

The effects of sampling and internal noise on the representation of ensemble average size

Hee Yeon Im · Justin Halberda

Published online: 28 November 2012
© Psychonomic Society, Inc. 2012

Abstract Increasing numbers of studies have explored human observers' ability to rapidly extract statistical descriptions from collections of similar items (e.g., the average size and orientation of a group of tilted Gabor patches). Determining whether these descriptions are generated by mechanisms that are independent from object-based sampling procedures requires that we investigate how internal noise, external noise, and sampling affect subjects' performance. Here we systematically manipulated the external variability of ensembles and used variance summation modeling to estimate both the internal noise and the number of samples that affected the representation of ensemble average size. The results suggest that humans sample many more than one or two items from an array when forming an estimate of the average size, and that the internal noise that affects ensemble processing is lower than the noise that affects the processing of single objects. These results are discussed in light of other recent modeling efforts and suggest that ensemble processing of average size relies on a mechanism that is distinct from segmenting individual items. This ensemble process may be more similar to texture processing.

Keywords Texture · Scene perception · Visual perception

Outside of the laboratory, our visual system often comes across a cluttered scene with thousands of objects. But, unlike a random scatter of items, this visual world often contains systematic structure and redundancy—such as collections of similar items (e.g., books on a shelf or faces in a

crowd). To make the most of its limited resources, the visual system may exploit this systematicity. One strategy is to rely on redundancy within the scene to collapse information across similar objects in order to represent the general description of the group rather than information for every single item in the group. This summary representation or statistic has been called an “ensemble feature.” For example, laboratory experiments suggest that humans can represent with little effort the mean size (Ariely, 2001; Chong & Treisman, 2003; Im & Chong, 2009), general direction of motion (Watamaniuk & Duchon, 1992), general location or “centroid” (Alvarez & Oliva, 2008), and approximate number of items in a group (Halberda, Sires, & Feigenson, 2006), as well as the average emotion of faces in a crowd (Haberma & Whitney, 2009, 2010).

Much of the controversy surrounding this emerging literature on ensemble features has focused on the question of whether representing such features requires a mechanism distinct from our ability to represent individual objects. For example, Myczek and Simons (2008) argued that relying on one or two individual items from a display to estimate the ensemble average of the entire display could result in levels of performance similar to those observed for human subjects in ensemble average size tasks. This would amount to there being no distinct mechanism for representing ensemble average size—that is, no mechanism distinct from simply sampling a few individual objects from the set. Average size was one of the first ensemble features to be investigated (Ariely, 2001), and it has perhaps drawn the most skepticism and widest criticism—owing, perhaps, to the suggestions that early visual areas have no “size-tuned” neurons (Myczek & Simons, 2008; Simons & Myczek, 2008) and that representing object size is traditionally thought to require selecting individual objects from the background (e.g., Bundesen & Larsen, 1975; Cave & Kosslyn, 1989). At the moment, the process that enables

H. Y. Im (✉) · J. Halberda
Department of Psychological and Brain Sciences, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA
e-mail: heeyeon.im@jhu.edu

the coding of average size remains to be determined, and criticism has primarily focused on the possibility that performance derives from object-based sampling strategies.

Another possible model for how humans represent ensemble features is texture processing—in which early feature information is pooled across regions without requiring segmentation of individual objects (Dakin & Watt, 1997; Malik & Perona, 1990; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001). It may be that information such as the approximate number of items and the average size of the items can be estimated by pooling early featural evidence such as filled region and density, without actually segmenting individual items (Tibber, Greenwood, & Dakin, 2012).

It remains to be determined whether representing ensemble features such as average size requires mechanisms similar to or distinct from those employed when selecting individual items or processing textures. One approach to informing such debates is to measure hallmarks or signatures of visual processing that may distinguish these modes of processing—such as internal noise and sample size. For example, several authors have suggested the possibility that the internal noise that affects ensemble representation may be lower than the internal noise that affects individual object representation, perhaps owing to ensembles relying on a distinct representational route (Alvarez, 2011; Ariely, 2001, 2008). If the internal noise that affects the representation of average size were to be significantly lower than the internal noise that affects the representation of individual object size, this would suggest that representing individual sizes is not the mechanism relied on for average size representation (though the observed thresholds could also be lower due to increased numbers of samples). Similarly, if the number of objects that appear to be averaged during average-size processing exceeded the broadly assumed limit of three to four objects for parallel attention (Oksama & Hyönä, 2004; Pylyshyn & Storm, 1988; Scholl, 2001), this would also suggest that representing individual sizes is not the mechanism relied on for average size representation. Here we explored both of these aspects of average-size representation to test the plausibility of object-based sampling strategies and to address several outstanding questions related to ensemble feature processing.

Several approaches have been undertaken to test the plausibility of object-based sampling strategies for ensemble processing. For average size, Chong, Joo, Emmanouil, and Treisman (2008) demonstrated that mixing trials together that require different object-based sampling strategies does not lead to a reduction in human performance overall, suggesting that either subjects do not use such strategies or that they are able to rapidly determine the optimal strategy for each display and switch strategies from trial to trial, given only 200 ms of display time. Because the latter option seems unlikely, the authors suggested that subjects must be relying

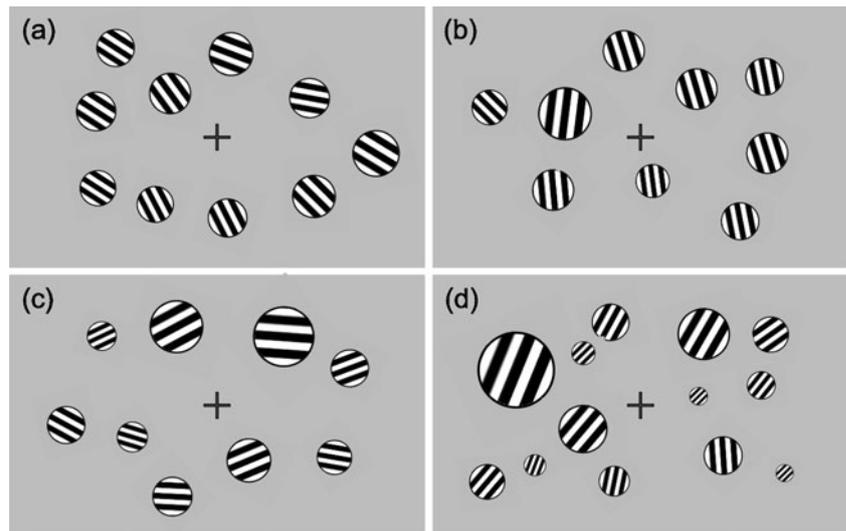
on a mechanism distinct from selecting individual objects. In average-orientation processing (Parkes et al., 2001), humans appear to be able to extract an ensemble average even when individual items are too crowded to allow for successful discrimination of individual orientations. This too suggests that ensemble features might be represented without selecting individual items. And, for representing the average emotion of a collection of faces (Haberma & Whitney, 2009, 2010) and the average location of a set of dots (Alvarez & Oliva, 2008), measuring subjects' response variability for recalling individual items has been used to estimate the internal noise that might affect ensemble processing and, via Monte Carlo simulation, to estimate the number of items that would need to be averaged in order to attain human-like levels of performance in ensemble feature tasks. It was argued that these estimates (e.g., eight objects for the centroid, and seven for faces) exceeded the expected limits of object-based parallel attention (Oksama & Hyönä, 2004; Pylyshyn & Storm, 1988; Scholl, 2001), suggesting that subjects did not rely on this type of object-based sampling strategy. Because previous methods have relied on the observed discrimination thresholds or response variability when subjects must represent, for instance, the size of a single item as a way of estimating the internal noise that might affect ensemble processing (Alvarez & Oliva, 2008; Haberman & Whitney, 2009, 2010), the question of whether the internal noise that affects ensemble representation is lower than the internal noise that affects individual-object representation has not yet been tested. The opportunity has arisen to use variance summation modeling to estimate both the internal noise and the sample size that affect ensemble feature processing from within the ensemble feature task itself.

The variance summation approach enables one to measure how observers' response variability may change as a function of stimulus variability. Intuitively, it should be easier for an observer to estimate the average size of an ensemble when variability in sizes is low, and performance should become poorer as variability increases (Fig. 1). The variance summation approach exploits a noise analysis that assumes the additivity of variances on the basis of convolution to model the data (Eq. 1) in terms of the local and global limits of the system and external noise. In the variance summation model, the local and global limits are characterized by the internal noise of the ensemble averaging mechanism and the sample size that the observer gathers from the stimulus, and the external noise is assessed by the variability embedded within the stimulus, such that

$$\sigma_{\text{obs}} = \sqrt{\sigma_{\text{int}}^2 + \sigma_{\text{ext}}^2/n}, \quad (1)$$

where σ_{obs} is the observed threshold, σ_{int} is the intrinsic

Fig. 1 Displays: Cartoon examples of arrays from Experiment 1a (ensembles) with different levels of external noise—(a) 1°, (b) 3°, (c) 6°, and (d) 10°



or internal noise, σ_{ext} is the external noise, and n is the number of samples being employed. In the present experiments, σ_{ext} is the variability of the sizes within an ensemble of sine gratings (Fig. 1), which is under the experimenter's control (i.e., the Gaussian distribution of object sizes in the display). σ_{int} is the noise or error inside the head of the observer (also assumed to be Gaussian) that affects their estimate of the ensemble average. Thus, Eq. 1 is simply a way of combining these two Gaussian sources of noise in order to fit the observed sensitivity of the observer.

By measuring observed thresholds (σ_{obs}) at multiple levels of external noise (σ_{ext}), it is possible to fit values for the internal noise affecting the ensemble averaging mechanism (σ_{int}) and the number of samples (n) that the observer seems to rely on (i.e., the number of individual gratings a subject averages). When the external variability (σ_{ext}) is lower than the internal noise (σ_{int}), the observed threshold (σ_{obs}) will derive almost entirely from the internal noise. But, as the external variability (σ_{ext}) increases, it will eventually come to exceed the internal noise (σ_{int}) to become the dominant force determining the observed threshold (σ_{obs}). Intuitively, the observed threshold will not increase rapidly until the external noise is greater than the internal noise.

Sample size will function to raise or lower the observed thresholds (σ_{obs}) across all levels of external noise (σ_{ext}), as pooling evidence from greater numbers of items will result in reduced observed thresholds (σ_{obs}). The pattern of reduction in the observed thresholds due to increased sample size is distinct and separable from the reduction that occurs from reduced internal noise.

Our approach here was inspired by previous research in which variance summation modeling has been used

to estimate the internal noise and efficiency (i.e., sample size) of texture discrimination mechanisms (Beaudot & Mullen, 2005; Dakin, 2001; Dakin, Bex, Cass, & Watt, 2009; Demanins, Hess, Williams, & Keeble, 1999; Heeley, Buchanan-Smith, Cromwell, & Wright, 1997). For the present case, a benefit of this approach is that, once generated, estimates of the internal noise affecting ensemble processing can be compared to behavioral estimates of the internal noise affecting individual-object processing to address the question of whether or not the internal noise for ensemble processing is lower than that for processing individual items.

The approach taken here for estimating internal noise and sample size can be used to study any ensemble process that pools evidence across multiple samples. In this way, the present work aims to introduce an approach (i.e., variance summation modeling) that can be used to study ensemble feature processing more generally. In the present case, we use this approach to address two important theoretical questions: Is the internal noise for ensemble processing lower than the internal noise for processing individual items, and does the number of samples required by ensemble processing exceed the three- to four-item limit of object-based attention?

The experiments

We relied on a standard two-alternative forced choice psychophysical discrimination task in which subjects had to judge which of two briefly flashed arrays had either the larger individual size or the larger average size. Experiments 1a (ensembles) and 1b (single gratings) were run within subjects, so as to allow comparisons across the two tasks.

Method

Subjects

The two authors and 14 naive subjects participated in the experiment. All of the subjects had normal or corrected-to-normal vision. The naive subjects received course extra credit for participating.

Apparatus and stimuli

The stimuli were generated using MATLAB software, together with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997), and were displayed on an LCD monitor driven by a Macintosh iMac computer (the viewable area was a gray central square window with a 17-in. diagonal). The subjects were seated approximately 50 cm from the screen and viewed the display binocularly. At this viewing distance, each pixel was approximately 0.04° of visual angle, and each grating subtended between 1.6° and 4.0° of visual angle. The stimuli were presented on a gray background and consisted of one or more sinusoidal gratings with a spatial frequency of 4 cycles/deg and a Michelson contrast of 99.8% (this is within the range of spatial frequencies that produce optimal orientation discrimination [Dakin, 2001], so as to allow us to compare performance with orientation processing in future work). In Experiments 1a (ensembles), multiple gratings (9, 11, 13, 16, 19, or 23 gratings) were randomly located on the display, subtending $56^\circ \times 40^\circ$ of visual angle. In Experiments 1b (single gratings), only one grating appeared within this same viewing area. The locations of gratings varied across the two stimulus arrays in order to minimize the masking of stimuli in the second flash by those in the first flash.

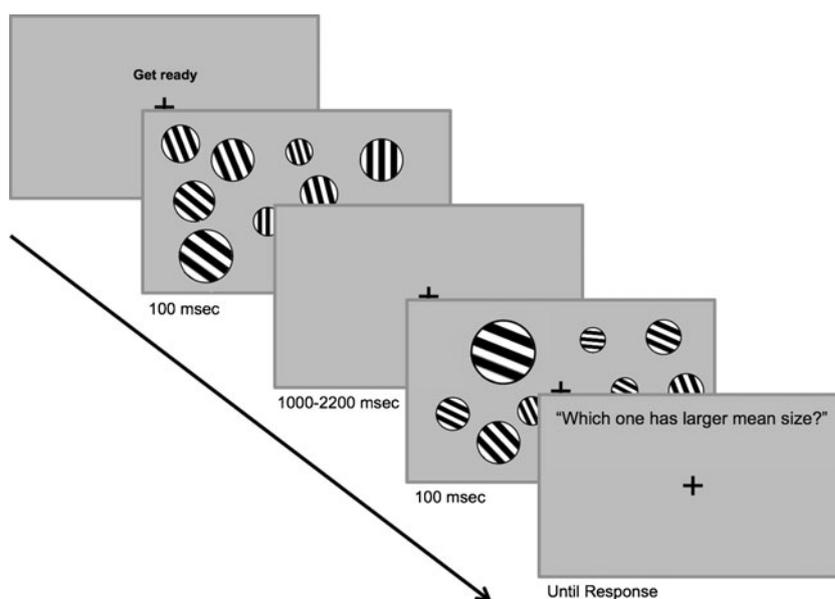
To focus on the ensemble processing of distinct objects, we ensured that each item in the array was a segmentable individual object. Each object on the display had a clearly drawn border that was salient in order to avoid blur and reduce the blending of gratings and background. Because crowding by adjacent stimuli occurs in a compulsory manner and may be equivalent to perceiving texture (Parkes et al., 2001), we also ensured that adjacent gratings were separated from each other by at least half of their eccentricity in the display—that is, crowding in foveal vision only occurs over very small distances (2–6 arcmin; Toet & Levi, 1992), whereas crowding in peripheral vision occurs over larger distances, roughly at about half of the eccentricity (Pelli & Tillman, 2008). This ensured that our displays were not crowded.

Procedure

The basic procedure required subjects to view two brief displays (100 ms each), one after the other, and then to judge which display, the first or second, contained either the larger average size (Exp. 1a, ensembles) or the larger individual grating (Exp. 1b, single gratings; see Fig. 2). Auditory feedback for errors was provided throughout. The short exposure duration of 100 ms was chosen to prevent scanning eye movements (Morgan, Ward, & Castet, 1998). The interstimulus interval was randomly varied from 1,000 to 2,200 ms, making the onset of the second flash unpredictable, in order to disrupt strategic planning in preparation for the second display. This delay also reduced any afterimage effects of the first stimulus display, as well as possible effects of the apparent rotation of the second display from the first.

We relied on a standard adaptive procedure (Watson & Pelli, 1983) to identify each subject's threshold. Six interleaved QUEST routines of 40 trials each were run for each

Fig. 2 Method: A sample trial from Experiment 1a (ensembles)



threshold, and the final threshold estimate was obtained by fitting the data with a Weibull function and evaluating thresholds at 75 % correct responses. The Weibull function was fitted to the behavioral data using the Psignifit toolbox, Version 3.0, for Python, which implemented the maximum likelihood method described by Wichmann and Hill (2001).

On each trial, the QUEST procedure determined the sizes for the gratings to be displayed in order to determine the size difference that resulted in 75 %-correct performance for the subject. There was a standard size and a comparison size for each trial. One of the two displays (first or second) always showed the standard size of 2.8° in diameter.

In Experiment 1a (ensembles), we systematically varied the external variability of sizes among the items within each ensemble. The size for each grating within a display was randomly drawn from a Gaussian distribution centered on the size that QUEST had specified for that display (e.g., 2.8° in diameter, or a comparison size) with a standard deviation ($\sigma = 1^\circ, 3^\circ, 6^\circ, \text{ or } 10^\circ$). These *SDs* were varied across four separate blocks of trials, their order randomized across subjects. We ensured that the individual sizes of gratings in each display adhered to the specified mean and *SD* for each trial, with tolerances of $\pm 0.16^\circ$ (i.e., four pixels) for the mean and $\pm 0.02^\circ$ for the *SD*. Though this restriction violated the assumptions of random sampling, the deviation from random sampling was small given the set sizes that we used, and this restriction ensured that the trials within each block accurately reflected the target mean and *SD* while still allowing for sufficient variability across trials to fit the variance summation model. To provide an accurate fit to the data, we ensured that all analyses took into account the actual average size of the individual elements presented to the subjects, instead of the values suggested by QUEST. Since the average orientation of the gratings was not a relevant feature in this experiment, the average orientation of the gratings for each display varied randomly among six values (30°, 60°, 90°, 120°, 150°, and 180°).

In Experiment 1b (single gratings), we measured size discrimination thresholds for a single grating at the standard size of 2.8° in diameter. We included two different blocks that varied the locations of the single gratings, one in a foveal region (4° × 4° around the fixation cross) and one in the periphery (from 56° × 40° of visual angle, but never appearing within the 4° × 4° foveal region). Including two blocks allowed us to estimate the reduction of size sensitivity as a function of eccentricity.

Each of the six blocks (four blocks for Exp. 1a and two blocks for Exp. 1b) lasted approximately 15 min, and the order of these blocks was randomized across subjects. All blocks were run during a single session.

Results

In Experiment 1a (ensembles), we systematically manipulated the variability of the sizes within each ensemble as a source of external noise (Fig. 1). If ensemble representations pool evidence across items that vary in size, the observed thresholds should increase systematically with increasing external noise. Figure 3 displays the observed thresholds (σ_{obs}). The thresholds smoothly increased as external variability (σ_{ext}) increased.

In an ensemble process that pools evidence across multiple samples, the specific pattern of increase in the observed thresholds as a function of increasing external noise should be systematically related to a relationship between sample size (n) and the internal (σ_{int}) and external (σ_{ext}) noise. This relationship can be formalized by the additivity of variances, as in Eq. 1 (Beaudot & Mullen, 2005; Dakin, 2001; Dakin et al., 2009; Demanins et al., 1999; Heeley et al., 1997). This approach has been used successfully to estimate the internal noise and sample size for average-orientation processing (Beaudot & Mullen, 2005; Dakin, 2001; Heeley et al., 1997). In this model, the manner in which the average-size thresholds (σ_{obs}) increase as the external variability in sizes increases (σ_{ext}) can be determined by a summation of noise processes.

We fit the data from the ensemble size blocks for each subject separately using the variance summation model (Eq. 1) to obtain estimates of the internal noise and the number of samples involved in the averaging process using least-squares estimation. The group fit can be seen in Fig. 4, where the observed thresholds from Fig. 3 are reprinted as data points and the model fit is a smooth curve. The model provided an accurate fit to the subjects' performance as a function of bandwidth ($R^2 = .99, p < .01$).

One of our central interests was in Myczek and Simons's (2008) suggestion that only one or a few individual objects need to be sampled in order to attain the performance of human subjects in average-size discrimination tasks. The group fit from the variance summation model in Fig. 4

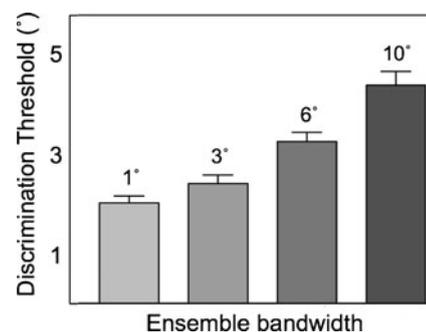


Fig. 3 Results: Discrimination thresholds for the ensemble conditions at each bandwidth, averaged across the 16 subjects. Error bars indicate the standard errors of the means

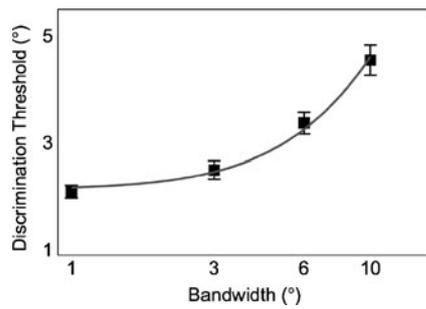


Fig. 4 Results and model: Discrimination thresholds for the ensemble conditions at each bandwidth, with the curve that indicates the average fit from the variance summation model

determined the estimate of 7.0 samples from each display. Table 1 presents the number of samples and the internal noise determined for each subject. An estimate of 7.0 samples per display exceeds the widely discussed (but not uncontroversial) estimate of a three- to four-item object-based limit of parallel attention (Oksama & Hyönä, 2004; Pylyshyn & Storm, 1988; Scholl, 2001). This analysis suggests that Myczek and Simons underestimated the number of samples that subjects rely on in ensemble feature tasks of average-size processing. One reason for their lower estimate

may be that their simulations did not take internal noise (σ_{int}) into account (Ariely, 2008; Haberman & Whitney, 2010). Estimating the internal noise that affects the processing of ensemble average size was our second major focus.

If the representation of ensemble features relies on sampling individual objects and then averaging them—and perhaps throwing away the original samples (Alvarez, 2011; Ariely, 2001)—then the internal noise (σ_{int}) estimated by the variance summation model should match the observed thresholds for processing individual items. If, instead, the ensemble pooling process relies on estimating scene statistics without individuating items (e.g., perhaps through mechanisms similar or identical to texture processing), then the internal noise that affects this process may be distinct from the noise affecting individual object representations. There has been some suggestion in the literature that the internal noise affecting ensemble averaging may be lower than the internal noise affecting representations of individual items (Alvarez, 2011; Alvarez & Oliva, 2008; Ariely, 2001; Chong & Treisman, 2003; Haberman & Whitney, 2009). Such suggestions were motivated by evidence that the observed thresholds in an ensemble averaging task tended to be lower than the observed thresholds for identifying individual items. But, observed

Table 1 Fitted parameters from the variance summation model (two leftmost columns) and the internal noise values empirically measured in the single-grating displays (two rightmost columns)

Subject	VSM		Single grating	
	Ensemble # of sample	Ensemble Internal noise	Periphery Internal noise	Fovea Internal noise
J.H.	4.55	1.80	4.41	1.91
H.I.	3.68	1.69	3.40	1.65
Sub3	7.86	2.47	3.08	2.15
Sub4	6.27	2.60	2.42	2.42
Sub5	9.86	1.54	6.74	1.89
Sub6	7.60	1.95	4.04	2.88
Sub7	2.61	1.92	4.85	2.14
Sub8	12.95	2.09	3.48	2.91
Sub9	8.60	1.58	4.43	1.25
Sub10	6.88	2.95	2.35	3.12
Sub11	4.78	2.24	4.08	2.20
Sub12	4.59	3.04	7.08	3.90
Sub13	16.85	1.08	2.53	1.07
Sub14	6.33	3.03	4.14	3.16
Sub15	3.56	2.64	3.07	3.00
Sub16	9.36	1.50	3.32	1.12

thresholds can be lower due either to decreased internal noise or to increased numbers of samples, and the previous work was unable to disentangle the potential contributions of these two sources. The variance summation model allows us some handle on this question, as it allows us to measure the contributions of each of these factors from performance within the ensemble task itself.

We determined subjects' thresholds for processing individual gratings within both the fovea region and the periphery. Internal noise for ensemble processing across the entire display (Table 1) was significantly lower than the observed thresholds for discriminating single gratings in the periphery [$t(15) = 4.77, p < .01$] and marginally lower than the thresholds for discriminating single gratings in the fovea [$t(15) = 1.59, p = .13$] (Table 1). Importantly, given the crowding controls in our ensemble displays, only one or two gratings could appear within the fovea region in ensemble displays—the remaining gratings would have appeared in the peripheral region (see the screenshot cartoons in Table 1). Because subjects relied on many more than one or two gratings during the ensemble feature task (Table 1), in order to approach the performance of the human subjects, any object-based sampling strategy would need to assume fovea-level noise across all sampled items, not just for the one or two that happened to fall within the fovea region (or, it would need to sample even more items than was suggested by the variance summation model; Table 1). Because the internal noise for representing individual gratings increases as one moves into the periphery (e.g., note the differences in the observed thresholds for the fovea and periphery in Table 1), it would appear that subjects do not rely on selecting and averaging individual gratings during the ensemble feature task.

Thus, variance summation modeling of performance in an average-size ensemble feature task suggests that the number of samples required for ensemble processing is greater than one or two items and that the internal noise affecting ensemble processing is lower than the internal noise for processing a single item presented in the periphery, and marginally smaller than the internal noise for processing a single item presented within the fovea.

Discussion

We systematically manipulated the external variability of sizes within a set of sine gratings in an ensemble feature task in order to empower variance summation modeling to estimate the sample size (n) and internal noise (σ_{int}) that affect subjects' processing of average-size information. We also compared these estimates to the observed thresholds for processing single gratings. We found that subjects appear to rely on many more than one or two individual gratings when

representing the average size of items in an ensemble (Table 1). We also found that the internal noise affecting the ensemble process is slightly lower than the internal noise that affects the representation of individual item sizes within the fovea (Table 1). These results suggest that ensemble processing relies on a mechanism that is distinct from the processing of single items, and variance summation modeling provides a means for studying internal noise and sampling procedures on the basis of performance within the ensemble feature task itself. We discuss these results in light of other recent modeling efforts.

Synthesizing our present study results with previous work on ensemble processing that has relied on Monte Carlo simulations provides greater insight into how sampling and internal noise may affect ensemble processing. Previous models (Haberman & Whitney, 2010; Myczek & Simons, 2008) can be distinguished on the basis of whether or not they included internal noise and whether the sampling was assumed to be random with or without replacement. In Fig. 5, we have placed these models in a 2×2 matrix that delineates how they differ from each other. We simulated each of these models in order to explore how observed thresholds may change with sample size. Figure 6 displays the average thresholds generated from a series of Monte Carlo simulations for each of the models in Fig. 5. In essence, these Monte Carlo simulations involve allowing a computer program to perform the ensemble size discrimination task by taking a sample of the gratings present in each display, applying or not applying internal noise to each sample, and then averaging these samples to estimate the average size of the display. Each simulation was a type of ideal observer analysis, in that only external noise, sampling, and internal noise could affect the model's performance—both the averaging process and the decision process were assumed to be without errors and without added noise (see also Haberman & Whitney, 2010; Myczek & Simons, 2008). For each simulation, the model completed 240 trials within the QUEST procedure, and QUEST provided an estimated threshold for that run—that is, the same methods that we used for the adults in Experiment 1a.

Previous Monte Carlo modeling efforts have relied on the internal noise for representing single items in order to

		Implementation of Internal noise	
		Yes	No
Sampling with Replacement	No	Haberman & Whitney, 2010	Myczek & Simons, 2008
	Yes	Variance Summation Model (Dakin, 2001; Current study)	No citation

Fig. 5 Comparison of existing models of ensemble processing

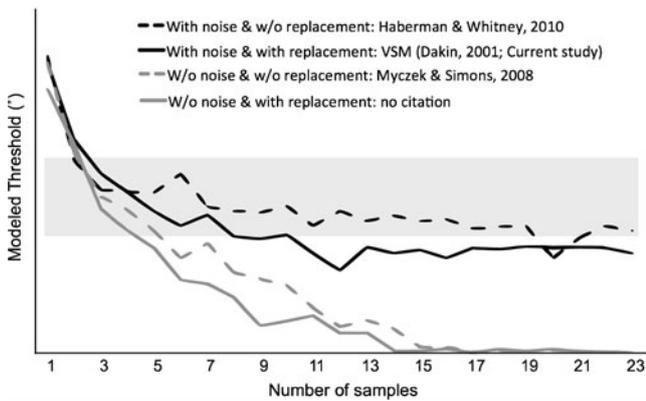


Fig. 6 Monte Carlo simulations: The average observed thresholds for discriminating ensemble average size generated by Monte Carlo simulations of existing models, with internal noise set to levels equivalent to the 16 subjects' single-grating thresholds and external noise set to $\sigma = 6^\circ$. The gray-shaded region indicates the mean (center of region) and standard deviation of human performance across the 16 subjects from the ensemble feature blocks at $\sigma = 6^\circ$

estimate the effects of sampling (Alvarez & Oliva, 2008; Haberman & Whitney, 2010). For comparison, we took this same approach and ran 16 simulations for each model, fixing the external noise to 6° and fixing the internal noise to the thresholds from the single-grating fovea blocks from the 16 subjects in Experiment 1b (note that the relative ordering of performance for these models would remain unchanged across all levels of internal noise >0 , making the selection of this parameter value somewhat arbitrary for the present purposes). The curves in Fig. 6 represent the observed thresholds averaged across the 16 simulations for each model. The standard deviation in human performance that we observed on these same trials for the ensemble feature blocks with $\sigma = 6^\circ$ is plotted (gray-shaded area), centered on the mean for human performance. From the simulations, it is clear that a major factor affecting the models is whether or not internal noise is implemented. When internal noise is included (e.g., Dakin, 2001; Haberman & Whitney, 2010; present experiments), the model's performance gracefully declines through the range of human performance. When it is not included (e.g., Myczek & Simons, 2008), the modeled threshold rapidly declines and soon outstrips human performance. The simulations in Fig. 6 also demonstrate the effects of sampling either with (e.g., variance summation model: Dakin, 2001; present experiments) or without (e.g., Haberman & Whitney, 2010; Myczek & Simons, 2008) replacement, though this effect is less dramatic than the effect of including internal noise. The reduced thresholds observed for sampling with replacement are due to the structure of the arrays—which present subjects with a finite set from a Gaussian distribution of sizes—in which sampling with replacement allows a model to sample more frequently from

the more-frequent items close to the true mean. Such effects will be reduced if the number of items in the array increases or if the distribution of the array becomes more uniform. Sampling with replacement is consistent with the assumptions of the variance summation model presented here, as well as with previous work in texture processing in which what has counted as a sample need not be a discrete individual object. In such cases, the estimates of sample size generated by the variance summation model can be understood to be estimates of the efficiency of the pooling process—for instance, in texture processing, the estimates of sample size from the variance summation model have been understood as estimates of the efficiency of evidence gathering, even when no segmentable objects are present in the stimulus (a collection of overlapping oriented Gabor patches; Beaudot & Mullen 2005; Dakin, 2001). To the extent that ensemble processing is similar to texture processing, the sampling and internal noise that characterize ensemble processing may be more similar to those of texture discrimination than of individual-object segmentation. In contrast, work from the ensemble feature literature has often taken a sample to be one of the segmentable objects within the ensemble and has implemented sampling without replacement (e.g., Haberman & Whitney, 2010; Myczek & Simons, 2008). We believe that extracting ensemble features may not rely on segmenting individual items, but rather may rely on mechanisms more similar to texture processing—a suggestion that has also been made elsewhere (Balas, Nakano, & Rosenholtz, 2009; Dakin, Tibber, Greenwood, Kingdom, & Morgan, 2011; Freeman & Simoncelli, 2011; Haberman & Whitney, 2010; Parkes et al., 2001).

While we have known for some time that the visual system can discriminate textures and can also segment individual objects from a background, ensemble feature processing might seem to fall somewhere between these two abilities and require pooling evidence across a group of clearly distinct but similar-looking items—for instance, representing the mean size, approximate number, average emotion (of faces), and centroid of multiple items. Here, we have used variance summation modeling to estimate the internal noise and sampling that affect ensemble processing. Because this method can generate estimates of internal noise and sample size from performance within the ensemble feature task itself, it has empowered the discoveries that subjects rely on many more than one or two samples when processing ensemble average size and that the internal noise that affects average size processing is markedly lower than the internal noise for processing a single item in the periphery, and even somewhat lower than the internal noise for processing a single item in the fovea. This

suggests that the mechanisms that support ensemble average-size representation are distinct from those that support our ability to segment and attend individual items, but it remains to be determined whether these ensemble mechanisms may be similar or identical to those of texture processing.

References

- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*, 122–131. doi:10.1016/j.tics.2011.01.003
- Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*, 392–398. doi:10.1111/j.1467-9280.2008.02098.x
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*, 157–162.
- Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & Psychophysics*, *70*, 1325–1326. doi:10.3758/PP.70.7.1325
- Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 1–18. doi:10.1167/9.12.13
- Beaudot, W. H. A., & Mullen, K. T. (2005). Orientation selectivity in luminance and color vision assessed using 2-D band-pass filtered spatial noise. *Vision Research*, *45*, 687–696.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436. doi:10.1163/156856897X00357
- Bundesden, C., & Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 214–220.
- Cave, K. R., & Kosslyn, S. M. (1989). Varieties of size-specific visual selection. *Journal of Experimental Psychology: General*, *118*, 148–164.
- Chong, S. C., Joo, S. J., Emmanouil, T.-A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics*, *70*, 1327–1334. doi:10.3758/PP.70.7.1327
- Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*, 393–404. doi:10.1016/S0042-6989(02)00596-5
- Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America: A*, *18*, 1016–1026.
- Dakin, S. C., Bex, P. J., Cass, J. R., & Watt, R. J. (2009). Dissociable effects of attention and crowding on orientation averaging. *Journal of Vision*, *9*(11), 1–16. doi:10.1167/9.11.28
- Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences*, *108*, 19552–19557. doi:10.1073/pnas.1113195108
- Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*, 3181–3192.
- Demianin, R., Hess, R. F., Williams, C. B., & Keeble, D. R. (1999). The orientation discrimination deficit in strabismic amblyopia depends upon stimulus bandwidth. *Vision Research*, *39*, 4018–4031.
- Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*, 1195–1201.
- Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 718–734. doi:10.1037/a0013899
- Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*, 1825–1838. doi:10.3758/APP.72.7.1825
- Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, *17*, 572–576.
- Heeley, D. W., Buchanan-Smith, H. M., Cromwell, J. A., & Wright, J. S. (1997). The oblique effect in orientation acuity. *Vision Research*, *37*, 235–242.
- Im, H. Y., & Chong, S. C. (2009). Computation of mean size is based on perceived size. *Attention, Perception, & Psychophysics*, *71*, 375–384. doi:10.3758/APP.71.2.375
- Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America: A*, *7*, 923–932.
- Morgan, M. J., Ward, R. M., & Castet, E. (1998). Visual search for a tilted target: Tests of spatial uncertainty models. *Quarterly Journal of Experimental Psychology*, 347–370.
- Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, *70*, 772–788. doi:10.3758/PP.70.5.772
- Oksama, L., & Hyönä, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition*, *11*, 631–671.
- Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*, 739–744.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442. doi:10.1163/156856897X00366
- Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*, 1129–1135. doi:10.1038/nm.2187
- Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 179–197. doi:10.1163/156856888X00122
- Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*, 1–46. doi:10.1016/S0010-0277(00)00152-9
- Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics*, *70*, 1335–1336. doi:10.3758/PP.70.7.1335
- Tibber, M. S., Greenwood, J. A., & Dakin, S. C. (2012). Number and density discrimination rely on a common metric: Similar psychophysical effects of size, contrast, and divided attention. *Journal of Vision*, *12*(6), 1–19. doi:10.1167/12.6.8
- Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, *32*, 1349–1357.
- Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research*, *32*, 931–942.
- Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*, 113–120. doi:10.3758/BF03202828
- Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*, 1293–1313. doi:10.3758/BF03194544