-set-only, or superset)

and each type of size distribution (two-peaked, smooth narrow, or smooth fat distributions). We first observed that the attend-subset condition showed greater response errors than the other two conditions for all three distribution types. A statistical test using the two-way repeated-measures ANOVA with the two factors of the task conditions (three levels: attended-subset, half-set-only, and superset) and the distribution types (three levels: two-peaked, smooth narrow, and smooth fat distributions) confirmed this observation with the strong main effect of the task conditions ( $F(2,40) = 94.45$ ,  $p < 0.001$ ,  $\eta^2_G = 0.50$ ). Further contrast analyses showed that when the participants did not have to segment subsets of circles as in the half-set-only and superset conditions, their accuracy was significantly better for all the three distribution conditions (attend-subset vs. half-set:  $t(20) = 11.56$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen’s  $d = 2.52$ ; attend-subset vs. superset:  $t(20) = 9.12$ ,  $p < 0.001$ , Bonferroni corrected- $\alpha = 0.017$ , Cohen’s  $d = 1.99$ ). We also found a significant but weak main effect of the distribution types ( $F(2,40) = 15.02$ ,  $p < 0.001$ ,  $\eta^2_G = 0.02$ ), and a significant but weak interaction between the two factors ( $F(4,80) = 5.19$ ,  $p < 0.001$ ,  $\eta^2_G = 0.04$ ).

The finding that the half-set-only and the superset tasks were performed better than the attend-subset tasks indicates that there was a cost of subset selection for later extraction of ensembles. This finding is in line with previous evidence that segmenting subsets of items based on size or orientation is a strong limiting factor for ensemble tasks (Inverso et al., 2016; Oriet & Brand, 2013; Utochkin et al., 2018; however, the cost may be minimal in the domain of color – see Sun, Chubb, Wright, & Sperling, 2016a). Note that in previous studies,



**Fig. 6** (A) Centroid positioning error as a function of the task and the size distribution in Experiment 2. Error bars denote the standard error of the mean. (B) The spatial distribution of correct answers (large dots) and participants' responses (small dots) on a screen in the various conditions of Experiment 2

category-defining features could be both extremely distinct between categories and homogeneous within the category to make segmentation easy (e.g., strictly vertical and strictly horizontal lines – Inverso et al., 2016; Oriet & Brand, 2013; or extremely short and extremely long lines – Utochkin et al., 2018). Even in these cases, however, observers were imperfect at reporting subset summaries. Compared to the previous studies, category-defining features in the current study (e.g., size) were distributed in a more heterogeneous and continuous manner; thus, observers were more frequently confused overall when segmentation of subsets was required.

Of principal interest to us, we observed that when participants had to segment one of the categories to extract and report the centroid of the category in the attend-subset condition, their accuracy for the two-peaked distributions was better than the other two distributions (two-peaked vs. smooth narrow:  $t(20) = 3.96$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ ,

Cohen's  $d = 0.86$ ; two-peaked vs. smooth fat:  $t(20) = 4.25$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen's  $d = 0.93$ ). There was no difference, however, between the two smooth distributions – narrow versus fat ( $t(20) = 0.68$ ,  $p = 0.51$ , Cohen's  $d = 0.15$ ). This result suggests that enhanced peak separation of the feature distribution (i.e., good segmentability) facilitates the extraction of ensemble features from the two independent categories. The two-peaked distributions with larger separation yielded the best accuracy for the attend-subset task, presumably providing the best categorical separation.

In the superset condition in which the participants were to extract the centroid of all the circles, without any segmentation of subsets of circles, we found that their precision for displays that contained the smooth fat distributions of circles was better than in the two other distributions (smooth fat vs. two-peaked:  $t(20) = 3.62$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ ,

Cohen's  $d = 0.79$ ; smooth fat vs. smooth narrow:  $t(20) = 6.78$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen's  $d = 1.48$ ). On the other hand, for the half-set-only condition, in which only a subset of circles sampled from one single subset was presented, we found that the smooth narrow distribution yielded an impaired precision compared to the two other distributions (smooth narrow vs. two-peaked:  $t(20) = 3.89$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen's  $d = 0.85$ ; smooth narrow vs. smooth fat:  $t(20) = 3.14$ ,  $p = 0.005$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen's  $d = 0.69$ ). This result suggests that the participants were still sensitive to the distribution of the subsets of circles even when segmentation of subsets was not required or helpful for them to extract the centroids of the entire sets. Moreover, the fact that the two-peaked distribution was not any better than the other two distribution types in the superset condition suggests that the two-peaked distribution improved participants' performance on ensemble extraction only when subset categorization was necessary. Further research beyond the main focus of the current article is necessary to understand the nature of these effects.

Although the centroid task has been utilized as a useful tool to measure ensemble perception in many previous studies (Alvarez & Oliva, 2008; Inverso et al., 2016; Sun et al., 2016a, b; Rodriguez-Cintron, Wright, Chubb, & Sperling, 2019), the use of this task in our experiment revealed some practical challenges that are worth noting. The first reasonable point is that participants' accuracy for extracting centroids in the three different (attended-subset, half-set-only, and superset) conditions could be affected differently by the overall spatial arrangements of objects in these conditions, not solely by how the objects were categorized in the size domain. This is particularly the case because the physically limited display was already quite packed and occupied by multiple circles, which inevitably resulted in the superset condition including all the circles in the visual array (thus twice as many circles compared to the pre-cued subset condition) to have centroids more toward the center of the screen, compared to the attended subset and half-set-only conditions (see Fig. 6B). However, our current findings cannot be merely explained by the concentrated pattern of centroids in the superset condition: If this was the case, participants would have been more precise in the superset condition than in both the attend-subset and half-set-only conditions by simply choosing just around the center. Instead, we did observe that participants' accuracy for centroid extraction in the superset conditions was somewhat comparable to that in the half-set-only subset condition despite their different profiles of centroid concentration (see Fig. 6A).

One might argue that the greater error in the attend-subset condition reflected the possibility that participants had just reported the centroid of everything (e.g., superset) instead of trying to categorize items to determine the centroid of the

attended subset only. To evaluate this possibility, we tested whether participants' centroid extraction in the attend-subset trials could lie in the same area as their estimated centroids in the superset trials. We found that the estimated centroid locations in the attend-subset trials were almost twice as far from the superset centroid as in the genuine superset trials ( $7.6^\circ$  vs.  $4^\circ$ , respectively; comparison:  $t(20) = 7.31$ ,  $p < 0.001$ , Cohen's  $d = 1.6$ ). This result suggests that, in the subset trials, observers did not merely report the centroid of all the items (e.g., superset), but marked locations toward the extracted centroid of the pre-cued subset by categorizing individual items and segmenting subsets. The following analysis further shows that the greater errors in the attend-subset conditions are more likely to reflect the additional processing noise and cost resulting from incomplete segmentation of the pre-cued subset from the superset, which is shown as a mixture of the centroid of the pre-cued subset and global centroid of the superset.

**Bias towards the center** Note that the unidimensional Euclidean distance as the measurement of the centroid localization errors does not necessarily capture the whole spectrum of changes in the two-dimensional space of the screen. Specifically, it does not introduce a single measurement of directionality that would still be useful to unambiguously estimate the trend of the bias: either towards or away from a certain location. In fact, the same quantitative change in the Euclidean distance can be caused either by attraction in one dimension or by repulsion in another. Therefore, here we used a "triangular" algorithm to evaluate the directionality of the bias. We calculated (1) the mean of the distances between participants' responses and the correct answers (the actual centroids of the attended subsets), here termed Response-Subset distance (mean =  $7.3^\circ$ ,  $SD = 5.3^\circ$ ), (2) the mean of the distances between participants' responses and the centroids of the supersets, here termed Response-Superset distance (mean =  $7.6^\circ$ ,  $SD = 5.0^\circ$ ), and (3) the mean of the distances between the correct answers (actual centroid locations of the attended subsets) and the centroid of the superset, here termed Subset-Superset distance (mean =  $10.2^\circ$ ,  $SD = 3.2^\circ$ ). We found that both Response-Subset distance and Response-Superset distance were shorter than Subset-Superset distance (Response-Subset vs. Subset-Superset:  $t(19) = 5.60$ ,  $p < 0.001$ , Cohen's  $d = 1.2$ ; Response-Superset vs. Subset-Superset:  $t(19) = 6.44$ ,  $p < 0.001$ , Cohen's  $d = 1.4$ ). This suggests that average response lies roughly in an area between the subset centroid and the superset centroid. In other words, participants marked the subset centroid in its real neighborhood but with a substantial shift towards the superset centroid. If we imagine a triangle with aspects corresponding to Subset-Superset, Response-Subset, and Response-Superset distances, then the aspect ratios found corresponds to a disposition of when the observer's response apex is projected onto the

Subset-Superset segment: This is a rough geometrical approximation of what the bias towards the superset centroid is. As a counter-example, if our observers had guessed random locations around the superset centroid, then Response-Subset distance would have been greater than or equal to Subset-Superset distance. As another counter-example, if our observers had been unbiased, then Response-Superset distance would have been equal to Subset-Superset distance. Finally, if our observers had been biased away from the superset (repulsion), then Response-Superset distance would have been greater than Subset-Superset distance.

One more reservation that can be made regarding Experiment 2 is that objects from the large and small categories were not mixed in space entirely randomly, unlike Experiment 1. To recap, two subsets had centroids that were manipulated to be significantly shifted such that some category members were intermixed within subset overlap, whereas others, outside the overlap, were not. This inherently resulted in a size gradient that could be used to judge the category center without category parsing (Rodríguez-Cintron et al., 2019). We acknowledge that this can be a plausible, alternative strategy. However, this does not undermine our main findings. We found that centroids of subsets drawn from two-peaked distributions were localized more precisely than those drawn from smooth distributions, regardless of the ranges of the feature distributions (narrow or fat). This pattern cannot be fully explained by the alternative strategy of gradient-based centroid localization because the smooth fat distributions provide a much stronger gradient (and therefore more efficient gradient-based centroid localization) than smooth narrow distributions. Rather, it seems to reflect the fact that the efficiency of centroid localization was dependent on how abrupt the difference between different subsets was, as defined by the shape of the feature distribution. Such an ability to see an abrupt transition between subsets in an uncertain spatial structure (like the overlapped region of both subsets in our experiment) is exactly what is meant by the concept of feature-based segmentability (Utochkin et al., 2018).

## Experiment 3

### Categorization in the domain of length and ensemble extraction in the domain of orientation

Given the practical limitations of the centroid paradigm (such as limited screen space or an inherent size gradient when the small and large subsets are shifted to provide a substantial centroid difference), we conducted Experiment 3 to further test our hypothesis that ensemble distributional properties determine segmentability of subsets for categorization and category-specific ensemble extraction in different feature domains. Here we used orientation instead of the centroid to

minimize the challenges caused by the limited space in the display and to replicate and generalize the previous findings of Experiment 2. Orientation and size are traditionally assumed to be separable feature dimensions, at least with a very weak interaction (e.g., Ashby & Lee, 1991; Garner & Felfoldy, 1970; Shepard, 1964; Ward, 1985; but also see Potts, Melara, & Marks, 1998). Unlike the centroid task, there is no need to create an arbitrary spatial shift between subsets to make subsets separable in the domain of orientation, unlike the domain of spatial location. Therefore, different subsets can be spatially intermixed completely, and no gradient cues are available. If we can replicate the effects of the types of size distributions on the precision and the bias in extracting average orientation, we can conclude that the segmentability of a subset based on a different feature dimension modulates the ease of extraction of ensemble representation from the subset.

## Participants

A new group of 20 undergraduate students (17 females; age range: 18–28 years) of the Higher School of Economics took part in the experiment for extra course credits. All reported having normal or corrected-to-normal vision, normal color vision, and no neurological problems. Written informed consent was obtained for the experiment from the participants in accordance with the Declaration of Helsinki.

## Apparatus and stimuli

The apparatus used was the same as in Experiment 2. Sets of 16 white lines were presented as stimuli within a  $16^\circ \times 16^\circ$  square visual array. This visual array was divided into 16 cells ( $4 \times 4$ ) with an invisible grid (each cell subtending  $4^\circ \times 4^\circ$ ). Each cell contained a single item of a set. In each cell, an item could be randomly jittered within  $\pm 0.8^\circ$  in both the horizontal and vertical directions. All lines had a fixed width of  $0.16^\circ$ . Lengths of the lines varied from  $0.6^\circ$  to  $3.7^\circ$  and were drawn from one of the length distributions described below. Overall, these distributions followed the three types from Experiments 1 and 2 (two-peaked, smooth narrow, and smooth fat). However, since our stimuli in Experiment 3 were the lines characterized by lengths rather than circles characterized by diameters, we changed specific values. These values were adjusted from the study by Utochkin et al. (2018), where length distributions were manipulated to provide high or low levels of segmentability of subsets.

- 1) *Two-peaked distribution (i.e., bimodal distribution with larger separation)*. The mean length of lines in the “long” subset was  $3.35^\circ$ , and the mean length of the “short” subset was  $0.7^\circ$ . For each subset of eight lines, individual lengths were approximately  $\pm 3\%$  and  $\pm 9\%$  of the subset (categorical) mean, each size assigned to two lines.

- 2) *Smooth narrow distribution.* The mean length of lines in the “long” subset was  $2.4^\circ$ , and the mean length of the “short” subset was  $1.7^\circ$ . For each subset of eight lines, individual lengths were approximately  $\pm 3\%$  and  $\pm 9\%$  of the subset (categorical) mean. Therefore, each subset had the same relative range as the categories of the two-peaked distributions, but the overall superset range was much smaller.
- 3) *Smooth fat distribution.* The mean length of lines in the “long” subset was  $2.8^\circ$ , and the mean length of the “short” subset was  $1.24^\circ$ . For each subset of eight lines, individual lengths were approximately  $\pm 10\%$  and  $\pm 35\%$  of the subset (categorical) mean. Therefore, each category had a greater relative range compared with the previous two distributions, but the overall superset range was comparable with the two-peaked one.

The grand mean orientation was randomly chosen in each trial from the range between  $1^\circ$  and  $180^\circ$ . The difference in the mean orientations between the “long” and “short” subsets was fixed to be  $30^\circ$ . Whether the average orientation of one subset would be more tilted than the other to a clockwise direction or counter-clockwise direction was randomly determined on each trial. Thus, the grand mean orientation of the superset always was  $\pm 15^\circ$  away from the mean orientation of each subset. The range of orientations within each subset was  $60^\circ$ , such that individual orientations were separated by  $\pm 30^\circ$ ,  $\pm 22^\circ$ ,  $\pm 14^\circ$ , and  $\pm 5^\circ$  from the subset mean. Within each subset, lengths and orientations of individual members formed random conjunctions, such that there were no strict and predictable correlations between these two features in a subset. As in Experiment 1, items from the two subsets were positioned randomly and spatially intermixed.

## Procedure

The overall design and procedure of Experiment 3 were similar to those of Experiment 2, except that participants had to report the mean orientation, instead of the centroid, of a specific set, depending on the cue provided before each trial. Here, we also had three conditions asking participants to attend to a subset of lines based on a median length (pre-cued with either “Long” or “Short”), attend to the superset of all the 16 lines, or attend to a half-set presented alone (both superset and half-set were pre-cued with “All”). A sample trial of Experiment 3 is illustrated in Fig. 7. After the pre-cue presentation for 1 s, a stimulus containing lines was presented for 500 ms. The stimulus was followed by a response display where observers were instructed to adjust the orientation of a probe stimulus to report the mean orientation of the pre-cued set. A probe stimulus was a line in its length of  $1.6^\circ$ , with an orientation that was randomly determined presented in the center of the empty screen. The circle was surrounded by a

black ring with a white slider. By dragging the mouse around the ring, participants could turn the line (which was accompanied by the slider moving along the ring) to choose any orientation between  $1^\circ$  and  $180^\circ$ . To record the answer, participants pressed a space bar and immediately got feedback showing their adjusted orientation and the correct answer. Experiment 3 consisted of 540 trials: 3 distributions (Two-peaked, Smooth narrow, and Smooth fat)  $\times$  3 conditions (attended-subset, half-set-only, and superset)  $\times$  60 repetitions. At the beginning of the experiment, participants completed 15 practice trials.

## Results and discussion

**Precision** On each trial, we calculated the circular error, the angular difference between participant’s response, and the correct response (Error = Response – Correct), with the circular error of  $0^\circ$  corresponding to the perfectly correct answer and the circular error of  $\pm 90^\circ$  corresponding to maximum possible error. From the distribution of errors, we calculated the standard deviation (*SD*) to use it as an estimate for the precision of orientation averaging such that a greater *SD* indicates less precise averaging.

First of all, we observed that the half-set condition showed better performance (i.e., decreased standard deviation) than the other two conditions for all three distribution types. In support of this, a statistical test using the two-way repeated-measures ANOVA with the two factors of the task conditions (three levels: attended-subset, half-set-only, and superset) and the distribution types (three levels: two-peaked, smooth narrow, and smooth fat distributions) showed a significant main effect of the task conditions ( $F(2,38) = 83.17, p < 0.001, \eta^2_G = 0.28$ ). The main effect of the distribution type was non-significant ( $F(2,38) = 1.30, p = 0.28, \eta^2_G = 0.004$ ). We also found a significant, though small, effect of the interaction between the two factors ( $F(4,76) = 5.91, p < 0.001, \eta^2_G = 0.03$ ). The effect of the task conditions on the *SD* of the circular error distribution was further tested using contrast analyses, with the half-set-only condition (average *SD* =  $21^\circ$ ) being more precise than the attend-subset and superset conditions (average *SDs* =  $28^\circ$  in both; comparisons: half-set vs. subset:  $t(19) = 14.19, p < .001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen’s  $d = 3.17$ ; half-set vs. superset:  $t(19) = 11.71, p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen’s  $d = 2.62$ , Fig. 8A). It is not surprising that the mean orientation of the half-set was estimated more precisely than the mean orientation of the superset because the physical orientation range of half-sets was smaller than that of supersets ( $60^\circ$  vs.  $90^\circ$ ). This is also consistent with previous reports of averaging error increasing with the range of feature values (Corbett et al., 2012; Dakin, 2001; Fouriez, Rubenfeld, & Capstick, 2008; Im & Halberda, 2013; Marchant, Simons, & de Fockert, 2013; Maule & Franklin, 2015; Utochkin &

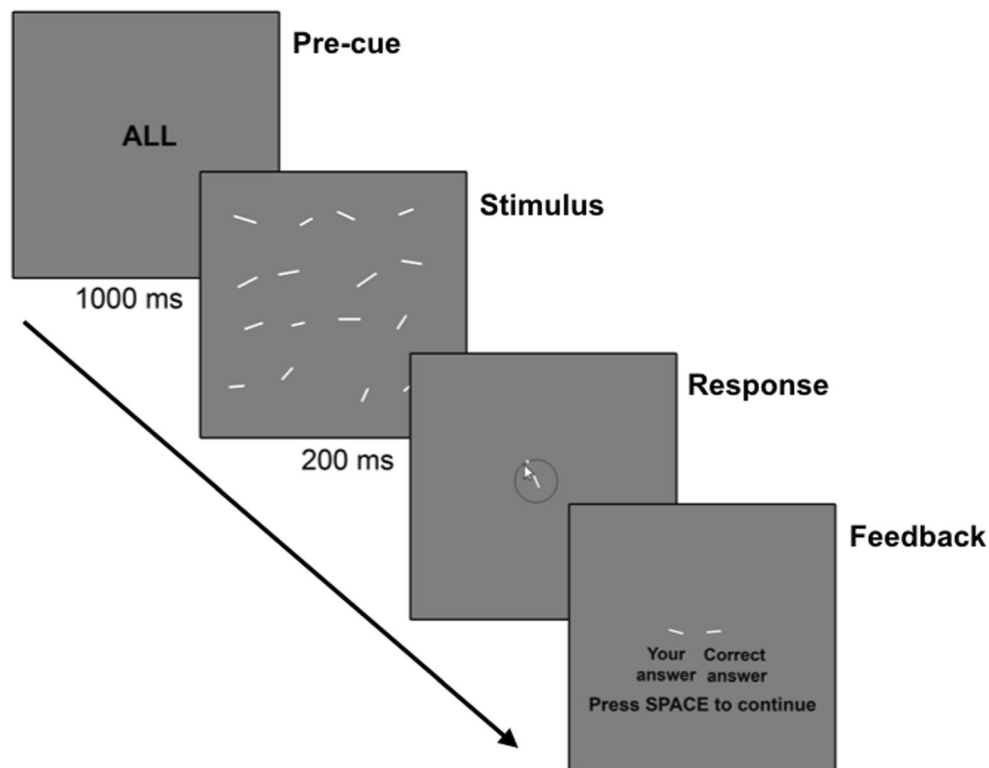


Fig. 7 An example trial of Experiment 3 demonstrating the “Superset” task

Tiurina, 2014). On the other hand, the precision in the attend-subset condition was worse than in the half-set-only condition, although the physical range of the relevant category was the same. This result replicates the findings reported in Experiment 2, suggesting that the ensemble extraction of a subset is noisy due to the imperfect selection and segmentation process.

We next examined the effects of shapes of feature distributions. We found that when observers had to report the mean orientation of the attend-subset, their *SD* was smaller in the

two-peaked distributions (average *SD* = 27°) than in the smooth narrow distributions (average *SD* = 30°; comparison:  $t(19) = 4.01, p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen’s  $d = 0.90$ ), although it did not reach a significant difference with the smooth fat distribution.

**Bias towards the mean of the superset** To examine whether participants’ mean orientation estimation of the attend-subset was biased towards the grand mean of the superset, we conducted an additional test by adjusting the error distributions.

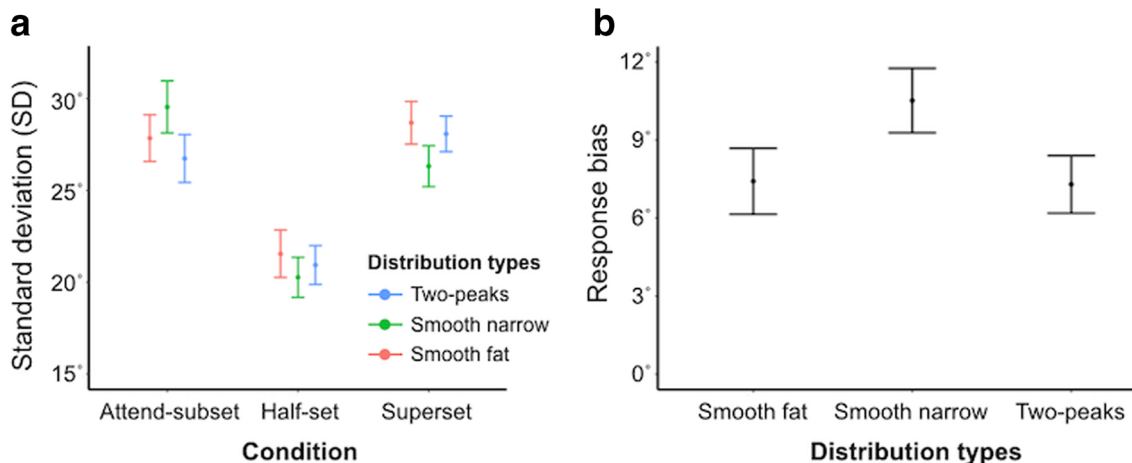


Fig. 8 The results of Experiment 3: (A) error standard deviation (*SD*) in the average orientation-adjustment task as a function of condition and length distribution; (B) bias towards the grand mean orientation as a

function of the length distribution in the “attend-subset” condition. Error bars denote the standard error of the mean

This adjustment was performed to evaluate the directionality of the mean orientation of the subset relative to the superset. We flipped the signs of raw error values (e.g., Error = Response – Correct answer) for the trials where the mean orientation of the subset was greater than the mean of the superset. Through this transformation, any responses that were biased *towards* the mean of the superset had a positive error value, whereas any responses that deviated *from* the mean of the superset had a negative error value. We then calculated the mean of the distribution of these transformed error values to quantify the magnitude of the systematic *bias towards* or *away from* the grand mean of the superset. The signed error values are particularly informative for our purpose because they could indicate both the directionality of the bias (e.g., towards/against) relative to the mean of the superset and the magnitude of the bias (if present).

Figure 8B summarizes the mean biases in the three different types of distributions. Overall, we found considerable positive biases in all these distributions (mean = 7–10°, one-sample t-tests for differences from 0°: all  $t_s(19) > 6$ , all  $p_s < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , all Cohen's  $d_s > 1.3$ ). Because the physical distance between the correct answer (true mean of the subset) and the mean of the superset was 15°, these biases were smaller than the distance between the correct answer and the grand mean (one-sample t-tests for differences from 15°: all  $t_s(19) > 4$ , all  $p_s < .001$ , Bonferroni-corrected  $\alpha = 0.017$ , all Cohen's  $d_s > 0.9$ ). Therefore, we can conclude that participants' estimation of the mean orientation of the subset was highly biased towards the mean of the superset, although their estimation was not completely based on the grand mean itself. This result is consistent with our finding in Experiment 2. Moreover, the participants' biases in mean orientation estimation were greater when the shape of the length distribution was smooth narrow (mean = 10.5°,  $SD = 5.4^\circ$ ) than when it was two-peaked (mean = 7.3°,  $SD = 4.8^\circ$ ; comparison:  $t(19) = 4.01$ ,  $p < 0.001$ , Bonferroni-corrected  $\alpha = 0.017$ , Cohen's  $d = 0.90$ ). However, a significant difference was not observed between the smooth fat distribution and the two-peaked distribution.

It is worth highlighting that there was a notable difference in participants' performance for the superset conditions of Experiment 3 compared to that of Experiment 2. In Experiment 3, we observed that the response accuracy for the superset condition was comparable to that for the attend-subset condition, although worse than for the half-set condition (Fig. 8A). In Experiment 2, however, participants' centroid positioning errors in the superset condition were consistently comparable to those in the half-set-only condition, significantly better than the attend-subset condition (see Fig. 6A). One reasonable consideration regarding this discrepancy is related to the practical limitations of the centroid paradigm we acknowledged in Experiment 2. Due to the limited space of the screen size, the centroids of superset condition tended to

be centralized, compared to the positions of centroids in the attend-subset condition in which the locations of two subsets were spatially separated (see Fig. 6B). In Experiment 3 in which this limitation in accommodating all the lines with enough distances in the same visual area was resolved, participants' response accuracy for the superset condition was equivalent to the attend-subset condition, rather than the half-set condition. Therefore, when the distribution of the superset was controlled better to be equated to the other two conditions as in Experiment 3, the half-set condition showed a clear advantage over the attend-subset and the superset conditions.

Our main finding of Experiment 3 was that the two-peak distribution in the domain of length allowed for more precise and easier segmentation than the smooth narrow distribution, which in turn resulted in more precise and less biased (although not perfect) ensemble estimates in another feature domain (e.g., orientation) from the segmented subsets. From this point of view, our results replicate the results of Experiment 2, providing generalizable results across different visual features.

## General discussion

The primary goal of this study was to test how the statistical structure of multiple intermixed objects (mean, median, range, and the shape of the feature distribution) affects the ease of making categorical discriminations (e.g., segmentation or parsing) between groups of objects and modulates the extraction of overall ensemble statistics from these groups. Instead of estimating categorical effects using indirect or implicit manipulations (e.g., Khayat & Hochstein, 2019; Utochkin et al., 2018; Utochkin & Yurevich, 2016), the current study strived to make the categorization task as explicit and direct as possible. Here participants were engaged in the categorization component more directly, following our instruction for them to use an overt categorization rule (half-split by size). Here we report three main novel findings: First of all, participants could categorize individual objects into categories (subsets) – large versus small subsets or long versus short subsets – based on the ensemble central tendency (presumably the grand mean extracted from the superset, Experiment 1) and use these segmentation rules for ensemble extraction in another feature domain: segmentation by size and ensemble extraction by location (Experiment 2) or segmentation by length and ensemble extraction by orientation (Experiment 3). Second, participants' categorization performance was sensitive to and dependent on the shape of the distribution in the feature domain to be used for segmentation. Finally, the ease of segmenting and categorizing subgroups based on one feature dimension also systematically affected the precision of extraction of ensemble summary statistics from the subgroups in another feature dimension (e.g., centroid or orientation).



## The mean value as a categorical boundary

In their recent work, Khayat & Hochstein (2018, 2019) made an insightful statement that an ensemble's central tendency (e.g., the mean feature) can be implicitly extracted from a series of briefly shown items and used as the best representative of the set most likely reported as having been presented (even if it actually has not). There is a great resemblance between the retrieval of the ensemble mean and the object with high typicality in a categorization task (e.g., Khayat & Hochstein, 2019). This resemblance is suggestive of demonstrating that ensemble statistics indeed can be naturally related to categorization and that the mean object can be considered a "prototype" of ensemble members.

We suggest that our findings (especially from Experiment 1) demonstrate the flip-side of the effects found by Khayat and Hochstein (2018, 2019). The most typical representative of *all* items is, at the same time, the point of maximum ambiguity (boundary) when you need to parse these items into *separate* categories. Whereas Khayat and Hochstein demonstrated that the probability of an item to be recognized as a set member decreases as a function of its distance to the mean in the feature space (or along a typicality scale), we demonstrated that the probability of accurate categorization increases as a function of the distance to the mean. We consider that our findings mirror the patterns observed in Khayat and Hochstein's (2018, 2019) with a reversal, possibly because of the different nature of the tasks used.

Although we explicitly used the median size as a boundary in our categorization rule, our stimulation with skewed superset distributions and the resulting asymmetry of the categorization function (see Fig. 3, and also Experiment 1 for a detailed explanation) showed that observers rather relied on the grand mean as a more natural representation of the categorical boundary. The fact that they did it unintentionally, contrary to the task instruction, and without extended practice, extends our idea of the functionality of ensemble mean not only as an explicit approximate of summarized ensemble properties (Chong & Treisman, 2003) but also as a powerful tool of "naturally" organizing visual and cognitive representations for many tasks.

## Role of feature distribution: Ensemble categorization as a probabilistic process

Our results showed that the distance effects on the categorization are strongly modulated by the shapes of feature distributions of the stimuli that differed in the range and shape (e.g., smoothness). In addition, Experiments 2 and 3 consistently showed two-peaked distributions facilitated the precision of ensemble estimates of segmented, two-peaked subsets compared to the smooth narrow distributions (note that there was no consistency in the smooth fat distributions). This finding

suggests important properties of ensemble-based categorization. First of all, ensemble-based categorization does not rely on the all-or-none process of definitive classification based on the boundary rule. Rather, it is a probabilistic and integrative decision that is strongly dependent on many factors including the resemblance of the items within the same subset, deviance from the items that belong to other subsets, and the disposition of the overall group that includes all the items (e.g., superset that provides their common mean "prototype"). This idea is not novel in the categorization literature (e.g., Rosch & Mervis, 1975). Second, the probability of categorizing items as large or small depended both on absolute and relative distance from the item to the categorical boundary. The role of *absolute* distance was supported by the following findings. The smooth narrow distributions, including only sizes densely distributed around the grand mean, yielded the lowest categorization accuracy, whereas the two-peaked distributions, including only sizes substantially separated from the grand mean, yielded much better categorization accuracy. A similar pattern was also found in subset summary judgments in Experiments 2 and 3. The smooth fat distribution containing both middle and extreme sizes yielded an intermediate categorization rate. The effect of *relative* distance from the categorical boundary can be demonstrated by the fact that items with extreme sizes (smallest or largest) were categorized almost with the same accuracy regardless of the distribution range and shape.

Our findings on the roles of deviations of individual items from the grand mean in determining the ensemble-based categorization also provide insights into the concept of segmentability. Previous work of Utochkin and colleagues (Utochkin, 2015; Utochkin et al., 2018; Utochkin & Yurevich, 2016) has suggested that the shape of the feature distribution – whether it is smooth/single-peak (non-segmentable) or bumpy/several-peaks (segmentable) – is critical for determining whether all items belong to the same or different categories. Such a segmentability "rule" can be a useful heuristic for categorization as it plausibly conveys the distributional properties of objects of the same and different kinds in the real world (Utochkin, 2015). However, our new data show that ensemble categorization can be based not only on the analysis of peaks and gaps in the feature distribution, but rather can be explained by the occurrence of elements similar to a common "prototype" (ensemble mean) and, thus, causing a lot of categorical confusion. There are many such elements in smooth distributions, especially in narrow-range ones, which makes categorization into two groups more difficult. In contrast, there is less categorical confusion in the two-peaked distribution because they do not contain a lot of "prototypical" elements: Moreover, the common "prototype" in the form of the grand mean might not be encoded at all in such ensembles (see Treue, Hol, & Rauber, 2000).

## Categorization and segmentation of spatially intermixed subsets based on size

In Experiments 2 and 3, we found that an ability to accurately extract ensemble statistics was significantly impaired and biased when the items from two subsets were presented and intermixed together, compared to when a set was presented in isolation (e.g., half-set-only condition). Our results suggest, in line with a number of other studies (Inverso et al., 2016; Oriet & Brand, 2013; Utochkin et al., 2018), that many feature dimensions such as size, length, or orientations are not perfect bases for global attentional “filtering” possibly due to the inherent noise. On the other hand, many other studies demonstrating a much better ability to compute ensemble statistics independently within spatially intermixed subsets (e.g., Chong & Treisman, 2005; Halberda et al., 2006; Im & Chong, 2014, etc.) used color segmentation. The visual system appears to be better tuned to even tiny color differences to use them for global selection of spatially intermixed subsets (Sun et al., 2016a, b), although there is also an indication that parsing solely by color is still imperfect and can be improved by an additional cue such as spatial separation (see Im, Park, & Chong, 2015). Various sensory and perceptual dimensions may differ in their potential to drive global segmentation and categorization in complex scenes. To better understand the hierarchical nature of visual grouping and scene perception of more complex and realistic images, future research will need to investigate how feature-based attention in various domains guides segmentation and categorization of objects, sets, and scenes. One promising direction is to establish effective measurements that quantify how imperfect segmentation leads to noisy ensemble representations of subsets. One potential attempt would be to rely on modeling approaches and simulations (e.g., Sun et al., 2016a, b) and evaluate how ensemble representations of subsets in a visual array are degraded by mistakenly including members of the distractor set, omitting members of the target set, or occurrence of both.

In Experiments 2 and 3, we observed that the superset condition in which participants did not have to categorize individual objects into subgroups at all also showed the effects of the shapes of distribution in another feature dimension. Although the specific patterns were not completely identical, both superset conditions in Experiments 2 and 3 showed that the two-peaked distribution of irrelevant feature to the ensemble extraction task (e.g., size in Experiment 1 and length in Experiment 2) was not helpful or even detrimental, relative to one fat distribution and two-peaked distributions. This finding suggests that even when they did not have to, participants’ ensemble representations were sensitive to the shape of the feature distribution of objects in a visual image. Previous work has only examined the effects of the shape of feature distribution *within the feature*. For example, the size variance of Gaussian distribution from which sizes of individual objects

were drawn systematically affected the precision of extracting mean size (Im & Halberda, 2013) and the magnitude of adaptation aftereffect to the mean size (Corbett, Wurnitsch, Schwartz, & Whitney, 2012); and discrimination threshold for mean size increased when the two sets to be compared had different size distributions (e.g., one set of objects from two-peaked distribution vs. the other set of objects from a uniform distribution), compared to when the two sets had the same size distributions. Thus, the current finding on the systematic modulation by the distribution shape of objects of an ensemble *across* the two different features is novel.

**Hierarchical coding empowered by rapid categorization for ensembles of objects** We propose that the segmentability found in our experiments can have more far-reaching implications for understanding the organization of visual perception. Segmentability reflects a biologically justified strategy of perceiving heterogeneous items, depending on the perceptual context. When highly dissimilar and, hence, segmentable items are mixed together, the visual system tends to emphasize their differences, given that it is more likely that they represent different types of objects. However, when a smooth transition is provided between the same dissimilar features, it is likely that they are merely extreme variants of the same object type. This perceptual principle tends to reproduce the principle of physical feature distribution among natural objects. Therefore, segmentability can be considered to be a somewhat low-level basis for the rapid perceptual categorization of multiple objects (Utochkin, 2015). Each such segmentable category also can be an appropriate unit of global scene analysis.

**Acknowledgements** This study was supported by the Russian Science Foundation (grant 18-18-00334). The authors thank Vladislav Khvostov for help in running experiments.

## References

- Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research*, 83, 25–39. <https://doi.org/10.1016/j.visres.2013.02.018>
- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3), 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(4), 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 of the United States of America*, 106(18), 7345–7350. <https://doi.org/10.1073/pnas.0808981106>
- Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science*, 12(2), 157–162. <https://doi.org/10.1111/1467-9280.00327>

- Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, *120*, 150–172.
- Bacon, W. F., & Egeth, H. E. (1991). Local processes in preattentive feature detection. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 77–90. <https://doi.org/10.1037/0096-1523.17.1.77>
- Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *The Psychological Record*, *59*(2), 171–186. Retrieved from <http://opensiuc.lib.siu.edu/tpr/vol59/iss2/2/%5Cnpapers2://publication/uuid/5CB02E77-1797-45CD-8557-A91D8E46328A>
- 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>
- Bravo, M. J., & Nakayama, K. (1992). The role of attention in different visual-search tasks. *Perception & Psychophysics*, *51*(5), 465–472. <https://doi.org/10.3758/BF03211642>
- Chetverikov, A., Campana, G., & Kristjánsson, Á. (2016). Building ensemble representations: How the shape of preceding distractor distributions affects visual search. *Cognition*, *153*, 196–210. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0010027716301159>
- Chetverikov, A., Campana, G., & Kristjánsson, Á. (2017). Rapid learning of visual ensembles. *Journal of Vision*, *17*(2), 21. <https://doi.org/10.1167/17.2.21>
- Chong, S. C., & Evans, K. K. (2011). Distributed versus focused attention (count vs estimate). *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 634–638. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/wcs.136/full>
- 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](https://doi.org/10.1016/S0042-6989(02)00596-5)
- Chong, S. C., & Treisman, A. (2005). Statistical processing: computing the average size in perceptual groups. *Vision Research*, *45*(7), 891–900. <https://doi.org/10.1016/j.visres.2004.10.004>
- Corbett, J. E. (2017). The Whole Warps the Sum of Its Parts: Gestalt-Defined-Group Mean Size Biases Memory for Individual Objects. *Psychological Science*, *28*(1), 12–22. <https://doi.org/10.1177/0956797616671524>
- Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica*, *138*(2), 289–301. <https://doi.org/10.1016/j.actpsy.2011.08.002>
- Corbett, J. E., Wumitsch, 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>
- Cowan, N. (2001). The magical number 4 in short term memory. A reconsideration of storage capacity. *Behavioral and Brain Sciences*, *24*(4), 87–186. <https://doi.org/10.1017/S0140525X01003922>
- Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of Optical Society of America A*, *18*(5), 1016–1026.
- Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*(22), 3181–3192. [https://doi.org/10.1016/S0042-6989\(97\)00133-8](https://doi.org/10.1016/S0042-6989(97)00133-8)
- Epstein, M. L., & Emmanouil, T. A. (2017). Ensemble coding remains accurate under object and spatial visual working memory load. *Attention, Perception, & Psychophysics*, *79*, 2088. <https://doi.org/10.3758/s13414-017-1353-2>
- Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. <https://doi.org/10.1016/j.tics.2004.05.002>
- Fouriezos, G., Rubinfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics*, *70*(3), 456–464.
- Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*, 225–241.
- Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753. <https://doi.org/10.1016/j.cub.2007.06.039>
- Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception & Performance*, *35*(3), 718–734. <https://doi.org/10.1037/a0013899>
- Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, *17*(7), 572–576. Retrieved from <http://pss.sagepub.com/content/17/7/572.short>
- Hochstein, S., Pavlovskaya, M., Bonneh, Y. S., & Soroker, N. (2018). Comparing set summary statistics and outlier pop out in vision. *Journal of Vision*, *18*(13), 1–13. <https://doi.org/10.1167/18.13.12>
- Im, H. Y., & Chong, S. C. (2014). Mean size as a unit of visual working memory. *Perception*, *43*(7), 663–676. <https://doi.org/10.1068/p7719>
- Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception & Psychophysics*, *75*(2), 278–286. <https://doi.org/10.3758/s13414-012-0399-4>
- Im, H. Y., Park, W. J., & Chong, S. C. (2015). Ensemble statistics as units of selection. *Journal of Cognitive Psychology*, *27*(1), 114–127.
- Im, H. Y., Zhong, S., & Halberda, J. (2016). Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision Research*, *126*, 291–307. <https://doi.org/10.1016/j.visres.2015.08.013>
- Inverso, M., Sun, P., Chubb, C., Wright, C. E., & Sperling, G. (2016). Evidence against global attention filters selective for absolute bar-orientation in human vision. *Attention, Perception, and Psychophysics*, *78*(1), 293–308. <https://doi.org/10.3758/s13414-015-1005-3>
- Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. <https://doi.org/10.1038/35058500>
- Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, *290*, 91–97. doi:<https://doi.org/10.1038/290091a0>
- Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, *18*(9), 1–14. <https://doi.org/10.1167/18.9.23>
- Khayat, N., & Hochstein, S. (2019). Relating categorization to set summary statistics perception. *Attention, Perception, and Psychophysics*. <https://doi.org/10.3758/s13414-019-01792-7>
- 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. <https://doi.org/10.1167/19.9.3>
- Knierim, J. J., & Van Essen, D. C. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *Journal of Neurophysiology*, *67*(4), 961–980.
- Leib, A., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications*, *7*, 13186. doi:<https://doi.org/10.1038/ncomms13186>
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. <https://doi.org/10.1038/36846>
- Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: effects of set size and item heterogeneity on average size perception. *Acta Psychologica*, *142*(2), 245–250.

- Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision*, *15*(4):6. doi: <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*, *8*(11), 9. <https://doi.org/10.1167/8.11.9>
- Nothdurft, H.-C. (1992). Feature analysis and the role of similarity in preattentive vision. *Perception & Psychophysics*, *52*(4), 355–375. <https://doi.org/10.3758/BF03206697>
- Nothdurft, H.-C. (1993). The role of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research*, *33*(14), 1937–1958. [https://doi.org/10.1016/0042-6989\(93\)90020-W](https://doi.org/10.1016/0042-6989(93)90020-W)
- 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>
- Oriet, C., & Hozempa, K. (2016). Incidental statistical summary representation over time. *Journal of Vision*, *16*(3). <https://doi.org/10.1167/16.3.3>
- Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739–744. <https://doi.org/10.1038/89532>
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Potts, B. C., Melara, R. D., & Marks, L. E. (1998). Circle size and diameter tilt: a new look at integrality and separability. *Perception & Psychophysics*, *60*(10), 101–112.
- Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197.
- Rodriguez-Cintrón, L. M., Wright, C. E., Chubb, C., & Sperling, G. (2019). How can observers use perceived size? Centroid versus meansize judgments. *Journal of Vision*, *19*(3). <https://doi.org/10.1167/19.3.3>
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. [https://doi.org/10.1016/0010-0285\(75\)90024-9](https://doi.org/10.1016/0010-0285(75)90024-9)
- Rosenholtz, R. (2000). Search for a unique orientation among heterogeneous distracters. *Investigative Ophthalmology & Visual Science*, *41*(4).
- Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, *39*(19), 3157–3163. [https://doi.org/10.1016/S0042-6989\(99\)00077-2](https://doi.org/10.1016/S0042-6989(99)00077-2)
- Rosenholtz, R. (2001). Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal-detection theory models of search. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(4), 985–999. <https://doi.org/10.1037/0096-1523.27.4.985>
- Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*, 54–87.
- Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, *10*(14), 19. <https://doi.org/10.1167/10.14.19>
- Sun, P., Chubb, C., Wright, C. E., & Sperling, G. (2016a). Human attention filters for single colors. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(43), E6712–E6720. <https://doi.org/10.1073/pnas.1614062113>
- Sun, P., Chubb, C., Wright, C. E., & Sperling, G. (2016b). The centroid paradigm: Quantifying feature-based attention in terms of attention filters. *Attention, Perception, & Psychophysics*, *78*(2), 474–515. <https://doi.org/10.3758/s13414-015-0978-2>
- Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology Section A*, *40*(2), 201–237. <https://doi.org/10.1080/02724988843000104>
- Treue, S., Hol, K., & Rauber, H. J. (2000). Seeing multiple directions of motion-physiology and psychophysics. *Nature Neuroscience*, *3*(3), 270–276. <https://doi.org/10.1038/72985>
- Utchkin, I. S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. *Journal of Vision*, *15*(4), 8. <https://doi.org/10.1167/15.4.8>
- Utchkin, I. S., Khvostov, V. A., & Stakina, Y. M. (2018). Continuous to discrete: Ensemble-based segmentation in the perception of multiple feature conjunctions. *Cognition*, *179*, 178–191. <https://doi.org/10.1016/j.cognition.2018.06.016>
- Utchkin, 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.
- Utchkin, I. S., & Vostrikov, K. O. (2017). The numerosity and mean size of multiple objects are perceived independently and in parallel. *PLoS ONE* *12*(9), e0185452. <https://doi.org/10.1371/journal.pone.0185452>
- Utchkin, I. S., & Yurevich, M. A. (2016). Similarity and heterogeneity effects in visual search are mediated by “segmentability”. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(7), 995–1007. <https://doi.org/10.1037/xhp0000203>
- Ward, T. B. (1985). Individual differences in processing stimulus dimensions: Relation to selective processing abilities. *Perception & Psychophysics*, *37*, 471–482.
- Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238. <https://doi.org/10.3758/BF03200774>
- Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. <https://doi.org/10.1038/nature06860>

**Open Practices Statement** All experimental data are available on Open Science Framework at: <https://osf.io/grbfe/>.

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