Attention, Perception, & Psychophysics (2020) 82:865–876
https://doi.org/10.3758/s13414-019-01938-7

40 YEARS OF FEATURE INTEGRATION: SPECIAL ISSUE IN MEMORY OF ANNE TREISMAN

Perception of means, sums, and areas

Aire Raidvee [1,2] · Mai Toom [1] · Kristiina Averin [1] · Jüri Allik [1,3]

Published online: 20 February 2020

Abstract
In this age of data visualization, it is important to understand our perception of the symbols that are used. For example, does the perceived size of a disc correspond most closely to its area, diameter, circumference, or some other measure? When multiple items are present, this becomes a question of ensemble perception. Here, we compare observers' performance across three different tasks: judgments of (i) the mean diameter, (ii) the total diameter, or (iii) the total area of (N = 1, 2, 3, or 7) test circles compared with a single reference circle. We draw a parallel between Anne Treisman's feature integration theory and Daniel Kahneman's cognitive systems, comparing the preattentive stage to System 1, and the focused attention stage to System 2. In accordance with Kahneman's prediction, average size (diameter) of the geometric figures can be judged with considerable accuracy, but the total diameter of the same figures cannot. Like the total length, the cumulative area covered by circles was also judged considerably less accurately than the mean diameter. Differences in efficiency between these three tasks illustrate powerful constraints upon visual processing: The visual system is well adapted for the perception of the mean size while there are no analogous mechanisms for the accurate perception of the total length or cumulative area. Thus, in visualizing data, using bubble charts proportional to area may be misleading as our visual system seems better adapted to perceive disc size by the radius rather than the area.

Keywords Mean size perception · Sum size perception · Perception of area · Ensemble characteristics · System 1 and System 2 · Kahneman's conjecture

Bubble charts were invented by a French cartographer Charles Joseph Minard (1781–1870) to convey numerical information on maps (Friendly, 2008). In these maps, data points are shown as bubbles or discs, the size of which represents the values on some relevant dimension (e.g. size of a population; Szafir, Haroz, Gleicher, & Franconeri, 2016). In principle, the size of discs can be specified in multiple ways, such as the radius, diameter, circumference, or area, to say nothing about other, less conventional measures. If one chooses a measure that is not directly proportional to the apparent size of these discs, then it could end in misleading results. Sometimes, it is recommended to choose area because it is believed that the human visual system naturally experiences disk's size in terms of its area.[1] This recommendation, however, goes astray because it was established long ago that a disc's size is not typically experienced in terms of its area. Although the subjective magnitude of size increases in direct proportion with the length of a straight line (Hartley, 1981; Stevens & Guirao, 1963), the apparent size of a disc is not proportional to its area (Ekman & Junge, 1961; Schneider & Bissett, 1988; Stevens, 1975; Stevens & Guirao, 1963). The apparent size of discs or circles can be described by a power function with an exponent close to .70 (Stevens, 1975) or even less (Li, Martens, van Wijk, & ACM, 2010). This means that the disk's size is perceived proportionally to the disc's diameter rather than to its area (for exceptions from this rule, see Cleveland, Harris, & McGill, 1982).

It was realized some time ago that a human observer is an intuitive statistician (Peterson & Beach, 1967). For example, visual displays may contain multiple similar objects or events, and the observer is capable of making reasonably accurate

✉ Aire Raidvee
aire.raidvee@ut.ee

[1] Department of Psychology, University of Tartu, Näituse 2, 50409 Tartu, Estonia

[2] New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[3] Estonian Academy of Sciences, Tallinn, Estonia

[1] http://visage.co/data-visualization-101-bubble-charts/ or https://en.wikipedia.org/wiki/Bubble_chart.

statistical summaries of them. Typically, participants are asked for estimates of the proportion, mean, variance, correlation, or some other descriptive statistic of some attributes of these objects or events. The correspondence between the estimates and the calculated statistics serves as the measure of accuracy (Peterson & Beach, 1967). Based on these ideas, it was demonstrated that the perceived average size of an array of lines is indeed proportional to the mean value of their actual lengths (Miller & Sheldon, 1969). This is an indication that the visual system is able to represent, beside isolated physical attributes, ensemble characteristics that are an abstract property of an incoming visual image that is computed from multiple individual measures (Alvarez, 2011; Whitney, Haberman, & Sweeney, 2014).

The concept of the intuitive statistician experienced a considerable surge of enthusiasm after widely acclaimed studies which showed that people can make surprisingly precise judgments about the average size of multiple geometric objects, typically lines or circles (Ariely, 2001; Chong & Treisman, 2003, 2005). These studies were mainly inspired by a consideration that the reduction of a set of similar items to a prototypical mean helps to economize on the limited capacity of the visual system by replacing multiple representations of individual elements with their statistical summary characterizing the set as a whole. It was proposed that the process of perceptual averaging is carried out automatically, presumably by an array of parallel "computers," the main function of which is to help bypass the bottleneck of focused attention (Alvarez, 2011; Ariely, 2001, 2008; Chong, Joo, Emmanouil, & Treisman, 2008; Chong & Treisman, 2003, 2005; Whitney et al., 2014; Whitney & Leib, 2018). However, even one of the core claims that mean size can be computed outside of focused attention did not meet the expectations, because it was demonstrated that all published evidence can be explained through various focused-attention strategies, without invoking a special mechanism for the average size perception (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013; Myczek & Simons, 2008; Simons & Myczek, 2008; Solomon, Morgan, & Chubb, 2011). Although the idea of effortless and massively parallel computation sounded attractive, a number of compelling demonstrations showed that the observer's decisions about the mean size had a precision that could have been achieved as if only a few elements had been taken into account (Allik et al., 2013; Myczek & Simons, 2008; Utochkin, 2015). Thus, the visual system can indeed make statistical summaries as if only a limited subsample of individual objects was used (Allik et al., 2013; Dakin, 2001; Legge, Gu, & Luebker, 1989; Solomon et al., 2011)

Researchers of the ensemble characteristics were able to describe many interesting properties of the mean size perception, such as how different visual cues (Chong & Treisman, 2005), item heterogeneity (Marchant, Simons, & de Fockert, 2013), grouping (Im & Chong, 2014), previous adaptation (Corbett, Wurnitsch, Schwartz, & Whitney, 2012), exposure time (Whiting & Oriet, 2011), or crowding (Banno & Saiki, 2012) affect the ability to estimate the mean size. However, surprisingly little attention has been paid to the question of what specific stimulus attributes the judgements of the statistical aggregation are based on. The fact that participants were instructed to judge the mean size or any other statistical attribute of a collection of objects is not a guarantee that the visual system is able to carry out the instructed task properly (Morgan, Hole, & Glennerster, 1990). For example, it was proposed that observers have no access to high-precision codes for a two-dimensional area, and that they base their decisions on various heuristics derived from linear measures (Morgan, 2005). In all these cases of perceptual constraints, subjects are relying (while not necessarily being aware of it) on some other visual attributes or their combination, which serve as a reasonable proxy for the instructed property (Morgan & Glennerster, 1991; Seizova-Cajic & Gillam, 2006; Westheimer, 2008). In other words, a stimulus attribute specified in the instruction can be substituted with other stimulus attributes that are only surrogates for the intended one.

In many cases, the instruction is not constraining enough to specify which out of many equally plausible attributes is actually used for the judgement. For example, the instruction to compare sizes of two circles is ambiguous because it may mean their diameters, but it could also mean their areas. With only two circles for comparison, it does not matter which of the two attributes—diameter or area—is used for the comparison. However, the situation becomes ambiguous if instead of a single circle the task is to judge the mean size of three or more circles. The mean diameter of a set of circles is not proportional to the total area they cover on the plain. As an example, let us suppose that we have four test circles with radiuses r = 1, 2, 4 and 5 distance units, respectively. If we look for a reference circle that matches the mean radius of these four test circles, then it would be a circle with the radius of 3 distance units. However, the total area occupied by these four test circles would be equal to the area covered by a single reference circle with the radius of about r = 6.78 distance units. Four identical circles with the equal radiuses r = 3.39 are needed to cover the equivalent total area. Consequently, it makes a difference in which terms—diameter or area—the size of circles is specified.

A recent study shows that the apparent increase or decrease in the mean size of four circles will be perceptually identical, whether we add, for instance, 4 distance units to the diameter of only one of four test circles presented on the display, or we add 1 distance unit to the diameters of all four circles (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2014). Intuitively, it is more likely that the human observer can more easily notice an outlier that is 4 radius units larger than the reference size, rather than four small increments of 1 radius unit added to each of the four test circles. However, the results showed, in

harmony with the rules of arithmetic, that these two cases result in an identical perceptual outcome, which indicates that the visual system is insensitive to the grouping of increments and is tuned only to their mean size (Allik et al., 2014). Thus, the visual system is indeed able to compute a characteristic that is sufficiently close to the mean value of the test circles. Interestingly, this result automatically excludes the possibility that the aggregate size of circles was perceived in terms of their area. Please note that if we added 4 distance units to the radius of only one of the four test circles, then the total area occupied collectively by four circles was increased by 16 area units, whereas adding 1 distance unit to the radiuses of all four circles increased the total area by only 4 area units. Consequently, it was almost certainly not the area that was judged when the mean size of the test circles was compared with the size of a reference circle (Allik et al., 2014).

An intriguing idea about constraints imposed upon visual processing was recently advanced by Daniel Kahneman (2011) in his influential book *Thinking Fast and Slow*. Among many interesting observations, Kahneman proposed that the average length of randomly positioned lines can be judged with a considerable accuracy, but the total length of these lines cannot (pp. 92–93). In a sharp contrast to the accuracy of the mean size discrimination, he predicted that the visual system is expected to perform very poorly when the total length of these lines would be judged. According to Kahneman's idea, the mean size of a collection of nearly identical geometric figures can be computed by an evolutionarily old System 1, which is producing rapid, parallel, and automatic analysis, where only the final product is accessible to the cognitive awareness. On the other hand, System 2 is evolutionarily more recent and performs slowly, using sequential processes typical of deliberate thinking (Kahneman, 2011). Kahneman proposed that the task of estimating the total length of multiple objects activates System 2, which will receive from System 1 the average size and thereafter multiply average size by the number of estimated objects in order to compute the sum of sizes (p. 93).

However, it is straightforward to notice that the idea of these two systems was inspired by Anne Treisman's feature integration theory, which postulates two different stages of integration—preattentive and focused attention (Treisman, 1988; Treisman & Gelade, 1980). In the first stage, visual features are combined into ensembles or conjunctions automatically, without effort or attention by the perceiver. The second stage is orchestrated by focused attention, the main function of which is to associate or "glue" each of these features with the object to which it belongs. Because these two feature integration mechanisms work with different speeds, capacity limits, and conjunctions they are capable of forming, they leave signatures which makes it possible to identify which of these two stages, preattentive or focused attention, was used. For example, if we know that one of the involved systems cannot multiply and we detect signs of multiplication, then it is a tight alibi that this system was not responsible. This is a good example of how one brilliant idea can stimulate other equally great ideas.

However, Kahneman's proposal that the sum is derived from the mean value is by all means unorthodox. Every technical definition of the "mean" presupposes summation: adding up values and then dividing by the number of added values. Kahneman seems to think that information about the total length may even exist somewhere in the visual system, but it is cognitively inaccessible. This seems feasible in light of the recent experiments (Chetverikov, Campana, & Kristjánsson, 2016, 2017) showing that explicit measures of probing the internal representation may not reveal them fully. To cope with the absence of cognitive access, the sum of individual sizes can be reconstructed by a reverse operation multiplying the mean value by the number of added elements. However, previous studies have shown that discrimination of only two shapes based on their linear dimensions is considerably better than discrimination that is based on comparing areas of these two stimuli (Morgan, 2005; Nachmias, 2011). It seems that the magnitude of a two-dimensional area is not directly represented in the perceptual system, and decisions about area must be based on inferences resulting from linear measures (Morgan, 2005). Moreover, there is no information about the observer's ability to combine various measures derived from various test stimuli. Without direct experimental results, we are left with mere speculations on why the visual system deals relatively well with averages, but poorly with sums (Kahneman, 2011, p. 93).

Nevertheless, the main idea of the current study was inspired by Kahneman's conjecture that gave us an insight to compare the precision of discrimination under three different instructions:

1. The first task was a usual assignment to compare the mean diameter of *N* randomly positioned circles with the size of a reference circle. Participants were instructed to indicate whether the mean diameter of *N* circles was smaller or larger than the size of a single reference circle.
2. The second task, carried out in a separate series, was to compare the area of the test circle with the apparent total area of *N* test circles.
3. Finally, in the third task, participants were asked to judge whether the diameter of the test circle was smaller or larger than the total diameter of *N* test circles.

There are many studies in which participants are instructed to judge the mean size but very few in which the total height/width or total area are estimated (for an exception, see Lee, Baek, & Chong, 2016). The right panel in Fig. 1 shows three randomly positioned circles with unequal diameter. On the left panel, three different reference circles are shown, each

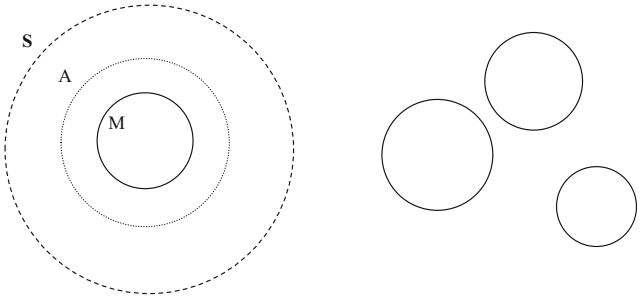

Fig. 1 The tasks are to estimate the mean diameter, the total area, and the total diameters of three test circles on the right compared with a left reference circle corresponding either to the mean diameter (*M*), total area (*A*), or summed diameters (*S*) of these three test circles

corresponding to the mean diameter (*M*), total area (*A*), and sum of diameters (*S*) of the three test circles shown in the right panel. It is important to notice that all three statistics—mean, area, and sum—are different from one another.

Previous studies have shown that the mean size can be estimated with the maximal precision of 4%–7% from the size to be judged (Allik et al., 2013; Myczek & Simons, 2008). If Kahneman's conjecture that total length is more difficult to judge than the mean size holds then the precision with which the size of the reference circle *S* could be discriminated from the total diameter of *N* test circles is considerably smaller. To our knowledge, there is only one study in which the mean size judgement is directly compared with the summed area judgement (Lee et al., 2016). Although Kahneman said nothing about the precision of determining the cumulative area, following his rationale, it would be logical to expect that determination of total area is also a difficult task, comparable to the finding of the total diameter. The results of Lee et al. (2016) suggest that the mean size can be judged more accurately than the cumulative area. However, in addition to the mean size and cumulative area, our study investigates the third logical option as well—the judgement of the total diameter of test circles.

Methods

Participants

Six participants, including authors of this paper, with normal or corrected-to-normal vision, participated in this study. The observers had various prior experiences with vision perception experiments. For organizational reasons, two observers were able to participate in two out of three series of experiments.

Apparatus

Stimuli were presented on various LCD monitors having resolution at least 1,920 × 1,080 pixels. The program was set to adjust stimulus resolution and calculate recommended viewing distance in order to compensate for possible variations in screen sizes. The adjustments in viewing distance were made to assure that one pixel would subtend to 2 minutes of arc for every participant. Experimental programs were written in MATLAB (The MathWorks, Inc.) using Cogent 2000, developed by the Cogent 2000 team led by John Romaya at the Laboratory of Neuroscience at the Wellcome Department of Imaging Neuroscience (http://www.vislab.ucl.ac.uk/cogent.php).

Stimuli and procedure

In each trial, the display consisted of two dark grey disks—stimulus areas—on a black background. Both of the areas were approximately 16.3° of visual angle in diameter and were presented on the left and right side from the central fixation mark with a gap of 1.5° between them. A set of target elements were presented in one of these background disks, and a reference on the other. The stimulus presentation time was 1 second. The location for the test and reference stimuli between the two background areas were chosen randomly before each trial. The test stimulus consisted of one, two, three, or seven ($N = 1, 2, 3$ or 7) randomly positioned, spatially nonoverlapping white unfilled circles with various sizes. The reference stimulus was a single circle with a size corresponding approximately to the mean diameter, the cumulative area, or the total diameter of the test circles. To assure that elements would not overlap or cross a border of a panel, inhibitory area was set around each circle and on panel borders. In one-element conditions ($N = 1$), the element was always presented at the center of the panel.

Participants were instructed to indicate by a corresponding mouse click which of the two stimulus areas, the right or left, had the greater magnitude on the designated attribute. After response, an auditory feedback about correctness of the answer was given. In case of a correct answer, a sound with high tone was played, and in case of an incorrect answer, a sound with low tone was played.

There were three different series of experiments corresponding to one of the three instructions:

1. In the mean diameter task, the observers were instructed to indicate which of the two stimuli had the larger mean diameter.
2. In the cumulative area task, the observers were asked to tell which of the two stimuli had the larger cumulative area.
3. In the total diameter task, the observer's instruction was to tell if the total diameter of the presented test circles was larger or smaller compared with the diameter of the single reference circle.

When there was only one test circle ($N = 1$), these three tasks were formally identical. Therefore, it is expected that psychometric functions from these three conditions are nearly identical, irrespective of different instructions.

In each trial, the mean diameter, cumulative area, or total diameter of the test elements was set to differ from the reference by increasing or decreasing the mean or summed size of the base-set elements by a variable delta *(Δd)* (for details, see Allik et al., 2013). The length of the diameter of the elements in the base set was determined by dividing the diameter (D) of an element in a single element condition [9.5°; 10.5°] into the number of diameters in range [$D/N \times 0.95°$; $D/N \times 1.05°$] equal to the number of elements in a set. The length of the diameter of the reference circle thus depended on the number of test elements and the task.

There was an inherent discrepancy between the mean diameter on the one hand, and the cumulative area and the total diameter tasks on the other. Unlike in the mean diameter task, the increase of the number of test circles N also increased the magnitude of the designated stimulus attribute, either the cumulative area or the total length of diameters. For instance, in the cumulative area task, the area of the reference was equal to the cumulative area of base-set elements. In order to equalize different tasks, it was decided that the size of the test circles was diminished proportionally to the number of test circles. This means that the size of the reference circle decreased proportionally to the number of the test elements. Thus, on average, in the cumulative area task, the diameter of the reference stimulus in the single element condition was 1.4 times longer than the diameter in a condition with two elements, 1.7 times longer than the size in a condition with three elements, and 2.6 times longer than in a condition with seven elements. In the mean diameter task, diameter of the reference stimulus in the single element condition was either 2, 3, or 7 times longer than the reference diameter in a condition with two, three, or seven elements, respectively (see Fig. 2 for clarification).

Sizes of base-set elements were increased or decreased to achieve changes in the diameter Δd of the test set from the corresponding reference. Average standard deviations of diameters in test sets with two, three, and seven elements were 3.8, 2.9, and 1.5 pixels, respectively. For every number of test elements, the average of diameters of elements in test sets differed by less than 4 pixels between the tasks. The deltas were selected randomly before each trial from a set of eight possible values symmetrically around a fixed reference value. In the mean size task $\Delta d = -12, -8, -4, -2, 2, 4, 8$, or 12 pixels, and in the total area and summed diameter $\Delta d = -36, -24, -12, -6, 6, 12, 24$, or 36 pixels.

Empirical data were approximated by a cumulative normal distribution after searching for the best fitting values for the mean (μ) and the standard deviation (σ). The mean μ of the psychometric function corresponds to the increment or decrement in the average Δd *relative* to the reference, which was judged subjectively equal to the size of the reference circle. The standard deviation marks the slope $1/\sigma$ of the psychometric function in pixels corresponding to the just noticeable difference (JND). In this particular case, the test elements deviating from the reference by the JND were correctly discriminated from the reference size in 84.1% of all trials.

Results

Probabilities with which the mean diameter of the test circles was chosen over the size of the reference circle together with the best fitting psychometric curves are shown in Fig. 3. As the relative size of the test elements Δd increased, the side accommodating the test stimulus was chosen more frequently as an answer. Each column corresponds to the number of the test circles ($N = 1$, 2, 3, or 7), and rows of the panels correspond to data from the six observers (AR, JA, KA, MS, MT, and RN). The bottom row corresponds to the aggregate results across all six participants. Dashed lines represent curves of cumulative normal distribution corresponding to the best fitting values of the mean μ and the standard deviation σ. In all cases, the fit of the psychometric curve to the data points was satisfactory. The correlation between the observed and predicted values was on average $r = .98$.

In line with previous research, there was a general tendency that the precision (as measured in pixels) of the mean diameter discrimination increased with the number of the test circles. However, in terms of the Weber fraction, the precision of the mean diameter discrimination decreased with the number of test circles. On average, the size of a single test circle ($N = 1$) was reliably discriminated from the reference when their diameters differed by approximately 12 pixels or 24 minutes of arc (or 4% of the size of the reference). However, a difference of about 3 pixels or 6 (or nearly 7% of the size of the reference) was necessary for the mean diameter of the seven circles to be confidently discriminated from a single reference circle (see the bottom row of panels in Fig. 3). It is well known that the discrimination precision in absolute units (here, pixels or angles of the visual field) increases as the length of the judged spatial interval becomes shorter (Allik et al., 2013; Burbeck & Hadden, 1993; Wolfe, 1923). Because the size of the test circles was given in absolute terms, discrimination precision improved with the decrease of the circles' diameter (and increase in the number of test elements). On the other hand, in several previous studies, the precision of the mean size discrimination remained approximately constant or increased with an increase in the number of test elements (Allik et al., 2013; Ariely, 2001; Marchant et al., 2013).

Figure 4 demonstrates data and psychometric curves for the total diameter judgement task. Like in Fig. 3, the columns of panels correspond to the number of elements and the rows of panels correspond to the observers, with the final row showing

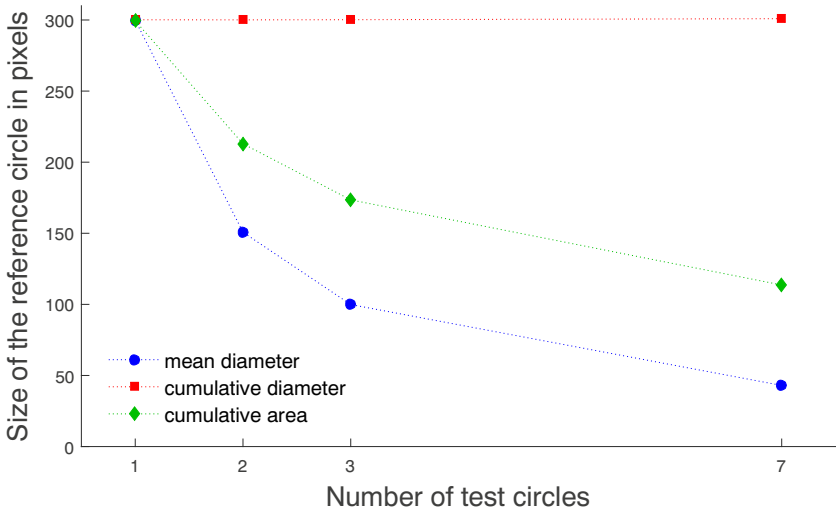

Fig. 2 Size of the reference circle (in pixels) in three different conditions where the task was to estimate the mean diameter (meanD), the total diameter (sumD), or the total area (sumA) of the test circles

aggregate data. Relative size of the reference circle to the total diameter of the test circles ($\Sigma d/\varepsilon$) is given on the horizontal axis, and probabilities of responding that the sum of diameters of the test circles is greater than the reference on the vertical axis. Unlike the mean size judgement task, the precision with which this task was solved decreased with the number of the test circles. For example, the sum of diameters of the seven test circles needed to be about 34 pixels (1.1 degrees, or 11.3% of the size of reference) smaller or larger than the diameter of a single test circle to be confidently discriminated. This is more than 11 times in absolute, and 1.62 times in relative (Weber fraction) terms less precise than the discrimination of the mean size.

Finally, Fig. 5 shows results for the third series of experiments in which participants were instructed to judge the total area covered by all test circles. Like the previous summed diameter task, the accuracy of the total area judgement deteriorated with the number of elements. On average, the area of the seven ($N = 7$) test circles was judged with the precision of about 34 pixels or 22% of the size of the reference (the rightmost panel in the last row of Fig. 5). Thus, judgments of the total area covered jointly by the test circles were more than 11 times in absolute and 3.14 times in relative (Weber fraction) terms less accurate compared with the precision with which the mean diameter could be discriminated.

Figure 6 summarizes results showing how the standard deviation of the psychometric function (σ) changes with the number of the test circles in the three tasks. Despite individual differences, a general pattern is obvious. The mean diameter discrimination precision generally improves with the number of elements. Conversely, the accuracy of the total area discriminations becomes worse in most observers as the number of judged elements increases. Unexpectedly, the precision of discriminating the total diameter is similar to the precision of discriminating total area occupied by the test circles in all but one observer. As we already mentioned, the precision of the mean diameter discrimination is on average more than 11 times more accurate than the discrimination of the total diameter or cumulative area. Even after normalization (σ divided by the diameter of the reference circle), the advantage of the mean diameter discrimination remained over the total diameter and cumulative diameter judgements.

Discussion

Researchers were not immediately excited about the discovery that the mean size of a group of geometric figures can be judged almost as precisely as the length of a single object (Miller & Sheldon, 1969). However, it was later considered sufficient proof of parallel mechanisms that are able to extract summary information about the sizes of all the objects in a display, essentially computing the mean size at a glance (Ariely, 2001; Chong et al., 2008; Chong & Treisman, 2003; Whitney & Leib, 2018). A rather basic statistical consideration predicts that mean size judgment is expected to improve with the square root of the elements to be judged (Fouriezos, Rubenfeld, & Capstick, 2008). Unlike many previous studies, we observed improvement. Unfortunately, we were unable to differentiate the improvement of the statistical efficiency with the increase of N from the decrease of the average size of the test objects that (according to Weber's law) also leads to the increase of discrimination accuracy. Because (unlike the total diameter and cumulative area tasks) only the mean diameter discrimination improved with N or the mean size of the test objects, we were not motivated to separate the effect of N from the effect of Weber's law.

As Kahneman predicted, the precision of the total diameter discrimination was many times worse than the judgement of their mean size. Likewise, our results replicated the recently reported findings that mean size discrimination is considerably more precise than total area discrimination (Lee et al.,

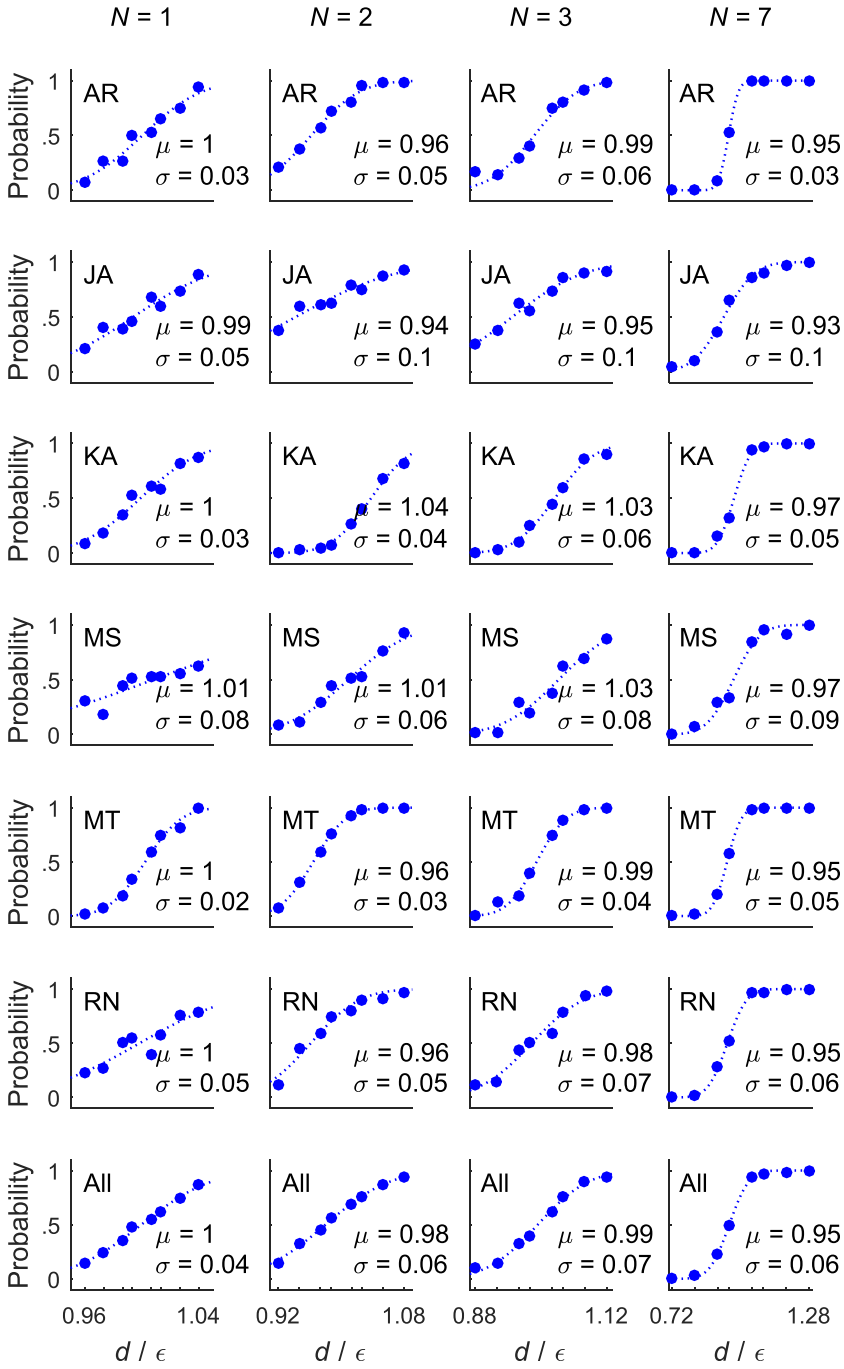

Fig. 3 Data and psychometric curves for the mean diameter discrimination task

2016). As there were multiple objects, their total diameter or area was determined with considerably less precision than their mean diameter. Kahneman's idea, if we understood correctly, was that the observed imprecision in the total size discrimination is caused by multiplication, which is used to reconstruct the summed size from the much more accurate mean size measurement (Kahneman, 2011, pp. 92–93). Because the area of single shapes is discriminated considerably worse than their linear measures (Morgan, 2005; Nachmias, 2011), there may be no need to involve multiplication as an explanation of decreased precision.

However, it is not the first time that mechanisms of mental multiplications have been postulated in visual processing. Although Stevens (1975) promoted the direct scaling methods, it was agreed that the fractionation methods are more reliable and accurate in the construction of psychophysical scales (Torgerson, 1958). The logic of the fractionation methods assumes that a subject is capable of reporting or producing the predetermined magnitude of sensory ratios. Producing or estimating sensory ratios presumes, of course, that the observer is capable of multiplying or dividing sensory magnitudes. However, it was Torgerson (1961)

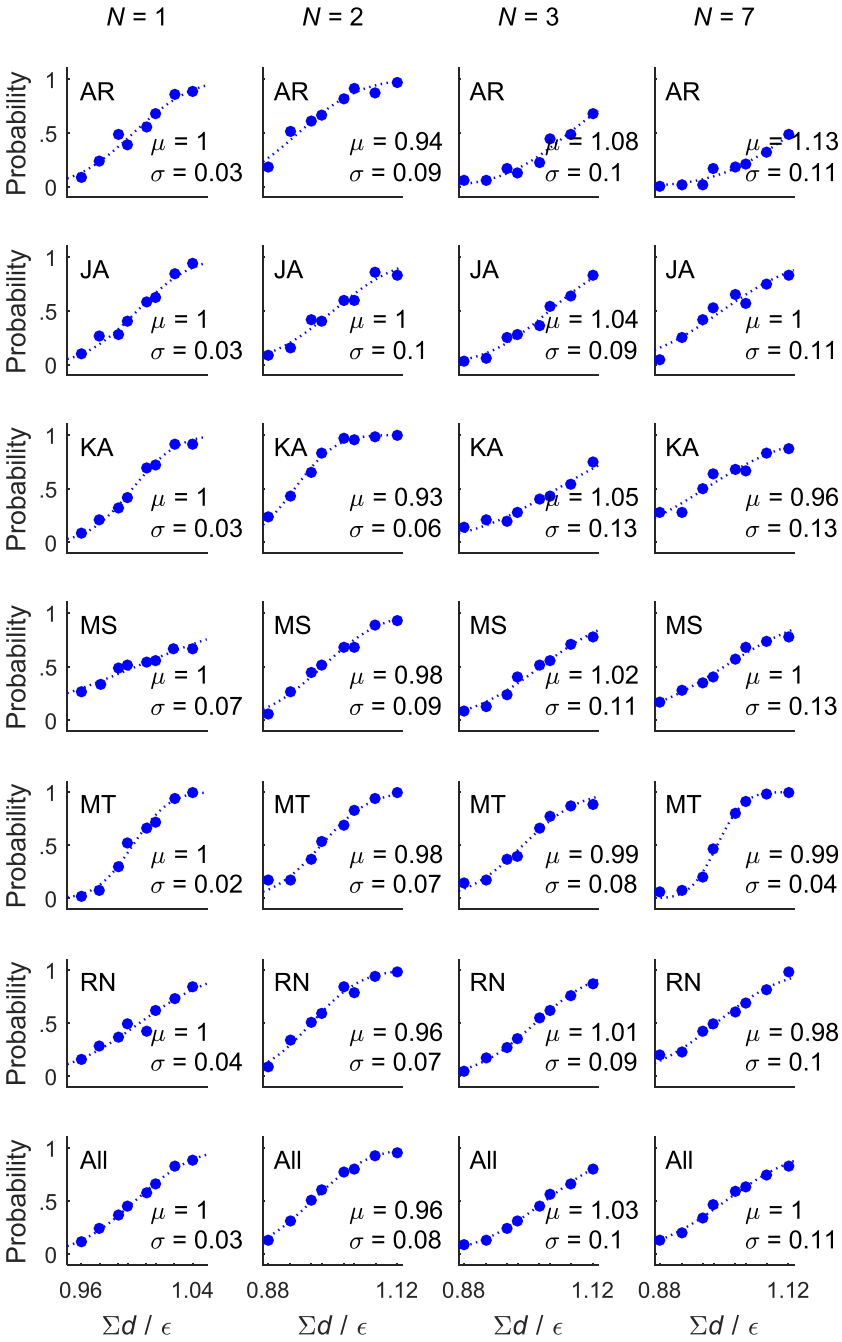

Fig. 4 Data and psychometric curves for the summed diameter discrimination task

who formulated a principle, known as the Torgerson's conjecture: The human observer is not able to distinguish between sensory ratios and sensory differences (Birnbaum & Veit, 1974). Although the Torgerson's conjecture was both supported (Masin, 2013; Narens, 2006) and not supported (Luce, 2012) by later studies, it is possible that multiplication can be substituted, in principle at least, by the judgement of sensory differences. Indeed, the most straightforward way to determine the total height is to measure diameters of each circle and determine the length of a spatial interval they appear to cover collectively. Therefore, there is no need for the multiplication mechanism. Obviously, this may have implications for the Kahneman's conjecture and also indicates that if the perceived diameters can be mentally added together into a straight line, then the measurement of the length of this line cannot be a precise operation. Thus, it is likely that manipulation of spatial intervals in the mind's eye is a noisy process that cannot be done accurately.

Intriguingly, the judgment of total diameters was almost as precise as the judgment of the total area occupied by the test circles. Unlike the cumulative height or width, it is impractical to determine the total area of the test circles from their mean diameter (cf. Lee et al., 2016). Returning to the above given example, four test circles with radiuses $r = 1, 2, 4$, and 5 cover

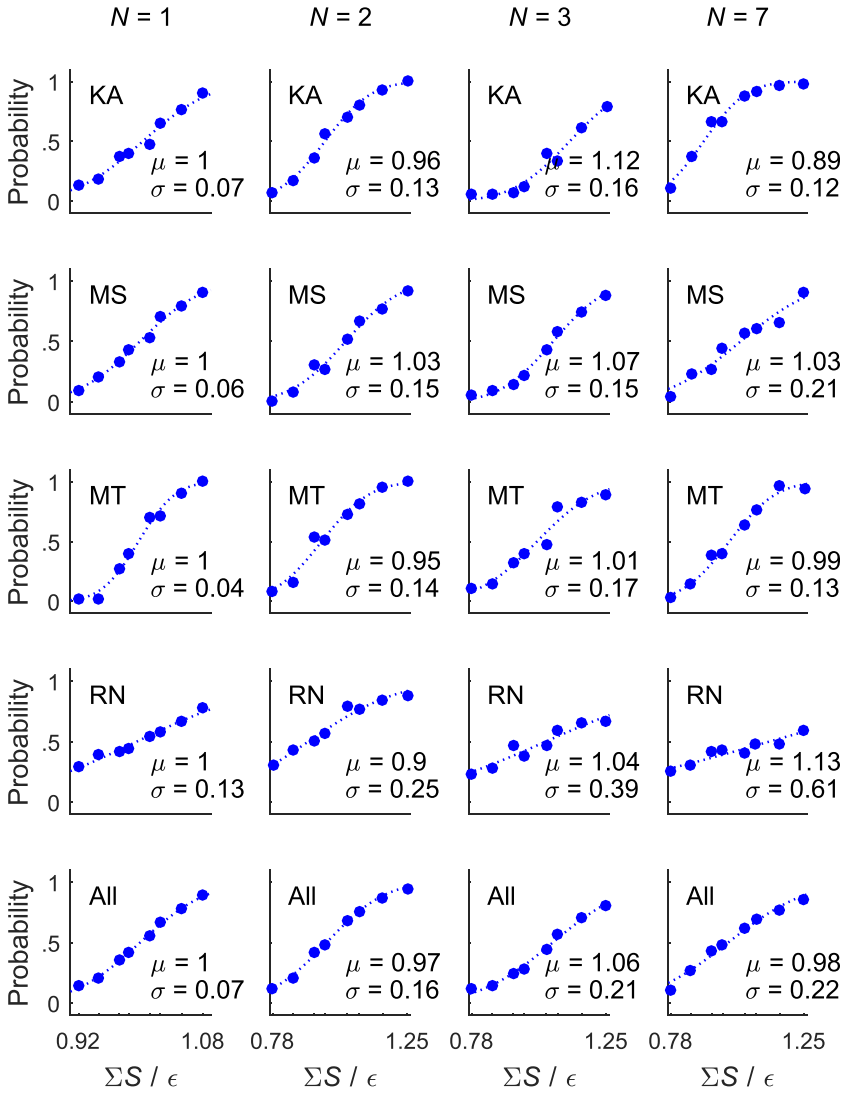

Fig. 5 Data and psychometric curves for the total area discrimination task

a total area which is equivalent to a single reference circle with the radius $r = 6.78$. If we take a circle with the mean size $r = 3$ and multiply its area by 4, we can find that these four circles have the total area equivalent to a circle with the radius $r = 6$, which is about 22% less than the actual area occupied by the aforementioned four test circles. Consequently, unlike the total height, it is at least impractical to use the mean diameter for the computation of cumulative area. Thus, Kahneman was right when he predicted that judging the mean size is an easy task, while determining the sum size is not. However, he was probably mistaken by proposing that such a dramatic difference between these two tasks is due to multiplication, which is allegedly used to solve the sum size tasks. Because there are more convenient ways to determine the sum size, it is problematic to propose a more complicated solution.

Because the total diameter and the cumulative areas were judged with approximately identical inaccuracies, the multiplication is not the only thing that could compromise precision. With only two circles to compare, the instruction to discriminate areas was as accurate as the instruction to discriminate diameters. It is likely that in these minimal cases, area was replaced with the height and width, which were directly proportional to area. However, if we are talking about shapes such as rectangles, then it is even logical to propose that their perceived area can be found through multiplication of their apparent height and width (Teghtsoonian, 1965). Although multiplication seems a likely option, it is not the only one. For example, it was proposed that an additive integration process can replace multiplication in judging two-dimensional area (Anderson & Weiss, 1971). Stepping aside from the world of simple geometric figures, analytically it is tedious to determine the area or volume of irregular shapes or bodies. Nevertheless, Archimedes solved a seemingly impossible problem posed to him by Hiero of Syracuse by immersing an irregular body into a vessel filled with a fluid. So did German mechanic Jakob Amsler, who invented a disarmingly simple device—planimeter—to measure the area of irregular shapes like pieces of land on a map (Runeson, 1977).

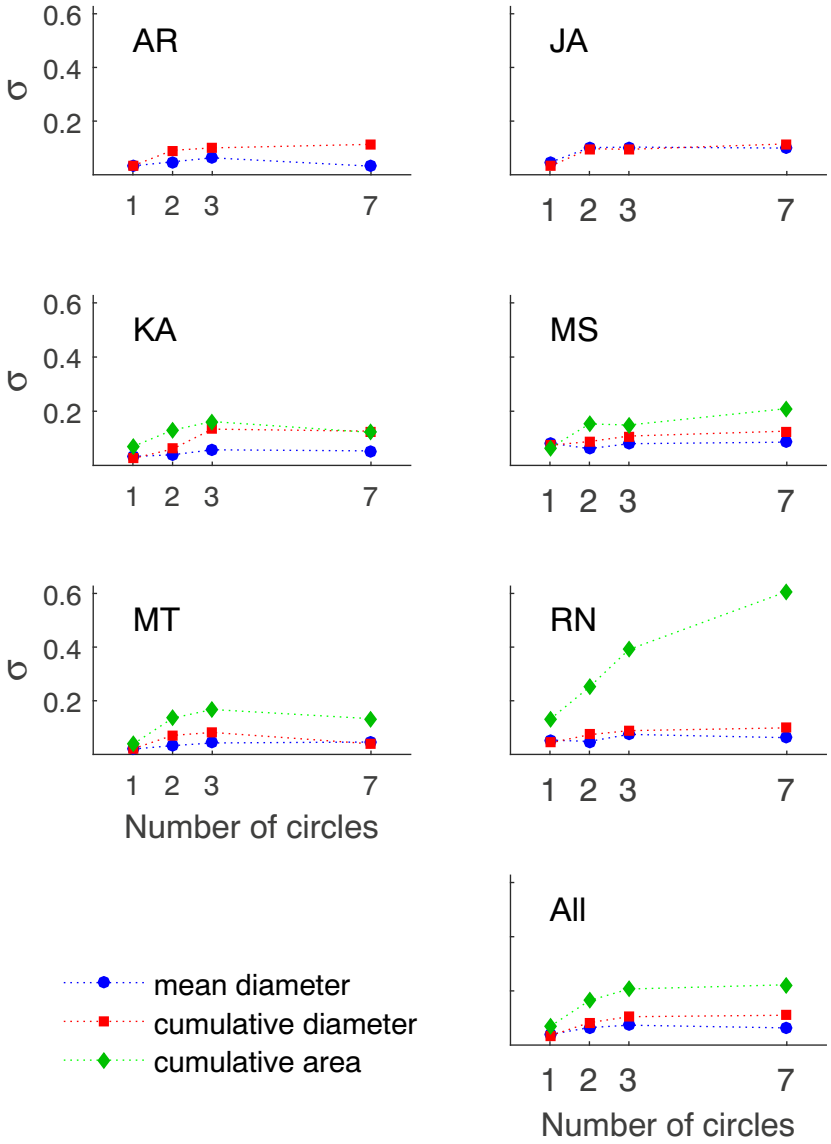

Fig. 6 The values of Weber fractions as a function of the number of the test elements (N = 1, 2, 3, and 7) for three different instructions: mean diameter, summed diameter, and summed area

Ironically, Runeson used the planimeter metaphor to convince his readers of the existence of smart perceptual mechanisms that are capable of measuring the area directly, not deducing it from some linear measures (Runeson, 1977). These smart, specialized mechanisms may exist for some complex perceptual attributes, but it seems unlikely that we have one of them for the perceived area itself.

Thus, the observed inaccuracy in the judging of total diameter and cumulative area may be an illustration of powerful constraints upon visual processing (Morgan, 2005; Morgan et al., 1990). It seems that there is no direct and high-precision perceptual representation for either two-dimensional area or total length of multiple spatial intervals. There seems to be only two ways of improving the precision of perception of these traits that are poorly represented in the visual system. One possibility, quite frequently used, is replacement with proxy variables. For instance, it is complicated to compute the position of each element in a cluster of dots, then replace them by the centroid, which is easy to determine for a cluster (Morgan et al., 1990, Experiment 2). Another option is to learn new perceptual routines, which can improve the precision with which decisions can be made. For example, it seems possible to put all test circles on an imaginary line and measure in this mental image the length of the occupied interval. Unfortunately, we have no information on techniques that could improve the precision of answers nor the ability of our participants to learn these techniques. It is possible that more implicit probes of internal representations will help here, as recent evidence suggests that the more explicit judgments may underestimate the level of detail in the internal statistical representations (Chetverikov et al., 2016, 2017).

This study has several limitations. One was our reluctance to disentangle the number of test elements from their mean size. It is well known that the precision of discrimination decreases with the length of the judged spatial interval: the JND increases in proportion with the judged length (Allik et al., 2013; Burbeck & Hadden, 1993; Wolfe, 1923). Thus, based on Weber's law alone, it is expected that the precision of discrimination increases with the decrease of the judged length (Allik et al., 2013). In order to cope with this deficiency, we attempted to make the mean diameter judgement task maximally similar to the cumulative area and the total length task. Results showed that the difference between these three tasks was rather obvious: The visual system deals relatively well with averages, but poorly with sums.

Another limitation is failing to test the Kahneman's conjecture more directly. If means were processed by a fast and sums by a slow system, as Kahneman proposed, then it would have been informative to compare processing times. Unfortunately, our instructions did not impose time limits. Participants were free to choose a pace most comfortable for them. For this reason, there was no direct link between the response and processing times. Indeed, for some participants, judging the mean size was an easy task, which took on average less time than deciding about cumulative area and total length (observers AR and MT). For another group of participants, the response time was practically identical under different instructions (e.g. JA and RN). However, for two participants (KA and MS) it was the mean size, which on average took more time to decide than area or total length. Thus, the response times tell us very little about how fast or slow the processing systems really are. The same uncertainty is about the number of elements. For example, it is tempting to propose that if the assumed process is serial, then processing time has to increase linearly with the number of processed elements. If analysis of the response times revealed anything, then a tendency that the decision times of seven elements are slightly shorter than decisions about two or three elements. In other words, based on response times alone, it is impossible to say anything specific on how fast or slow the processing system really is.

Finally, we have a practical recommendation for those who would like to use bubble charts. Although human observers can estimate the area of bubbles or discs, as we saw, it cannot be done accurately. Unlike popular wisdom, the human visual system is not predisposed to experience bubbles or discs in terms of their area. Thus, choosing bubble charts proportional to area could be misleading. It seems that discs and circles are experienced closer to their radius or diameter, for which the visual system is better adapted.

Acknowledgements Aire Raidvee was supported by the Mobilitas Pluss Returning Researcher Grant (MOBTP91) by Estonian Research Council. We thank Jeremy Wolfe, Nancy Kanwisher, and anonymous reviewers for critical comments and suggestions. We are extremely grateful to Mait Samuel and Richard Naar for their help.

References

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research, 83*, 25–39. doi:https://doi.org/10.1016/j.visres.2013.02.018

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2014). Obligatory averaging in mean size perception. *Vision Research, 101*, 34–40. doi:https://doi.org/10.1016/j.visres.2014.05.003

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131. doi:https://doi.org/10.1016/j.tics.2011.01.003

Anderson, N. H., & Weiss, D. J. (1971). Test of a multiplying model for estimated area of rectangles. *The American Journal of Psychology, 84*(4), 543–548. doi:https://doi.org/10.2307/1421171

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162. doi:https://doi.org/10.1111/1467-9280.00327

Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & Psychophysics, 70*(7), 1325–1326. doi:https://doi.org/10.3758/pp.70.7.1325

Banno, H., & Saiki, J. (2012). Calculation of the mean circle size does not circumvent the bottleneck of crowding. *Journal of Vision, 12*(11). doi:https://doi.org/10.1167/12.11.13

Birnbaum, M. H., & Veit, C. T. (1974). Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks. *Perception & Psychophysics, 15*, 7–15.

Burbeck, C. A., & Hadden, S. (1993). Scaled position integration areas: Accounting for Weber law for separation. *Journal of the Optical Society of America A—Optics Image Science and Vision, 10*(1), 5–15. doi:https://doi.org/10.1364/josaa.10.000005

Chetverikov, A., Campana, G., & Kristjánsson, Á. (2016). Set size manipulations reveal the boundary conditions of perceptual ensemble learning. *Vision Research, 140*, 144–156.

Chetverikov, A., Campana, G., & Kristjánsson, Á. (2017). Rapid learning of visual ensembles. *Journal of Vision, 17*(2), 21, 1–15.

Chong, S. C., Joo, S. J., Emmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics, 70*(7), 1327–1334. doi:https://doi.org/10.3758/pp.70.7.1327

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404. doi:https://doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900. doi:https://doi.org/10.1016/j.visres.2004.10.004

Cleveland, W. S., Harris, C. S., & McGill, R. (1982). Judgments of circle sizes on statistical maps. *Journal of the American Statistical Association, 77*(379), 541–547 .

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition, 20*(2), 211–231. doi:https://doi.org/10.1080/13506285.2012.657261

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A—Optics Image Science and Vision, 18*(5), 1016–1026.

Ekman, G., & Junge, K. (1961). Psychophysical relations in visual perception of length, area and volume. *Scandinavian Journal of*

Psychology, 2(1), 1–10. doi:https://doi.org/10.1111/j.1467-9450.1961.tb01215.x

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics, 70*(3), 456–464. doi:https://doi.org/10.3738/pp.70.3.456

Friendly, M. (2008). The golden age of statistical graphics. *Statistical Science, 23*(4), 502–535. doi:https://doi.org/10.1214/08-sts268

Hartley, A. A. (1981). Mental measurement of line length: The role of the standard. *Journal of Experimental Psychology: Human Perception and Performance, 7*(2), 309–317. doi:https://doi.org/10.1037/0096-1523.7.2.309

Im, H. Y., & Chong, S. C. (2014). Mean size as a unit of visual working memory. *Perception, 43*(7), 663–676. doi:https://doi.org/10.1068/p7719

Kahneman, D. (2011). *Thinking fast and slow*. New York, NY: Farrar, Straus and Giroux.

Lee, H., Baek, J., & Chong, S. C. (2016). Perceived magnitude of visual displays: Area, numerosity, and mean size. *Journal of Vision, 16*(3), 12–12. doi:https://doi.org/10.1167/16.3.12

Legge, G. E., Gu, Y., & Luebker, A. (1989). Efficiency of graphical perception. *Perception & Psychophysics, 46*(4), 365–374. doi:https://doi.org/10.3758/bf03204990

Li, J., Martens, J.-B., van Wijk, J. J., & ACM. (2010). A model of symbol size discrimination in scatterplots. *Proceedings of the 28th Annual Chi Conference on Human Factors in Computing Systems* (pp. 2553–2562). doi:10.1145/1753326.1753714

Luce, R. D. (2012). Torgerson's conjecture and Luce's magnitude production representation imply an empirically false property. *Journal of Mathematical Psychology, 56*(3), 176–178. doi:https://doi.org/10.1016/j.jmp.2012.02.002

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica, 142*(2), 245–250. doi:https://doi.org/10.1016/j.actpsy.2012.11.002

Masin, S. C. (2013). On the ability to directly evaluate sensory ratios. *Attention, Perception, & Psychophysics, 75*(1), 194–204. doi:https://doi.org/10.3758/s13414-012-0382-0

Miller, A. L., & Sheldon, R. (1969). Magnitude estimation of average length and average inclination. *Journal of Experimental Psychology, 81*, 16–21. doi:https://doi.org/10.1037/h0027430

Morgan, M. J. (2005). The visual computation of 2-D area by human observers. *Vision Research, 45*(19), 2564–2570. doi:https://doi.org/10.1016/j.visres.2005.04.004

Morgan, M. J., & Glennerster, A. (1991). Efficiency of locating centers of dot-clusters by human observers. *Vision Research, 31*(12), 2075–2083. doi:https://doi.org/10.1016/0042-6989(91)90165-2

Morgan, M. J., Hole, G. J., & Glennerster, A. (1990). Biases and sensitivities in geometrical illusions. *Vision Research, 30*(11), 1793–1810. doi:https://doi.org/10.1016/0042-6989(90)90160-m

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70*(5), 772–788. doi:https://doi.org/10.3758/pp.70.5.772

Nachmias, J. (2011). Shape and size discrimination compared. *Vision Research, 51*(4), 400–407. doi:https://doi.org/10.1016/j.visres.2010.12.007

Narens, L. (2006). Symmetry, direct measurement, and Torgerson's conjecture. *Journal of Mathematical Psychology, 50*(3), 290–301. doi:https://doi.org/10.1016/j.jmp.2005.12.007

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29–46.

Runeson, S. (1977). On the possibility of "smart" perceptual mechanisms. *Scandinavian Journal of Psychology, 18*, 172–179.

Schneider, B., & Bissett, R. (1988). 'Ratio' and 'difference' judgments for length, area, and volume: Are there two classes of sensory continua? *Journal of Experimental Psychology: Human Perception and Performance, 14*(3), 503–512. doi:https://doi.org/10.1037/0096-1523.14.3.503

Seizova-Cajic, T., & Gillam, B. (2006). Biases in judgments of separation and orientation of elements belonging to different clusters. *Vision Research, 46*(16), 2525–2534. doi:https://doi.org/10.1016/j.visres.2006.02.010

Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics, 70*(7), 1335–1336. doi:https://doi.org/10.3758/pp.70.7.1335

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision, 11*(12), 13, 11–11. doi:https://doi.org/10.1167/11.12.13

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York, NY: Wiley.

Stevens, S. S., & Guirao, M. (1963). Subjective scaling of length and area and the matching of length to loudness and brightness. *Journal of Experimental Psychology, 66*(2), 177–186. doi:https://doi.org/10.1037/h0044984

Szafir, D. A., Haroz, S., Gleicher, M., & Franconeri, S. (2016). Four types of ensemble coding in data visualizations. *Journal of Vision, 16*(5), 11–11. doi:https://doi.org/10.1167/16.5.11

Teghtsoonian, M. (1965). The judgment of size. *American Journal of Psychology, 78*, 392–402. doi:https://doi.org/10.2307/1420573

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: Wiley.

Torgerson, W. S. (1961). Distances and ratios in psychophysical scaling. *Acta Psychologica, 19*, 201–205.

Treisman, A. M. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology Section A, 40*(2), 201–237. doi:https://doi.org/10.1080/02724988843000104

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136. doi:https://doi.org/10.1016/0010-0285(80)90005-5

Utochkin, I. S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. *Journal of Vision, 15*(4), 1–14. doi:https://doi.org/10.1167/15.4.8

Westheimer, G. (2008). Illusions in the spatial sense of the eye: Geometrical-optical illusions and the neural representation of space. *Vision Research, 48*(20), 2128–2142. doi:https://doi.org/10.1016/j.visres.2008.05.016

Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review, 18*(3), 484–489. doi:https://doi.org/10.3758/s13423-011-0071-3

Whitney, D., Haberman, J., & Sweeney, T. D. (2014). From textures to crowds: Multiple levels of summary statistics perception. In J. S. Werner & L. M. Chalupa (Eds.), *The new visual neuroscience* (pp. 695–710). Cambridge, MA: MIT Press.

Whitney, D., & Leib, A. Y. (2018). Ensemble Perception. *Annual Review of Psychology, 69*, 105-129. doi:https://doi.org/10.1146/annurev-psych-010416-044232

Wolfe, H. K. (1923). On the estimation of the middle of lines *American Journal of Psychology, 34*, 313–358. Retrieved from www.jstor.org/stable/1413954

Open practices statements The data and materials for all experiments are available. None of the experiments was preregistered.