Available online at www.sciencedirect.com
SCIENCE DIRECT®
Vision Research 45 (2005) 891–900
Vision Research
www.elsevier.com/locate/visres

Statistical processing: computing the average size in perceptual groups

Sang Chul Chong [a,*], Anne Treisman [b]

[a] Department of Psychology, Vanderbilt University, 301 Wilson Hall, 111 21st Avenue South, Nashville, TN 37203, United States
[b] Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544-1010, United States

Received 1 January 2004; received in revised form 23 September 2004

Abstract

This paper explores some structural constraints on computing the mean sizes of sets of elements. Neither number nor density had much effect on judgments of mean size. Intermingled sets of circles segregated only by color gave mean discrimination thresholds for size that were as accurate as sets segregated by location. They were about the same when the relevant color was cued, when it was not cued, and when no distractor set was present. The results suggest that means are computed automatically and in parallel after an initial preattentive segregation by color.

Keywords: Size; Perceptual groups; Mean; Automatic processing

The trees in a forest, the grass in a field, a flock of birds, the cars in a parking lot, are seen as groups of similar but not identical objects about which we may not need to store individuating information. For most purposes a description of their general statistical properties, such as the mean value, the range, the variance and the distribution on a number of dimensions, will meet our everyday needs. Ariely (2001) proposed that the visual system represents overall statistical properties when sets of similar objects are present. He showed that the mean size is perceived more accurately than the individual sizes in a display of disks of varied sizes, and that there is little effect of the number of disks.

Our hypothesis is that statistical descriptors are computed automatically when attention is distributed over the display and the scale is set to that of individual elements (Chong & Treisman, 2003). We showed that judgments of the mean size of a set of circles are almost as accurate as judgments of the size of a single circle presented alone, and that they are little affected either by exposure duration or by delay, suggesting an automatic and parallel process. We confirmed that the judgments involved computing the mean size of an array by showing that comparisons were almost as accurate when the distributions differed as when they were the same, using sets drawn from normal distributions, rectangular distributions, distributions with just two equal peaks, or homogeneous distributions. More recently we have tested the automaticity of this averaging process using another criterion,—the absence of interference from a concurrent task. Judgments of mean size could be combined without decrement with tasks requiring either distributed attention (search for an open circle among closed circles) or global attention (discriminating the orientation of a large rectangular frame around the display). On the other hand, tasks requiring either focused attention to individual circles in the relevant set (search for a closed circle among open circles) or

* Corresponding author.
E-mail address: sangchul.chong@vanderbilt.edu (S.C. Chong).

0042-6989/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2004.10.004

focused attention to an irrelevant stimulus (discriminating the orientation of a small foveal rectangle) did interfere with judgments of mean size (Chong & Treisman, in press).

In these experiments, we controlled the density and the number of elements, and we restricted the display to just the relevant elements. In natural scenes, the elements may vary not only in size but also in other quantitative attributes. In computing the mean size, can we ignore other parameters like the density and the numerosity of the elements, or are these different quantities pooled in some aggregate description of quantity? The effects of density and numerosity on statistical processing have been studied in other domains of visual perception. Dakin (1997) explored the effect of density in computing the average orientation of Glass patterns with a dipole separation of 8′. Discrimination whether the average orientation was clockwise or anticlockwise relative to the vertical was poor for very sparse patterns (8 dipoles/deg^2), but rapidly improved with an increasing number of dipoles, showing little effect of density above about 64 dipoles/deg^2. Allik, Tuulmets, and Vos (1991) investigated possible effects of size on visual number discrimination using two random dot-patterns. Participants compared a reference pattern that was always composed of 32 randomly distributed dots to a test pattern with one of five magnifications and with a slightly smaller or larger number of dots. They found that participants could accurately judge the number of items irrespective of the size of the stimulus pattern, suggesting size invariance in number discrimination. In our first experiment we explored the effects of density and of number on judgments of the mean size of sets of circles, to see whether participants could abstract the average size from other measures of quantity like the ratio of filled to unfilled area or the numerosity of the displays.

Natural scenes usually contain many disparate sets of elements. It might be meaningful to compare the sizes of pebbles in a dense pile with those scattered more sparsely around the area, but it would hardly be useful to average the sizes of the pebbles with the sizes of the grains of sand in an adjoining area, or with the fallen leaves scattered amongst the pebbles. In summarizing the sizes, we must pre-sort and select the items that should and that should not be pooled. By attending to a defined area, we may be able to generate statistical descriptors specifically for the elements it contains. Indeed this was the task we used in our earlier experiments (Chong & Treisman, 2003). Participants had no difficulty comparing mean sizes across the left and right visual fields. But what if the sets are composed of two types of elements that are spatially intermingled? Perceptual grouping based on differences in orientation and shape can occur even with randomly mixed sets (Beck, 1966). Does the computation of mean size follow perceptual segregation of the scene into separate groups, or are all the items in a given area pooled together? Can we selectively average a subset of randomly mixed elements defined by particular features such as color, shape, orientation or motion? Does this happen automatically and in parallel for all the different perceptual subsets in a scene, or must we choose in advance? In Experiments 2 and 3, we explore the perceptual structuring that constrains the averaging process and makes it useful to us in the real world.

1. Experiment 1

In the first experiment, we tested the effects of number and of density on discriminations of the mean size of circles in two spatially segregated arrays. One possibility is that the visual system forms a general representation of the total stimulation coming from a given area. This will be perfectly correlated with the mean size if the number and density are held constant (as in our earlier experiments), but such a correlation is seldom present in the real world. It is important to find out to what extent we are capable of separating out these various descriptors when they vary either independently or in partially correlated fashion. We presented displays of 8 circles in either a dense array (0.139 circle/deg^2) or a sparse array (0.075 circle/deg^2) and displays of 16 circles in a dense array (0.149 circle/deg^2). Participants compared the mean sizes of elements in two arrays (presented in the right and the left visual field) that were either matched in number and density or mismatched. To ensure that they were computing the mean size rather than, for example, the largest size or the mode, one array varied the number of instances of two fixed sizes and the other varied the sizes of two sets with equal numbers of instances.

2. Methods

2.1. Participants

Seven participants including the first author participated in the experiment. All were members of Princeton University. All had normal or corrected-to-normal vision.

2.2. Apparatus and stimuli

The stimuli were created with the Psychophysics Toolbox (Brainard, 1997) and presented on the screen of an Apple 17″ Monitor. The monitor was driven by a Macintosh G3, which also performed all timing functions and controlled the course of the experiment. Participants were seated approximately 76 cm from the screen, at which distance a pixel was approximately

0.02° of visual angle, and they viewed the screen with both eyes. The stimuli were white outline circles. Each display was divided into two halves vertically, each containing either 8 or 16 circles in a mixture of two sizes. Examples are shown in Fig. 1. The possible sizes were equally spaced on a power function with an exponent of 0.76 (the psychological scale for size, Teghtsoonian, 1965). The diameters ranged from 1° to 1.9° and the means of the diameters in each subset ranged from 1.4° to 1.6°. The luminance of the stimuli was 49.9 cd/m^2 and the luminance of the gray background was 26.8 cd/m^2.

We used two different ways of varying the mean sizes. The first was to vary the frequencies of two sizes (1° and 1.6°), holding the range constant at 0.6°. The two sizes could appear in the following frequencies: 2 with 6, 3 with 5, 5 with 3 or 6 with 2, making displays of 8 items, and 4 with 12, 6 with 10, 10 with 6, and 12 with 4, making displays of 16 items. The second was to vary the two sizes, but present equal numbers of each. Holding the range constant (0.6°) and the frequencies equal (4 with 4 for the set size of 8, or 8 with 8 for the set size of 16), we varied the smaller size from 1° to 1.3° and the larger size from 1.6° to 1.9°, to generate mean sizes that were either 7% smaller or 7% larger or 13% smaller or 13% larger than the mean sizes obtained by the first method. To generate a display, we randomly chose one of these two methods and assigned it to one of the two sides, then used the other method for the other side of the display.

Density was varied as follows: Each visual field was divided into an imaginary 4×7 matrix where each cell measured $2.6° \times 2.6°$. The left and right displays were separated by 2.6° between their near edges. The locations of the circles within the displays of 16 and the sparse displays of 8 were randomly selected in the matrix and they were randomly jittered within a range of 0.32° in each cell of the matrix. For the dense displays of 8 circles, we used an imaginary 3×5 matrix that was randomly positioned within the 4×7 matrix, keeping the cell sizes and the jitter the same. Thus the density was approximately matched for the 8 dense and the 16 element displays. There was no sparse condition for the set size of 16.

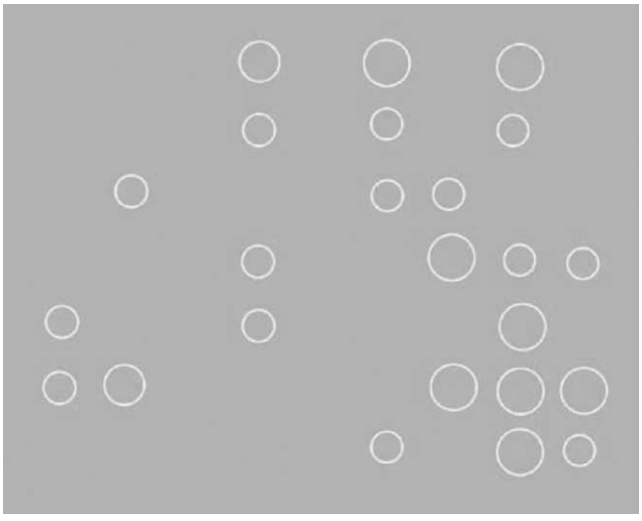

Fig. 1. The stimuli for Experiment 1. The left side has 6 large circles and 2 small circles and they are sparsely presented. The right side has 8 large and 8 small circles and they are densely presented.

2.3. Design

There were two independent variables, which were both varied within participants. The first variable was the mean size difference—large (13% diameter difference between the means of the two visual fields) or small (7% diameter difference); the second variable was the type of size comparison—either 8 sparse with 8 sparse, 8 dense with 8 dense, 16 dense with 16 dense, 8 sparse with 8 dense, 8 dense with 16 dense, and 8 sparse with 16 dense. All these conditions were randomly mixed within blocks. There were 48 trials in the practice block, 384 trials (2 mean size differences $\times$ 6 types of size comparison $\times$ 32 repetitions) in the experimental block. The order of trials within each block was randomly selected, under the constraint that each condition (2 mean size differences $\times$ 6 types of size comparison) was presented once before any condition was repeated.

2.4. Procedure

Each trial started with a fixation cross for 500 ms. The displays of circles were then presented for 200 ms. Participants' task was to decide which visual field had the larger mean size. When they thought that the left visual field had the larger mean size, they pressed '1'. When they thought that the right visual field had the larger mean size, they pressed '2'. If their decision was incorrect, they heard a short high-pitched tone.

3. Results and discussion

The results of Experiment 1 are shown in Fig. 2. We first compared large and small mean size difference conditions. The percent correct in the large condition (83%) was higher than the percent correct in the small condition (71%; $F_{(1,6)} = 93.8$, $p < .01$). However, the interaction between the mean size difference and the types of size comparison was not significant ($F_{(5,30)} = .595$, $p = .7$). Consequently, we merged the large and the small mean size difference conditions for further analysis.

We compared trials in which the two visual fields had the same type of display (matched, e.g. 8 dense and 8 dense) and trials in which the two visual fields had different types of displays (non-matched, e.g. 8 dense and 16 dense). A t-test showed that performance was better for

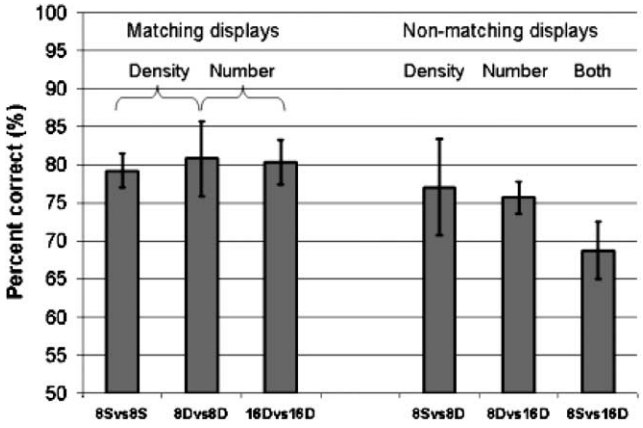

Fig. 2. The results of Experiment 1. 8S vs 8S stands for the 8 sparse with 8 sparse condition, 8D vs 8D stands for the 8 dense with 8 dense condition, 16D vs 16D stands for the 16 dense with 16 dense condition, 8S vs 8D stands for the 8 sparse with 8 dense condition, 8D vs 16D stands for the 8 dense with 16 dense condition, and 8S vs 16D stands for 8 sparse with 16 dense condition. The y axis starts at 50% correct because this was the chance level. The error bars indicate the confidence interval.

the matched displays (80%) than for the non-matched displays (74%), $t_{(6)} = 4.353$, $p < .01$. An ANOVA on the matched displays showed no significant differences due either to numerosity or to density ($F_{(2,12)} = .264$, $p = .77$). An ANOVA on the non-matched displays showed a significant overall effect of conditions ($F_{(2,12)} = 4.806$, $p < .05$). Subsequent pair-wise t-tests showed that the 16 dense with 8 sparse condition gave significantly lower accuracy than either the 8 sparse with 8 dense condition ($t_{(6)} = 2.68$, $p < .05$) or the 8 dense with 16 dense condition ($t_{(6)} = 2.965$, $p < .05$). However, the 8 sparse with 8 dense condition did not differ significantly from either the 8 sparse with 8 sparse condition ($t_{(6)} = .927$, $p = .39$) or the 8 dense with 8 dense condition ($t_{(6)} = 1.13$, $p = .30$). Thus there is no effect of density differences either within a display or across displays. It is only when differences in density and in number are combined, that performance begins to be slightly impaired.

The equal accuracy we observed for displays of 8 and 16 items confirms the earlier finding by Ariely (2001) that averaging is unaffected by display size. These results are also consistent with the very small effects of density on mean orientation discrimination (Dakin, 1997) and with the size invariance found in number discrimination (Allik et al., 1991). Statistical processing seems to be robust against variations in density and numerosity.

The fact that there was little difference in accuracy in comparing displays that were matched or non-matched in either number or density is a critical observation for the claim that participants were indeed averaging sizes. Simply summing the areas covered by elements in the displays to be compared would not help in estimating the mean when the displays differ in the number of elements they contain. Simply summing within equal sample areas would not help either when density differs. We also ruled out a direct comparison of individual element sizes by mixing frequencies and sizes in determining the means. The fact that all our displays were composed of only two sizes makes it very unlikely that participants compared the mode rather than the mean. The displays with equal frequencies actually had two modes and no other elements. Choosing the larger of those would give the wrong answer on half of the trials. The result confirms that at least in these conditions the displays are statistically analyzed and compared.

If there were substantial internal noise in encoding the individual sizes, it might mask any effects of our density variations. The noise would have to be very large to mask an effect of doubling the density and we have some evidence suggesting that this is unlikely to be the case. Chong and Treisman (2003) found that the threshold for judging the size of an individual circle was the same as the threshold for judging the mean size of 12 circles, suggesting that internal noise contributes little to the averaging process. Ariely (2001) also found little or no effect of the number of elements, from 4 to 8 to 16, on the accuracy of judging the mean sizes.

4. Experiment 2

In Experiment 2, we test whether participants can select which subset of items to average together, or whether all items are automatically pooled to form a single mean. Chong and Treisman (2003) found that thresholds for discriminating the mean size of heterogeneous sets in two spatially segregated arrays were as accurate as thresholds for discriminating the size of elements in homogeneous arrays or the size of two single elements. The purpose of Experiment 2 was to see whether the same would be true when the elements were spatially intermingled and the two sets were defined only by a color difference.

In addition, we used a larger number of size differences to allow a test of the idea that participants compute the median size rather than the mean. Including some small differences allowed us to compare trials on which the median gave an answer that was inconsistent with the mean and trials on which they gave consistent answers.

5. Methods

5.1. Participants

Five participants including the first author participated in the experiment. All were members of Princeton University. All had normal or corrected-to-normal vision.

5.2. Apparatus and stimuli

The stimuli and the apparatus were the same as in Experiment 1 except for the following changes. Each display consisted of circles in two different colors (blue and green). There were either 1 or 12 circles in each color. When there were 12, they were either homogeneous in size within the color sets, or a mixture of two different sizes. For the heterogeneous sets, we used two different ways of varying the mean sizes, as we did in Experiment 1. (1) Holding the difference in size constant at 0.5°, we varied the numbers of circles of each of the two sizes. Each of the two sizes could appear from 2 to 10 times, giving mean sizes from 1.1° to 1.5°. (2) Holding the range (0.5°) and the numbers (6 vs. 6) constant, we varied the smaller size from 0.8° to 1.2° and the larger size from 1.3° to 1.7°, to generate mean sizes that matched each of the mean sizes obtained by method (1). In generating a display, we randomly assigned one of the two colors to one of the two methods, then used the other method for the set in the other color. The difference between the mean sizes of the two different colored sets in any given display could take on any of nine different values—0%, 3.0%, 6.1%, 9.5%, 13.0%, 16.6%, 20.5%, 24.7%, and 29.1% diameter difference.

The displays were divided into an imaginary 8×7 matrix where each cell measured 2.6° × 2.6°. In all three conditions, (homogeneous, heterogeneous and single circle) the locations of the circles within the displays were randomly selected in the matrix and they were randomly jittered within a range of 0.32° in each cell of the matrix. The luminance of the stimuli for both colors was 18.1 cd/m^2 and the luminance of the gray background was 26.8 cd/m^2. An example of the heterogeneous displays is shown in Fig. 3.

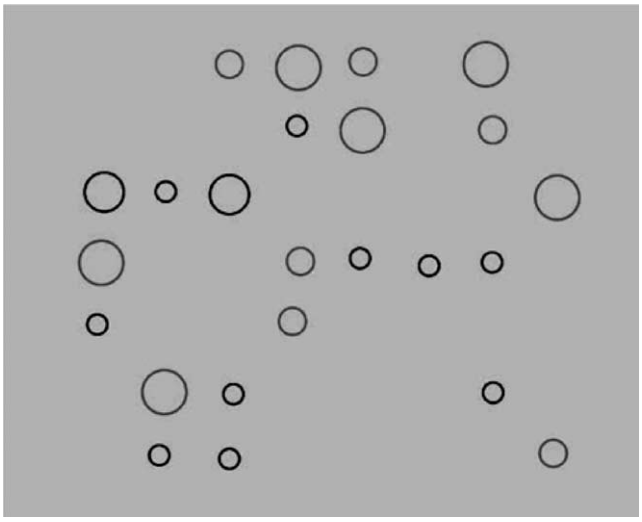

Fig. 3. The stimuli for Experiment 2. The black line indicates blue and the gray line indicates green.

5.3. Design

The task was to say which colored circles had the larger size or the larger mean size. The one independent variable in the experiment, which was varied within participants, was the type of display—heterogeneous sizes, homogeneous sizes, and single circle. Each participant served in two sessions containing three practice blocks followed by three experimental blocks (one for each type of size comparison). Each experimental block was preceded by a practice block of the same type. The order of blocks was counterbalanced within and across participants. There were 36 trials in the practice blocks, and 210 trials in each experimental block (9 comparison stimuli × on average 23 repetitions, ranging from 18 to 32). The order of trials within each block was randomly selected.

Thresholds were measured using 2AFC, in which participants decided on each trial which colored circles had the larger size or the larger mean size. Probit analysis (Finney, 1971) was used to determine the thresholds. This procedure plots the proportion of correct judgments against each difference between the means of the two different colored circles. The threshold was defined as the percent diameter difference between the means that gave 75% accuracy in this graph.

5.4. Procedure

Each trial started with a fixation cross for 500 ms followed by a display, presented for 200 ms. Each display consisted of 12 blue and 12 green circles, or a single circle in each color. The 12 circles in a given color were either in 2 different sizes (heterogeneous), or all the same size (homogeneous). Participants' task was to decide either which colored circles had the larger mean size or which colored circles had the larger size. When they thought that the blue circles had the larger mean size or the larger size, they pressed '1'. When they thought that the green circles had the larger mean size or the larger size, they pressed '2'. If their decision was incorrect, they heard a short high-pitched tone.

6. Results and discussion

The results of Experiment 2 are shown in Fig. 4. The thresholds were quite low for all three types of size judgment. A diameter difference of only 8%–12% was required for 75% accuracy in mean judgments. An ANOVA indicated significant effects of discrimination type ($F_{(2,8)} = 13.998$, $p < .01$). t-tests showed that the heterogeneous condition gave a significantly higher threshold than either the homogeneous condition ($t_{(4)} = 4.059$, $p < .05$), or the single item condition ($t_{(4)} = 3.313$, $p < .05$). The single item condition gave a

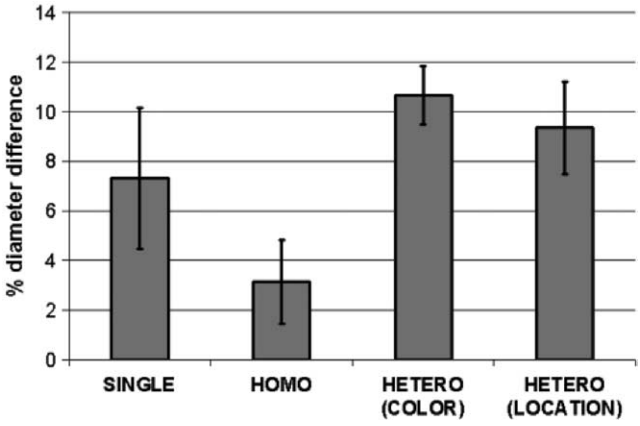

Fig. 4. The results of Experiment 2. The x axis indicates size judgment categories. The y axis indicates the thresholds defined as the percent diameter difference between the two sets on any given trial. HETERO stands for heterogeneous condition, HOMO stands for homogeneous condition, and SINGLE stands for single size condition. The error bars indicate the confidence interval. Thresholds for judging the mean size based on location came from Chong and Treisman (2003).

significantly higher threshold than the homogeneous condition ($t_{(4)} = 3.268$, $p < .05$). The thresholds were quite consistent. Standard errors of the mean across the five participants for heterogeneous, homogeneous, and single circle conditions were 1.5, 0.9, and 0.6 respectively.

We compared the thresholds for judging the mean size in the present experiment with those from Experiment 3 in Chong and Treisman (2003), to see how selection based on color compared with selection based on location. The methods used in the two experiments were identical except for small differences in the sizes tested and except for the fact that the present experiment used two different ways of varying the mean while the earlier experiment used only one (varied sizes rather than frequencies). Surprisingly, the thresholds for these intermingled color sets (10%) were about the same (9%) as those for sets segregated by location from Experiment 3 in Chong and Treisman (2003), despite the fact that the means for the color sets were varied in two different ways (frequencies for one set and sizes for the other), eliminating some indirect cues to the mean that might have been used in the earlier experiment. Performance on the single items and homogeneous items was better in the present experiment than in Chong and Treisman (2003). This could be due to the smaller spatial separation between pairs of circles to be compared in the present relative to the earlier experiment (10° compared to 16° for the single circles, and 2.6° compared to 4.8° for the nearest pair with the homogeneous circles).

The threshold for the homogenous condition in the present experiment was lower than that for the heterogeneous condition. One reason why the homogeneous displays might be easier than the heterogeneous in the present experiment (and much less so in the earlier experiment by Chong & Treisman, 2003) is that there was no need to locate and select more than one of the homogeneous circles to do the task whereas all the heterogeneous circles had to be considered. In the earlier experiment, selection (of one side of the display) was as easy for the heterogeneous as for the homogeneous circles. The single circle was also probably harder to locate in the present experiment where it was presented in random locations, whereas it was always in the center of the field in the earlier experiment.

Could the judgments have been based on the median rather than the mean size? To explore this possibility, we took all the trials on which the mean and the median would have given a different answer to the question 'which set has the larger average size?' These amounted to 11% of the total trials for each participant. Because these inconsistent trials all had small differences between the means, ranging from 3%, to 9.5%, we equated the difficulty of the consistent trials to which we compared performance, selecting only those 29% of the total trials on which the mean differences were also between 3% and 9.5%. In these selected subsets of trials, the proportion of easier trials (with the larger mean differences) was larger for the consistent trials. Yet accuracy did not differ significantly. In fact it was, if anything, higher for the inconsistent trials (68% compared to 64% in the consistent trials; $t_{(4)} = 0.841$, $p = .45$). Accuracy was significantly above the chance level both in the consistent trials ($t_{(4)} = 3.81$, $p < .05$) and in the inconsistent trials ($t_{(4)} = 7.84$, $p < .05$). Clearly, participants were not relying on the median values instead of the means.

7. Experiment 3

In the final experiment, we explore the degree of automaticity with which statistical descriptions of subsets of elements in a scene can be formed. How efficiently can participants select one subset to average? Are these descriptors computed for more than one group of elements at a time? Using a cueing paradigm, we compared thresholds for computing the mean size of a cued subset presented intermixed with other elements and the threshold for the same set presented alone. We also compared thresholds when the relevant subset was cued before the mixed display was presented and when it was not cued until the mixed display was presented.

8. Methods

8.1. Participants

The same five participants as in Experiment 2 were tested in this experiment.

8.2. Apparatus and stimuli

The stimuli and the apparatus were the same as in Experiment 2 except for the following changes: The colors of the circles were changed to green and red and their luminance to 16 cd/m^2 on the same background luminance of 26.8 cd/m^2 as in the previous experiment. Only the heterogeneous sets were tested in this experiment. The task was changed from a simultaneous discrimination of the mean sizes of the two subsets to a subsequent forced choice judgment of the mean of one subset. Two intermingled sets of 12 circles were presented for 200 ms, followed by two test circles, presented 3 degrees to the right and left of fixation. They overlapped the previous locations of randomly placed display circles on only 6% of trials, minimizing the risk of successive masking. Participants were asked to decide which of the two test circles matched the mean size of the circles in the designated-color. The relevant color was either cued before the presentation of the 24 colored circles, or cued after the presentation. In a single color condition, we presented only the relevant subset. In the single color displays, the 12 circles were randomly located within an imaginary 6×5 matrix which itself was randomly located within the 8×7 matrix to ensure a similar overall density of circles across all three conditions. This resulted in a greater density of *relevant* circles in the single color condition than in the heterogeneous conditions. However, Experiment 1 showed that density makes little difference to judgments of mean size.

8.3. Design

The task was to say which of the two test circles matched the mean size of the circles in the designated color. The one independent variable in the experiment was the type of cue-cued, non-cued, or single color. This was varied within participants.

Each participant served in two sessions, each containing three practice blocks followed by three experimental blocks (3 types of cue). Each experimental block was preceded by a practice block of the same type. The cue type (cued, non-cued, or single color) was blocked and the order of conditions was counterbalanced within and across participants. There were 36 trials in the practice blocks, 210 trials in each experimental block (9 comparison stimuli $\times$ on average 23 repetitions; comparison stimuli that were close to the thresholds were repeated more often). The order of trials within each block was randomly selected.

Thresholds were estimated using the same method as in Experiment 2 except that the differences between the mean sizes of the two sets of different colored circles were varied across participants. Participants were pretested to find the best range to test their individual thresholds. Two participants had differences ranging from 0% to 40% in steps of 5%, another two had differences ranging from 0% to 48% in steps of 6%, and the last participant had differences varying from 0% to 24% in steps of 3%. The numbers of trials per step varied less than in Experiment 2; there were 46 each for the first and the last three and 48 for the other three steps.

8.4. Procedure

The displays were preceded by a fixation cross presented for 500 ms. The circle displays then appeared for 200 ms. The cued and non-cued displays consisted of 12 circles in two different sizes for each color. The single color displays contained only the twelve circles in the relevant color. The displays were immediately followed by two test circles in the relevant color for that trial. Participants' task was to decide which of the two test circles matched the mean size of the circles in the relevant color. When they thought that the left test circle was the mean size, they pressed '1'. When they thought that the right test circle was the mean size, they pressed '2'. If their decision was incorrect, they heard a short high-pitched tone.

The cues signaling which was the color of the relevant subset were as follows: In the cued trials, the fixation cross was preceded by two parallel lines in the color designated for that trial, presented just above and below the fixation cross for 500 ms. The lines disappeared at the same time as the fixation cross. In the non-cued trials, the color of the test circles indicated which had been the relevant set of circles for the mean size judgment.

9. Results and discussion

The results of Experiment 3 are shown in Fig. 5. The mean size thresholds did not differ significantly across conditions ($F_{(2,8)} = 1.796$, $p = .23$). t-tests showed no significant differences among any possible pair-wise comparisons. Standard errors of the mean for cued, non-cued and single color conditions were 2.7, 2.6, and 2 respectively.

The fact that there was no difference between the cued and the single color condition suggests that participants could efficiently select the relevant subset of the display. The fact that there was no difference between the cued and the non-cued conditions indicates that they could register the mean sizes of both subsets as accurately as the mean of a single subset with displays that were present only for 200 ms., suggesting parallel, or very rapid serial, extraction of the mean.

Before drawing these strong conclusions, however, we wanted to make sure that participants really selected a subset and calculated its mean rather than pooling across the colors and giving the mean of the whole display. We conducted a number of tests to rule out the

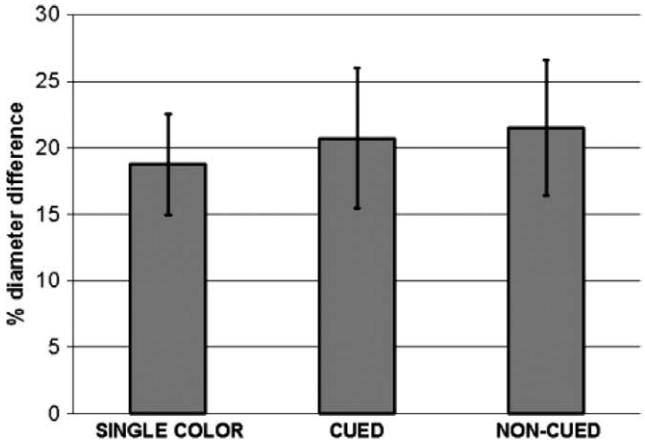

Fig. 5. The results of Experiment 3. SINGLE COLOR stands for mean discrimination without a distractor set, CUED stands for mean discrimination with a cue, and NON-CUED stands for mean discrimination without a cue. The error bars indicate the confidence interval.

latter strategy. First, we sorted the data by the differences between the means of the two sets of colored circles on any given trial. If participants were able perfectly to select the relevant subset and ignore the other, there should be no effect of the difference between their means. We regressed the percent correct on these mean differences and found that none of the five participants produced a significant positive slope in either the pre-cued or the post-cued condition. (One actually produced a significant negative slope in the post-cued condition.) The results imply that the distractor subset had little or no effect on accuracy.

Secondly, we divided the trials into two sets: the first set was those in which the mean of the whole display would give the correct answer in the 2AFC task, (i.e. it was closer to the test circle that matched the mean of the relevant subset than it was to the alternative test circle); the second was the reverse. Participants did better when the mean of the whole set was closer to the correct answer in both the cued condition ($t_{(4)} = 4.005$, $p < .05$, mean = 79.4%) and the non-cued condition ($t_{(4)} = 3.184$, $p < .05$, mean = 78.6%). However, even when the mean of the whole display was closer to the incorrect answer in the 2AFC task, participants still chose the mean of the relevant subset 55.5% of the time in the cued condition and 52.6% of the time in the non-cued condition. If they had relied solely on the mean of the whole display, they would have chosen the test circle that matched the mean size of the whole display on 100% of those trials. It seems then that participants were trying to select the relevant circles, as asked, but that their selection was imperfect.

The thresholds were higher in this experiment than in Experiment 2. This is probably due to the difference in method used. In Experiment 2, participants directly compared the mean sizes of two subsets of circles, whereas in Experiment 3, they were asked to select which of two later presented circles matched the mean of one selected subset. This may be more similar to an absolute judgment task than to the discrimination task used in Experiment 2. Chong and Treisman (in press) also found high mean discrimination thresholds (around 25%) using a task in which participants saw one display of mixed sizes, followed by a forced choice between two test circles of the one that matched the mean size of the preceding display.

There are some possible alternative accounts of the absence of difference between the cued and non-cued conditions. It could be that the test circles appearing immediately after the stimulus presentation interfered with a persisting or iconic representation of the previous stimulus. This should affect all three conditions, since the test circles were always present, but perhaps it imposed a ceiling effect such that the cue, when present, could not be used effectively. Iconic persistence seems unlikely because there is little if any iconic persistence with exposures of 200 ms (Di Lollo, 1980) and this exposure duration should allow adequate time to use the cue when it was given in advance. Interference from the test circles to a persisting representation of the previous display is also unlikely because the thresholds for mean size in Chong and Treisman (2003) without the test circles were similar to those found by Ariely (2001) with a test circle. Another account might be that color-based selection is inefficient. Moore and Egeth (1998) found that color-based attention helped only resource-limited search tasks and not data-limited ones. It is not clear which ours would be. However, the fact that we found no difference between the single color and the cued condition suggests either that our participants could select efficiently by a cued color, or that the circles in the non-cued color produced no interference because they were automatically segregated before the mean sizes were computed. Furthermore, in Experiment 2 we showed that intermingled sets of circles segregated only by color gave mean thresholds that were as accurate as sets segregated by location, suggesting efficient segregation by color.

10. General discussion

These three studies have explored some aspects of structural constraints on the statistical averaging process. We looked both at the abstraction of sizes from number and density and at the segregation of different subsets for statistical description. We also provided further evidence that participants do compute the mean size, as requested, rather than relying on some other cue. We had previously shown that mean size thresholds are similar when comparisons of mean size are made between samples drawn from the same distribution and

samples drawn from different distributions (Chong & Treisman, 2003). Three further tests are described in the present paper: First, we showed that participants can accurately discriminate mean sizes between sets in which the mean is varied by changes in the relative frequencies of different sizes, keeping the sizes themselves constant, and sets in which the mean is varied by changing the actual sizes presented, keeping the frequencies equal. This ensures that participants cannot rely on comparisons of individual sizes or on comparisons of particular subsets. The use of displays with only two sizes also made it highly unlikely that participants used the modal size rather than the mean.

Secondly, the results of Experiment 1 showed that the mean could be judged equally well across sets with different densities and with different numerosity, and almost as well across sets in which both the density and the numerosity differed. Finally, the additional analysis of Experiment 2 confirmed that participants indeed calculated the mean rather than the median. On trials for which the two gave inconsistent results, their judgments reflected the mean and not the median.

The results also suggest some quite sophisticated segregation of processing in which size is separated from other quantitative variables. One way in which this could be achieved is within specialized modules that abstract particular features for processing, both of group statistics and of individual discrepant elements. Studies of selective attention suggest that selection follows a stage of preattentive segregation that groups elements by salient feature differences, producing candidate sets for subsequent selective processing. Treisman and Gormican (1988) used the idea of pooling and averaging within the separate feature maps proposed in feature integration theory in accounting for the detection of outliers in visual search tasks, and more particularly for search asymmetries. Does the extraction of statistical parameters operate on the same preattentive groupings? A feature map for sizes would automatically separate responses to the mean size from responses to the number and the density of elements.

Further evidence from Experiments 2 and 3 can clarify the question whether the mean size is computed on segregated feature maps or whether heterogeneous sets are pooled in some earlier representation, for example the map of filled locations proposed in feature integration theory (Treisman & Gelade, 1980). To be useful, it would make more sense for perceptual statistics like the mean size to be selective within groups comprising entities that are likely to constitute parts of the same real world object, or set, or region. In Experiments 2 and 3, we tested whether means can be computed separately for spatially intermingled subsets defined only by color differences. We already knew that means can be computed separately for different spatial areas, specifically for the left and right sides of the visual field (Chong & Treisman, 2003), but selection there could be based on spatial attention or on left versus right hemisphere stimulation. Our new results with intermingled colors were surprising. The thresholds for discriminating the mean sizes of sets differentiated by color were as accurate as those for sets differentiated by location, in Chong and Treisman (2003).

Is the averaging process restricted to one subset at a time, or can it be applied in parallel to two or more separate subsets? If the computation is carried out automatically across each of the perceptual groups present in the display, the timing of cues to the relevant subset, and even the removal of the irrelevant subset should make little difference to performance. A striking finding in our experiment is that the presence of an irrelevant set of circles had little impact on accuracy, comparing the non-cued and single color conditions. It seems that means are computed automatically for both color sets, so that the cue and even the single color presentation confer no advantage over the no-cue. By "automatically" here we mean "in parallel" and probably "without intention" rather than "without attention", since attention was certainly directed to the mixed display. Chong and Treisman (2003) provide evidence that the averaging process is also free of interference from a dual task if the concurrent task requires global attention to the display as a whole. The present results suggest that the averaging process follows, and is constrained by, an early perceptual grouping by color, but it precedes the limited capacity bottleneck that forces selective attention.

This result is quite surprising in the context of research on binding. The task of averaging sizes selectively for color subsets would seem to require binding of sizes and colors. This would be the case if the averaging process were applied to individuated objects. However there may be an alternative strategy, similar to that proposed for guided search (Treisman, 1988; Treisman & Sato, 1990; Wolfe, Cave, & Franzel, 1989). If selection is based on selective activation or inhibition from different feature maps, it could produce candidate sets of items to be averaged independently, without separately binding a color and size to every item. Performance in the present averaging tasks suggests that this is the strategy used.

Research in other paradigms supports the view that the averaging process is automatic. For example, it does not depend on conscious access to the individual items to be averaged. Crowding in the visual periphery, a form of attentional overload, can eliminate perception of particular individual items (He, Cavanagh, & Intriligator, 1996). Yet Parkes, Lund, Angelucci, Solomon, & Morgan (2001) showed that humans could reliably estimate the average orientation even in conditions in which they were unable to report the orientation of any individual patch. Again this suggests automatic averaging of feature information. Using a display containing many different local directions of motion, Watamaniuk and McKee (1998) found that participants could either form a unified global

percept of motion in the direction of the mean or focus on one local direction of motion. Direction discrimination thresholds in the post-cued condition were not significantly higher than those obtained in the pre-cued condition, suggesting that direction information for both global and local motion is encoded in parallel. However, Watamaniuk and McKee did not test whether the two unified global percepts representing local directions of motion could be simultaneously extracted, analogous to our comparison of cued and non-cued displays for mean size judgments. Another form of automatic computation of a spatial mean comes from Melcher and Kowler (1999)'s study. They showed that saccades accurately landed near the center-of-area of the target shape, rather than at the center-of-gravity of the target or on the symmetric axis. Furthermore, accuracy of landing near the center-of-area was not affected either by changes in the spacing of the dots or by added dot clusters as long as they did not change the shape of target.

Our findings contribute to this exploration of preattentive processing, showing that average information is extracted for subsets of location or color-defined elements as part of an early global structuring of the visual scene. Future research will test what other features can define the separate sets for the averaging process, and what other statistics, besides the mean size, are automatically computed.

Acknowledgement

This research was supported by NIH grant number 1 RO1 MH58383, by Israeli Binational Science Foundation grant # 1000274, and by Conte Center Grant, number MH062196.

References

Allik, J., Tuulmets, T., & Vos, P. G. (1991). Size invariance in visual number discrimination. *Psychological Research, 53*, 290–295.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*, 157–162.

Beck, J. (1966). Perceptual grouping produced by changes in orientation and shape. *Science, 154*, 538–540.

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. (in press). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*.

Di Lollo, V. (1980). Temporal integration in visual memory. *Journal of Experimental Psychology: General, 109*, 75–97.

Dakin, S. C. (1997). The detection of structure in glass patterns: Psychophysics and computational models. *Vision Research, 37*, 2227–2246.

Finney, D. J. (1971). *Probit Analysis*. Cambridge, UK: Cambridge University Press.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature, 383*, 334–337.

Melcher, D., & Kowler, E. (1999). Shapes, surfaces and saccades. *Vision Research, 39*, 2929–2946.

Moore, C. M., & Egeth, H. (1998). How does feature-based attention affect visual processing? *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1296–1310.

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.

Teghtsoonian, M. (1965). The judgment of size. *American Journal of Psychology, 78*, 392–402.

Treisman, A. (1988). Features and objects: the fourteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology A, 40*(2), 201–237.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 16*, 97–136.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological Review, 95*, 15–48.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 459–478.

Watamaniuk, S. N. J., & McKee, S. P. (1998). Simultaneous encoding of direction at a local and global scale. *Perception & Psychophysics, 60*, 191–200.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model of visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 419–433.