

Statistical processing: Not so implausible after all

SANG CHUL CHONG AND SUNG JUN JOO
Yonsei University, Seoul, Korea

AND

TATIANA-ALOI EMMANOUIL AND ANNE TREISMAN
Princeton University, Princeton, New Jersey

Myczek and Simons (2008) have shown that findings attributed to a statistical mode of perceptual processing can, instead, be explained by focused attention to samples of just a few items. Some new findings raise questions about this claim. (1) Participants, given conditions that would require different focused attention strategies, did no worse when the conditions were randomly mixed than when they were blocked. (2) Participants were significantly worse at estimating the mean size when given small samples than when given the whole display. (3) One plausible suggested strategy—comparing the largest item in each display, rather than the mean size—was not, in fact, used. Distributed attention to sets of similar stimuli, enabling a statistical-processing mode, provides a coherent account of these and other phenomena.

There exists in perception a puzzling and striking discrepancy between two sets of findings: evidence for severe limits to attention, on the one hand, and evidence of a rich and detailed representation of the visual scene, on the other. For example, the span of apprehension is limited to three or four items (Woodworth, 1938), conjunctions of features in many search tasks seem to require serial search (Treisman & Gelade, 1980), and change detection in natural scenes often takes many seconds (Rensink, O'Regan, & Clark, 1997). On the other hand, people can get the gist of a scene in very brief presentations (e.g., Oliva & Torralba, 2006; Potter & Faulconer, 1975) and can detect instances of a semantic category, such as an animal, with little cost to a concurrent attention-demanding task (Li, VanRullen, Koch, & Perona, 2002; Thorpe, Fize, & Marlot, 1996). One idea that might contribute toward reconciling these apparently contradictory sets of findings is the hypothesis that we summarize complex scenes by rapidly forming statistical descriptions of sets of similar items (Ariely, 2001; Chong & Treisman, 2003; Treisman, 2006). The hypothesis is that by extracting the average value on a number of dimensions, as well as the range and variance, we can form an economical description without individuating each member of the set. Chong and Treisman (2003) proposed a separate mode of processing that accompanies distributed attention to sets of similar items, as contrasted with focused attention to individual items. We reported a number of experiments in which participants were able to identify the mean size of heterogeneous sets of circles with very brief exposures and with no effect of their number.

Myczek and Simons (2008) have questioned the plausibility of averaging as a processing operation. Of course,

performance in statistical processing is seldom perfect. Our suggestion is not that every item is correctly included in the calculation of the mean. This would not be a plausible claim. We suggest, instead, that when attention is distributed over a set of similar items, participants take in the properties of the set as a whole, rather than those of individual items within it, and that an estimate of the mean size is one such property. We believe that there is a meaningful distinction between attending to a set as a whole and attending to a small subset of individuated items and that this distinction may explain the dissociations that Ariely (2001) and we found between the perception of individual objects and the perception of group properties.

Myczek and Simons (2008) questioned the evidence on which we based this hypothesis. They pointed out that many of the experiments could, instead, be explained by a sampling procedure in which participants direct focused attention to one to four items, rather than averaging them all. They showed, with a number of simulations, that the level of performance reported in the studies so far could have been achieved by averaging just a small sample of each display. For cases in which the sample had to be larger than four (the putative limit of focused attention), they proposed some alternative strategies that could be used to get the average from, at most, four items. They did not claim that these strategies are, in fact, used, but they showed that they could be and, therefore, questioned the need to postulate a separate mode of statistical processing. This is an important claim, and it has led us to look for arguments and new evidence with which to check our hypothesis.

This article will be divided into two sections. First, we will briefly question the adequacy of Myczek and Simons's

(2008) account of the previously available data, and second, we will report some new findings that we think raise further problems for an account based exclusively on focused attention.

Earlier Studies of Statistical Processing

The first set of published data that seem to raise problems for the focused attention sampling strategy are found in Chong and Treisman (2005a), where the deployment of attention was specifically manipulated to see the effects on the averaging task. We showed that performance in the statistical-averaging task was more compatible with distributed or global attention than with focused attention. One experiment used the difference previously found between search for a closed circle among circles with gaps, which requires focused attention (Treisman & Souther, 1985), and search for a circle with a gap among closed circles, which reflects parallel processing with distributed attention. We compared performance both in a concurrent averaging task and in concurrent judgments on individual size. Whereas the individual size judgments were better when combined with the serial search task, performance on the averaging task was better when combined with the distributed attention parallel search task, even though the exposure duration was considerably shorter than it was with the focused attention task. These findings are directly relevant to Myczek and Simons's (2008) claims. There is no obvious sampling strategy that would benefit from a shorter exposure and concurrent parallel processing of the whole display. To explain the poorer performance with the longer displays in the focused attention task, Myczek and Simons suggested that serial search might interfere with sampling. But according to their account, both require focused attention to individual items. Performance on individual items was better when combined with serial search, so it seems implausible that sampling would be *harder* with the longer exposures of the focused attention task, relative to the brief popout displays and parallel processing used in the distributed attention task.

In another experiment, we combined statistical averaging with judgments of the aspect ratio of either a small foveal rectangle or a large rectangle that surrounded the display of circles to be averaged. Again, it is hard to see why attending to a large rectangle around a display should be more compatible with sampling a few items than focusing attention on a small rectangle in the center would be, given that the difficulty of the rectangle tasks was matched. The size of the attention window would have to be reset from the large rectangle to the sampled items, whereas it would not have to be with the smaller local rectangle.

In another experiment in which focused was compared with distributed attention, Chong and Treisman (2005b) either precued or postcued which subset of circles should be averaged (the red or the green) and compared accuracy for both with the accuracy when only a single set was presented. There was no significant difference in performance across the three conditions. Either memory for the whole display is perfect, or samples would need to be

taken from both sets in the postcued intermingled display, doubling the number of items to be attended, relative to the precued or the single set. The fact that there was no significant decrement in the postcued condition argues strongly against the sampling hypothesis.

Myczek and Simons (2008) suggested that the thresholds of mean size judgments were surprisingly high in these divided attention experiments, relative to the earlier experiments in which the two sides of the display were compared. However, the experiments also differed in the type of averaging task used. In the divided attention experiments, the task was a memory-based comparison of the mean size with that of a probe circle, whereas the earlier experiments used relative judgments of the two sides of the display while both were present. The divided attention tasks included a memory requirement that was not present in the simultaneous comparison task. Thus, performance cannot be directly compared across the experiments.

Another puzzle for the Myczek and Simons (2008) account is the inconsistency across different experiments in the sample sizes assumed. Performance in Ariely's (2001) experiments can be modeled with samples of one or two. Yet, in other experiments, at least four are needed. But if samples of four are possible with focused attention, why was performance not better in the Ariely experiments, where participants were averaging only 75% correct? If the capacity is available, why is it not used?

Myczek and Simons (2008) suggested that the mechanism we proposed is a new one that "has striking implications for scene perception" (p. 773) and that we should, therefore, be cautious in adopting it. We disagree that the hypothesis is so novel. As the authors say in their introduction, averaging has been shown for motion, orientation, and other variations of texture. Although the neural mechanism for coding size with outline circles has not been identified, either for single shapes or for sets of shapes, unlike the neural basis of motion or orientation, we assume that there must be one, perhaps related to the spatial frequency channels that code solid figures and textures on the basis of size. We proposed the averaging of responses in receptive fields and in voxels in fMRI as effects that are possibly analogous to the statistical processing that we have studied in psychophysical experiments, and we referred back to the pooling of responses in feature maps proposed in Treisman and Gormican (1988). The mechanism of neural averaging is unlikely to rely on individualizing each element, summing all of them together, and then dividing by N . This serial process of adding would be inconsistent with the speed of averaging and the insensitivity to exposure duration that we observed.

The existence of a separate distributed attention mode is also consistent with other findings. We know that attention to a global stimulus is possible (e.g., the global letter made of local letters in experiments by Navon, 1977, among others) and that textures can be easily discriminated on the basis of orientations, speeds or directions of motion, spatial frequencies, and other local variations. Another example suggesting differences between the local and the global attentional modes is the perception of

motion in the Ternus stimulus. When attention is spread over the display, the pairs of stimuli are grouped and are seen to move together, whereas when attention is focused on a small part of the display, one object is seen to oscillate to and fro (Hock, Park, & Schöner, 2002). Thus, separate modes of processing individual stimuli and sets of stimuli as global entities have been proposed in a variety of contexts.

There is one distinction, however, between global attention in experiments like those of Navon (1977) and the deployment of attention that we suggest is used in statistical processing, and that is the scale at which the perceptual processing is set. In Navon's experiments and others like them, participants attend to the global shape of the complete display as a unitary object; in the statistical-averaging experiments, participants attend to a property of the individual items and form a representation of the prototypical or average item—presumably, at the scale of the instances, rather than of the display as a whole. To distinguish this deployment of attention from the global one, we use the term *distributed attention*. It is applied to the set of items as a whole, but with the scale of processing set to that of the instances it contains.

New Studies

Next, we will report some new findings that we think are more consistent with our account than with the limits of focused attention. Note that it is, in fact, quite difficult to devise tests that require samples of more than four or five items, simply because samples of increasing size quickly approximate the population mean and because the performance of human observers is noisy, usually reaching levels between 75% and 80% correct. This means that converging evidence must be sought from other tasks.

In order to help readers to track the methods and the rationale we used in the new experiments, Table 1 summarizes the issues raised and the ways in which we tackled them.

**EXPERIMENT 1
Choice of Strategies for Averaging Size**

In the first experiment, participants' ability to select one of several optional sampling strategies on the fly was investigated. Myczek and Simons (2008) offered the following set of strategies as possible accounts of the various experiments published so far.

1. For Ariely's (2001) data for which a small range of sizes was used, sampling either one or two elements was sufficient. However, for the larger range of sizes, the mean of the largest and smallest is needed to simulate participants' performance.

2. For Chong and Treisman (2003) with different distributions, the strategy they suggested is to find the largest (or smallest) element and respond to the side on which that is located.

3. For Chong and Treisman (2003), Experiment 3, homogeneous/uniform condition, their strategy is to sample one or two on the mixed side and compare the mean of that sample with any circle on the homogeneous side.

4. For Chong and Treisman (2005b), the strategy is either to sample two items per side, average those, and then compare the two sides, or to find the largest or smallest circle and compare that with the most frequent circle on the side on which the frequencies are unequal.

Our impression is that implementing these strategies would require a number of separate operations in several cases. For example, to carry out Strategy 4, participants would need to determine which side is which, then find the largest or smallest circle on one side, then find the most frequent circle on the other side, and then compare the two. The feasibility of doing all this within a 200-msec exposure might be questioned.

A stronger prediction is that using these varied strategies should depend critically on the test conditions' remaining fixed throughout a block of trials, so that participants can discover which one works for a given type

**Table 1
Summary of Three New Experiments**

Purpose	Task	Displays	Independent Variables of Interest	Result
Experiment 1: Do participants perform worse when different display types, requiring different focused attention strategies, are mixed rather than blocked?	Indicate the side with the larger mean size	32 circles (16 on each side) Sizes: 0.8°–2.1° 200 msec	1. Display types: uniform, homogeneous, frequency 2. Session: mixed, blocked	There was no difference in accuracy between the two sessions, suggesting that the participants were not switching between focused attention strategies in the mixed session.
Experiment 2: Is the statistical processing of groups as accurate when participants view samples of one to two items from each group as when they view the entire display?	Indicate the side with the larger (mean) size (assuming all the items were shown)	16 circles (8 on each side) Sizes: 1°–1.9° 200 msec	1. Number of items shown on each side: one, two, or all eight items 2. Location of samples: foveal, random	There was no effect of sample location. Performance was much better with the whole display, indicating that the participants did not perform the task by sampling.
Experiment 3: Do participants determine the side with the largest mean by comparing the largest items on each side?	Indicate the side with the larger mean size	16 circles (8 on each side) Sizes: 0.7°–3.7° 500 msec	1. Side of the largest item, relative to the larger mean (<i>same, opposite</i>)	Performance was similar for <i>same</i> and <i>opposite</i> trials, suggesting that the strategy of comparing the largest items was not used.

of display. Experiment 1 tested this prediction by mixing trials that would use three different strategies randomly within the same blocks and comparing performance with blocked presentation of the same conditions. We chose the following tasks, which we had previously tested in separate blocks or experiments: Chong and Treisman (2003), Experiment 3, nonhomogeneous condition, where the strategy Myczek and Simons (2008) proposed is to find the largest circle on each side and compare them; Chong and Treisman (2003), Experiment 3, homogeneous condition, where the proposed strategy is to find the mean for the largest and smallest circles on the heterogeneous side and compare that with the size of one of the homogeneous circles; and finally, Chong and Treisman (2005b), Experiment 1 (with arrays of 16 per side), where the proposed strategy is to sample either 2 or 3 items per side.

Method

We used three different sets of displays modeled on selected conditions from our previous experiments. Sixteen circles were always present on each side of the visual field. First, in the uniform condition (uniform/uniform condition in Experiment 3 of Chong and Treisman, 2003), we used displays containing four sizes (0.96°, 1.19°, 1.49°, and 1.87°) four times each. The sets of sizes on one side of the display were either increased or decreased by 7% or 13%, as compared with those on the other side. Second, in the homogeneous/uniform condition from the same earlier experiment, one side of the display had 16 circles of 1.40°, whereas the other side again showed 16 circles of the four sizes from the uniform condition. Mean differences of 7% and 13% were also used in this condition. In the third condition (modeled on Chong and Treisman, 2005b, Experiment 1), one side of the display contained the two sizes, 1.34° and 1.63°, with different frequencies (4 with 12, 6 with 10, 10 with 6, and 12 with 4, for the small and large sizes, respectively). The other side of the display had equal frequencies of two sizes (8 with 8) and mean sizes that were 7% smaller, 7% larger, 13% smaller, or 13% larger, as compared with those on the opposite side.

Design. The participants completed two sessions. In the first session, we tested the effects of two factors within participants. The first was display type: the uniform, the homogeneous, and the frequency conditions. The second was level of difficulty: 7% or 13% mean size difference. These were randomly intermixed within both the practice and the experimental blocks. Each participant was tested in 16 practice trials and 288 experimental trials (3 display types \times 2 levels of difficulty \times 48 repetitions). In the second session, we blocked the display types to compare the results with those in the first session. The order of the display types was counterbalanced across participants.

Procedure. The procedure for Experiment 1 is shown in Figure 1. Each trial started with a fixation cross for 500 msec. Sixteen circles on each side of the visual field were presented for 200 msec. The task was to judge which side of the visual field had the larger mean size. The participants pressed “1” if the left side of the visual field had the larger mean size and “2” otherwise. A high tone was provided as feedback when the response was incorrect.

Participants. Seven Yonsei University students, including one of the authors, participated in the experiment. All had normal or corrected-to-normal vision, and all except the author were naive as to the purpose of the experiment.

Apparatus. The stimuli were displayed on a 21-in. Samsung SyncMaster monitor, which was controlled by MATLAB Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) on a PC. The participants were seated approximately 70 cm from the monitor. The distance was maintained by a head- and chinrest.

Results and Discussion

Figure 2 shows the results. For the first session, an ANOVA indicated significant effects of both type of dis-

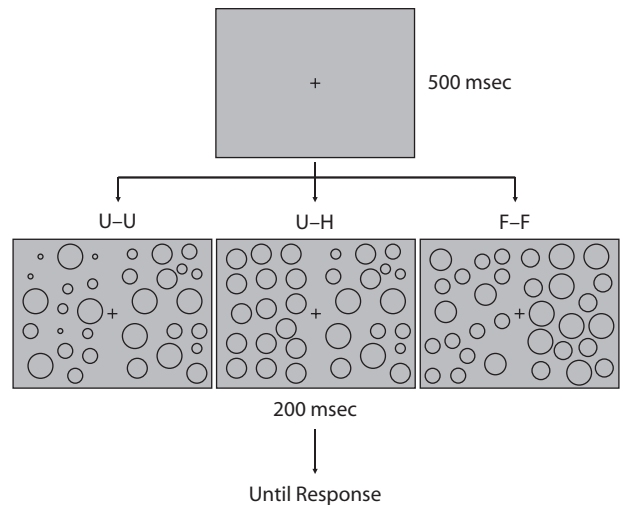


Figure 1. The procedure of Experiment 1. “U” stands for the uniform condition, “H” stands for the homogeneous condition, and “F” stands for the varied frequency condition.

play [$F(2,14) = 10.183, p < .01$] and level of difficulty [$F(1,7) = 46.738, p < .01$]. However, there was no significant interaction [$F(2,14) = 2.047, p = .166$].

The main interest was in how well the participants could perform in mixed blocks, where, according to the focused attention hypothesis, they would have to switch strategies on a trial-by-trial basis. We tested this by comparing the results of the first random session with those of the second blocked session. An ANOVA showed that performance did not differ significantly between the random and the blocked sessions [$F(1,6) = 3.272, p = .12$]. Also, paired-sample *t* tests indicated no significant differences between each pair of the conditions (all *ps* $> .05$).

Another way of testing whether the participants switched strategies on a trial-by-trial basis was to analyze separately the trials that switched conditions from the trials that repeated a condition. If the participants switched strategies on a trial-by-trial basis, they would be less accurate in the former trials than in the latter, because of switch costs. The accuracy in the switch trials (79.9%) did not differ significantly from that in the no-switch trials (82.5%) [$t(6) = 1.22, p = .27$], suggesting that the participants did not switch strategies on a trial-by-trial basis.

The level of accuracy is very similar to that obtained in the earlier experiments with blocked presentation. In the previous study of uniform and homogeneous displays, we measured the diameter difference at which the participants averaged 75%. These were 8% for the uniform and 11% for the homogeneous/uniform conditions. For the frequency condition, the averages were 71% correct for the 7% difference and 83% for the 13% difference.

There is clearly no decrement in the present study resulting from mixing the conditions randomly. Note that the blocked condition could also have benefited from greater practice, since it was run second. Again, the account in terms of distributed attention to the display as a whole seems more parsimonious and more plausible than assuming that, in the 200-msec presentation time, participants can

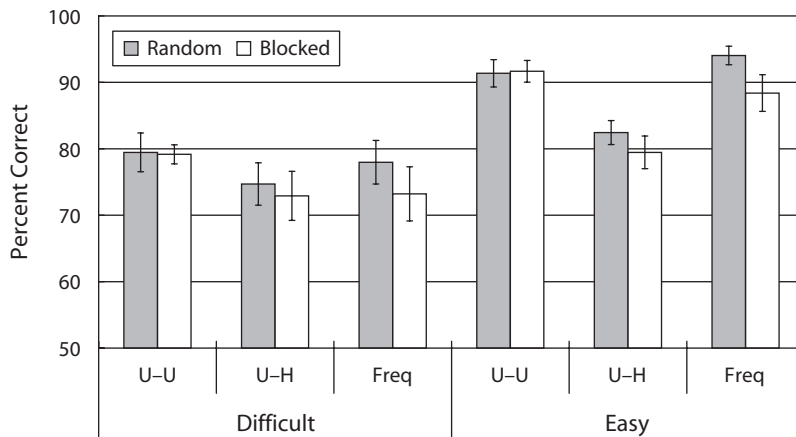


Figure 2. The results of Experiment 1. The white bar indicates the blocked conditions and the gray bar indicates the randomly mixed conditions. “U” stands for the uniform condition, “H” stands for the homogeneous condition, and “Freq” means the varied frequency condition. Error bars indicate standard errors of the means.

first determine which display type is present on each side of the display and then switch between several different focused attention strategies to match the type of display.

EXPERIMENT 2 Use of Samples to Judge Mean Size

In Experiment 2, we tested how well participants are, in fact, able to use samples of different sizes to estimate the mean size of a set of circles. Myczek and Simons (2008), in their simulation, assumed perfect use of each sample. The only limit they allowed to performance was the limited number of items in the sample. To the extent that there are other limits, their estimates of the performance possible with each sample size will be higher than that of human participants. We decided to replicate Experiment 1 from Chong and Treisman (2005b), except that we tested participants with actual physical samples of one or two items from each side of the displays, as well as with the whole displays (as in the original study).

We also tested another possibility in this experiment. The simulations selected random samples of items to test the predicted optimal performance, but human participants would be more likely to select the most visible or salient items in sampling the displays. One plausible strategy would be to select the items nearest the fovea, so we compared the performance of human participants both with a random selection of one or two items from each side of the display and with the one or two items nearest the fovea on each side.

Method

The method was taken from Experiment 1 of Chong and Treisman (2005b) and was modified to test different sample sizes. We used the sparse displays of eight from the original article. On one side of the display, we varied the frequencies of two sizes (1° and 1.6° diameters), using the following combinations (small:large): 2:6, 3:5, 5:3, and 6:2. On the other side, we varied the sizes, from 1° to 1.3° for the smaller size and from 1.6° to 1.9° for the larger, keeping the frequencies fixed at 4:4.

The display was divided into two visual fields separated by 2.6° at their nearest edges, each containing an imaginary $7 \text{ rows} \times 4 \text{ columns}$ matrix, with each cell measuring $2.6^\circ \times 2.6^\circ$. In the foveal condition, the position of the items was selected from among the cells near fixation. For samples of one and two items per side, the position was randomly selected from among column 4, rows 3, 4, and 5 in the left visual field and the mirror positions for the right visual field. For the whole display, the positions were randomly selected among the nine cells nearest fixation on each side: column 4, rows 2, 3, 4, 5, and 6, column 3, rows 3, 4, and 5, and column 2, row 4 in the left visual field and mirror cells in the right visual field. In the random condition, the positions were randomly selected in the imaginary matrix. The positions were randomly jittered within a range of 0.32° .

Design. Three factors were varied within participants. The first was the number of items randomly sampled from each side of the visual field: one, two, or all eight circles. The second was the level of difficulty: 7% or 13% mean size difference between the two sides. The third was the choice of items to present: either a random selection or the items nearest the fovea. The number of items in each sample was blocked, and the order of the blocks was counterbalanced across participants. Trials with random and trials with foveal samples were randomly intermixed within each block. Each participant was tested in 16 practice trials and 288 experimental trials (3 sample sizes \times 2 levels of difficulty \times random vs. foveal samples \times 24 repetitions).

Procedure. In each trial, a fixation cross was presented for 500 msec, followed by a 200-msec presentation of the visual display. The participants were instructed to report which side had the larger size (one item on each side) or the larger mean size. The participants pressed “1” when the left side had the larger (mean) size, and they pressed “2” when the right side had the larger (mean) size. No feedback was given.

Participants. Eight students at Yonsei University, including one of the authors, participated in the experiment. All except the author were naive as to the purpose of the experiment. All had normal or corrected-to-normal vision.

Apparatus. The apparatus was the same as that in Experiment 1.

Results and Discussion

Figure 3 shows the results of Experiment 2. The responses were scored as accurate or not, relative to the imaginary complete display, when only a smaller sample was shown. We first collapsed the data over the difficulty

level for further analysis, because the interaction (sample size \times difficulty level) was not significant [$F(2,14) = 3.729, p = .05$]. An ANOVA indicated a significant effect of the number of items sampled [$F(2,14) = 51.534, p < .01$]. However, neither the effect of foveal/random condition [$F(1,7) = 0.686, p = .435$] nor the interaction between foveal/random and number of items [$F(2,14) = 0.532, p = .599$] was significant, although for the full display, it came close [$t(7) = 2.04, p = .081$]. The important finding is that for the samples, proximity to the fovea made no difference at all. This led us to collapse the data over foveal/random condition for further analysis.

Accuracy was significantly higher in the whole-display condition (83.3%) than with samples of one and two items [62.2%, $t(7) = 8.153, p < .01$; and 66.8%, $t(7) = 9.425, p < .01$, respectively]. The difference in accuracy between samples of one and two did not quite reach significance [$t(7) = 2.132, p = .070$], showing that at the putative limit of focused attention (four items), performance remains pretty low.

The performance of our participants with the samples was also lower than that in Myczek and Simons's (2008) simulation [74.5% in the two items per side condition; $t(7) = 9.652, p < .01$]. The participants certainly focused attention as best they could on the limited number of circles presented in the sampling conditions, yet they were unable to use this information as effectively as they did the full displays or to reach the optimal results that could, in theory, be obtained from the simulations. These results suggest that sampling one or two items from each side and performing the task on the basis of those items was not equivalent to judging the mean size by attending to the whole display, even when the most visible items were sampled.

It is of some interest to measure the accuracy when scored relative to the actual presented items. These averaged 86.8%, 83.3%, and 83.3% correct for displays of one, two, and eight, respectively, which were not significantly different (all $ps > .05$). Thus, accuracy was no lower for larger displays (as shown also by Ariely, 2001).

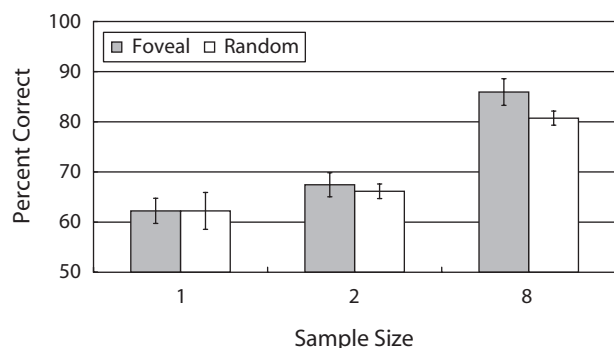


Figure 3. The results of Experiment 2. The sample size indicates the number of circles randomly selected and displayed out of eight circles on each side of the visual field. The white bar stands for the random selection and the gray bar stands for the foveal selection. Error bars indicate standard errors of the means.

EXPERIMENT 3 Comparing the Largest Sizes

Another way to test the focused attention account is to ask, for each strategy, whether it is, in fact, working in the data that participants provide. In the next experiment, we did this for the simplest strategy that could be applied to experiments that compare displays on two sides of the visual field—that is, to compare the size of the largest item on each side, rather than computing the mean. (This is one of the strategies proposed by Myczek & Simons, 2008, in their article.) In many cases, this strategy could work, but do participants in fact use it when asked to compare the average sizes?

Method

The stimuli were 16 circles, 8 on the left and 8 on the right side of the screen. There were two instances of each of four sizes on each side of the display. On each side, stimuli were placed within a 2×4 matrix, with each cell subtending 4° . The circles appeared centered within their cells, and there was a 4° blank area between the left and the right matrices. All the stimuli, as well as the fixation cross (0.2°), were drawn in white on a gray background. On one side, randomly chosen on each trial, the four sizes were randomly generated within a prespecified range (0.7° – 3.4° diameter). On the opposite side of the screen, the mean diameter was either 10% smaller or larger. The largest item was also 10% smaller or larger than the largest item on the opposite side. Importantly, mean size and largest item size were selected orthogonally, so that for half the trials, the largest item was on the side of the largest mean (*same* trials), whereas for the other half, the two were on opposite sides (*opposite* trials). The remaining three sizes were generated randomly, with the constraint that the smallest item be on the side of the smallest mean on half of the *same* and half of the *opposite* trials. Within each side, the sizes were at least 0.3° (10 pixels) apart.

Design. Each participant completed 16 practice trials, followed by 640 experimental trials. The variable of interest was whether the largest item in the display was on the same side as the largest mean (*same* condition) or on the opposite side (*opposite* condition). There was an equal number of *same* and *opposite* trials, and they appeared in random order. During the practice block, the participants received feedback on each trial (they heard a short beep when they chose the wrong side or failed to respond on time). In the experiment, the participants received feedback only at the end of each block (32 trials), by seeing their accuracy for that block displayed on the screen.

Procedure. On each trial, the display of circles appeared for 500 msec. The participants were asked to select the side of the screen (left or right) with the largest mean size. If the participants responded within 3,000 msec from the onset of the display, the trial was recorded; otherwise, the experiment moved on to the next trial. The interval between the end of a trial and the beginning of the next was 1,000 msec. The response keys for the left and the right sides of the screen were the numbers “1” and “2” on the keyboard number pad.

Participants. Eight Princeton University undergraduate students participated in this experiment for course credit. All the participants had normal or corrected-to-normal vision.

Apparatus. All the stimuli were generated using MATLAB and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and were presented on a 17-in. Mitsubishi DiamondPlus 74SB monitor, using a Macintosh G3. The participants were seated approximately 60 cm away from the monitor.

Results

The results of Experiment 3 are shown in Figure 4. The participants were, if anything, more accurate in the *op-*

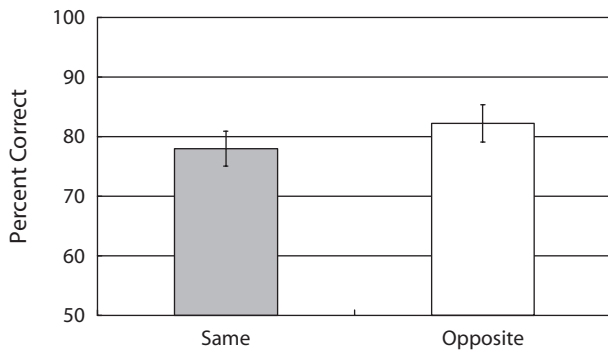


Figure 4. Average accuracy for Experiment 3 when the largest size was on the same side as the larger mean and when they were on opposite sides. Error bars indicate standard errors of the means.

posite than in the same condition (78% mean accuracy for same and 82% for opposite trials), although the difference was not significant in a repeated measures ANOVA [$F(1,7) = 2.06, p = .19$]. We also compared the first 30 trials for each condition in order to make sure that the participants had not learned to respond accurately to the opposite condition on the basis of the feedback given at the end of each block. Accuracy for the first 30 same trials was higher than that for the first 30 opposite ones (82% for same, 76% for opposite), but the difference again was not significant [$F(1,7) = 1.48, p = .26$]. It is clear that the participants were not using the strategy of responding on the basis of the largest size on each side. This factor had no significant effect on their performance.

CONCLUSIONS

We have shown a number of ways in which human performance seems poorly adapted to carry out the strategies proposed by Myczek and Simons (2008). One problem with the account based on a variety of different focused attention strategies is to specify how participants select the correct strategy for each task. In the previously published articles, conditions were typically blocked, allowing an opportunity for the participants to discover the optimal method to solve the task. However, in Experiment 1, the participants did just as well when we presented a random mixture of conditions requiring different strategies as when the different types of display were blocked. If they had been using focused attention, the mixed blocks would have required them to switch repeatedly between different strategies. There is no evidence here that they were using small samples in a variety of different ways, rather than simply doing as they were asked and statistically averaging the set as a whole.

Second, when we directly tested the sampling strategy by giving the participants samples of two or four circles to average, their responses gave significantly worse estimates of the invisible full display than would follow from the optimal performance implemented in the simulations. One possible reason why performance might have been worse for the participants when they were given a random

sample than it was for the simulations is that the sample might have included items that happened to be difficult to see. We ruled out this possibility by testing the participants with samples from the area closest to the fovea, where acuity would have been highest and where attention was most likely to be directed on each trial. Performance was no better with these samples than with the random ones.

Finally, when we checked the simple strategy of comparing the largest items on each side, instead of the means, we found that on trials in which the two strategies gave opposite answers, our participants did just as well as on trials in which the two strategies gave the same answer and far better than would be predicted by the *compare the largest* strategy.

Myczek and Simons (2008) raised an important question. They clearly demonstrated that alternative strategies are available that could mimic the performance of human participants in statistical-averaging tasks without going beyond the generally accepted limited focus of attention. What we have tried to do, in turn, is to raise some doubts about whether those strategies are actually used, given a number of limits that we have observed to human performance. The three experiments reported here gave results more consistent with the idea that statistical averaging is a natural perceptual function that can be invoked when sets of similar items are present in a scene. The means are not usually error free; they are approximations, rather than an accurate sum of the sizes divided by the number of items in the display. But they seem to reflect a perceptual operation on the set as a whole, rather than on a small subset of the items. We suggest that this statistical processing mode provides a partial solution to the question of how we can rapidly assess the general properties and layout of a natural scene and extract the gist, despite the stringent limits of focused attention that are needed for binding features, localizing changes, and individuating and identifying specific objects.

AUTHOR NOTE

This work was supported by Korea Research Foundation Grant KRF-2007-101039003-2007-8-0746, funded by the Korean Government (MOEHRD), by NIH Grant 2R01 MH 058383-04A1, and by Israeli Binational Science Foundation Grant 2003148. We are grateful to Shlomo Sher for his helpful comments and suggestions. Correspondence concerning this article should be addressed to S. C. Chong, Department of Psychology, Yonsei University, 262 Seongsanno, Seodaemun-gu, Seoul 120-749, Korea (e-mail: scchong@yonsei.ac.kr).

REFERENCES

- ARIELY, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, **12**, 157-162.
- BRAINARD, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433-436.
- CHONG, S. C., & TREISMAN, A. (2003). Representation of statistical properties. *Vision Research*, **43**, 393-404.
- CHONG, S. C., & TREISMAN, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, **67**, 1-13.
- CHONG, S. C., & TREISMAN, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, **45**, 891-900.
- HOCK, H. S., PARK, C. L., & SCHÖNER, G. (2002). Self-organized pattern

- formation: Experimental dissection of motion detection and motion integration by variation of attentional spread. *Vision Research*, **42**, 991-1003.
- LI, F. F., VANRULLEN, R., KOCH, C., & PERONA, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, **99**, 9596-9601.
- MYCZEK, K., & SIMONS, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, **70**, 772-778.
- NAVON, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, **9**, 353-383.
- OLIVA, A., & TORRALBA, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, **155**, 23-36.
- PELLI, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437-442.
- POTTER, M. C., & FAULCONER, B. A. (1975). Time to understand pictures and words. *Nature*, **253**, 437-438.
- RENSINK, R. A., O'REGAN, J. K., & CLARK, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, **8**, 368-373.
- THORPE, S., FIZE, D., & MARLOT, C. (1996). Speed of processing in the human visual system. *Nature*, **381**, 520-522.
- TREISMAN, A. [M.] (2006). How the deployment of attention determines what we see. *Visual Cognition*, **14**, 411-443.
- TREISMAN, A. M., & GELADE, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, **12**, 97-136.
- TREISMAN, A. [M.], & GORMICAN, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, **95**, 15-48.
- TREISMAN, A. [M.], & SOUTHER, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, **114**, 285-310.
- WOODWORTH, R. S. (1938). *Experimental psychology*. New York: Holt.

(Manuscript received August 14, 2007;
revision accepted for publication April 15, 2008.)