The numerosity and mean size of multiple objects are perceived independently and in parallel

Igor S. Utochkin*, Konstantin O. Vostrikov

National Research University Higher School of Economics, Moscow, Russia

* isutochkin@inbox.ru

Abstract

It is well documented that people are good at the rapid representation of multiple objects in the form of ensemble summary statistics of different types (numerosity, the average feature, the variance of features, etc.). However, there is not enough clarity regarding the links between statistical domains. The relations between different-type summaries (numerosity and the mean) are of particular interest, since they can shed light on (1) a very general functional organization of ensemble processing and (2) mechanisms of statistical computations (whether averaging takes into account numerical information, as in regular statistics). Here, we show no correlation between the precision of estimated numerosity and that of the estimated mean. We also found that people are very good at dividing attention between numerosity and the mean size of a single set (Experiment 1); however, they show some cost of dividing attention between two same-type (two numerosities or two mean sizes, Experiment 2) and two different-type (one numerosity and one mean size, Experiment 3) summaries when each summary is ascribed to a different set. These results support the idea of domain specificity of numerosity and mean size perception, which also implies that, unlike regular statistics, computing the mean does not require numerosity information. We also conclude that computational capacity of ensemble statistics is more limited by encoding several ensembles than computing several summaries.

Introduction

Numerous studies have shown that the visual system is very efficient at rapid representation of multiple objects in the form of *ensemble summary statistics* such as numerosity [1–5], the mean along a dimension–size ([6,7], etc.), orientation [8–10], color [11,12], speed [13], brightness [14], even emotional expression and gender [15,16], as well as variance [17–19]. There is evidence that the summary statistics are encoded directly as visual properties rather than inferred by any means, as they are prone to adaptation aftereffects [1,18,20]. The ensemble summaries are computable when conscious access to individual objects is limited [6,10,21,22]. Neurophysiological data also support the notion that the representation of ensembles as summaries is not based on the properties of individual objects [23,24].

Ensemble summary statistics give us some evident advantages. One such advantage is that they serve as better estimates of the set in general than each single item does [25]. Another advantage is that the summaries are efficient means for organizing multiple items into chunks [26–29] and for surviving the severe limitations of the processing bottleneck. Although it is a subject of debate whether extracting summary statistics is based on a genuine parallel representation of all presented items at one time [6,25,30–36] or on limited capacity sampling of just a few objects [37–41], it is generally admitted that it leads to fairly efficient approximation of multiple properties in the entire display in terms of both precision [6,25,32,42,43] and speed [7,34,44].

The functional structure of ensemble statistics: Correlation and parallelism

Many of the previously reported studies tested the properties and mechanisms of ensemble perception within separate domains (e.g., average size, average orientation, and numerosity). Together, they showed robustness in core phenomenology–an ability to judge generalized statistical features at a brief glance with relatively good accuracy. This finding naturally leads to the question of whether these local domains are linked together. One possibility is that there is high processing specialization between domains (for example, retinotopic feature maps, which are supposed to be rather independent in the visual cortex, or *modular* [45]), so computing statistical summaries is a part of processing within each module. Alternatively, it is possible that there is a locus within the cognitive system where different domains overlap, and some shared factors affect the resulting ensemble perception.

One approach to probing the functional relations between the domains is measuring *correlations* between tasks across observers [46,47]. The logic of measuring correlations is straightforward: if the tasks are correlated, then the domains addressed by these tasks have some overlap; if they are not correlated, they are likely separate and not supported by any common mechanism. In their recent study, Haberman, Brady, and Alvarez [48] implemented this approach in the field of ensemble perception. Specifically, they were interested in whether there are any common mechanisms/factors underlying ensemble averaging across a variety of visual domains. The authors tested the precision of reporting the average color; average orientation in two types of objects (Gabor patches and isosceles triangles); average emotional expression and average identity of human faces. In each experiment, Haberman et al. [48] ran a pair of tasks using the same participants. In total, eight experiments allowed the estimation of pairwise correlations between the majority of the tested summaries. Haberman et al. [48] found high correlations between the 'low-level' summaries (color and orientations), as well as between the 'high-level' summaries (emotional expression and face identity). However, cross-level correlations were low: The precision of color or orientation averaging did not correlate with that of emotion or face identity. Moreover, these cross-level correlations did not exceed a correlation between a summary statistics task and a verbal working memory task requiring no ensemble representation. This result led Haberman et al. [48] to conclude that there is no single domain-general averaging mechanism that could analyze information from all visual modules. At the same time, their results show some narrower generality spreading within the levels of the visual hierarchy (for example, for computing mean color and mean orientation). However, the authors admit that the conclusion drawn from the latter result is ambiguous. On one hand, this result can indicate the existence of a common computational statistical mechanism within each level. On the other hand, the correlations can be explained by some shared source of noise arising between features lying closer in the representational space (e.g., color and orientation) but not between remote features (e.g., orientation and facial expression).

Apart from testing correlations between tasks, there is another useful approach that is relevant for studying the functional organization. We will refer to it as testing *parallelism*. While correlational studies concern the overlap between different domains, parallelism studies raise a rather different question. Whether two domains overlap or not, it is important to know how they are coordinated in gaining access to conscious perception. For example, if we look at cars on a road trying to estimate their mean speed and mean size at one time, how would the information from the two domains be processed? Will we observe the same efficiency as reported in the numerous ensemble studies within separate domains–unlimited capacity, or parallel processing? Or, will we observe decrement in judging one or both summaries, suggesting interference between the domains? In terms of the functional organization, parallelism addresses the question whether a *limited-capacity bottleneck* is present anywhere in the system and makes various domains compete to be fully implemented.

Studies on task parallelism for probing functional organization have a long history in the field of attention [49–52]. As the role of attention in ensemble perception is a subject of ongoing discussion [8,21,25,53], a number of studies have also addressed the existence of the limited-capacity bottleneck in processing several summaries. Some of these studies have focused on an ability to process more than one summary along a single dimension (for example, the mean sizes of two color sets of dots). Chong and Treisman [54] have shown that it is possible to estimate the average sizes of two ensembles in parallel as accurately as the average size of only one ensemble: participants in their experiment were equally good at judging the average size of a relevant color set when the color was both precued (attention is paid to a relevant set) and postcued (attention is divided between the sets). Later, other studies have shown that no more than two [28] mean sizes can be stored in working memory and that probably only one mean size or mean orientation can be perfectly attended to at one time [30,31]. Halberda et al. [5] have also shown the two-set limit for parallel processing of several numerosities (plus superset numerosity that is always available, regardless of the number of sets). Note that, based on these results, it is difficult to say whether observed limitations in parallel computation reflect the bottleneck in multiple statistical computations or in the division of attention between different sets. From the object perception literature, it is known that it is harder to attend to different objects than to different features of one object [55].

More important for the issue of functional organization are the studies where division of attention between two different summaries are tested. Emmanouil and Treisman [56] showed that dividing attention between the mean size and the mean speed or between the mean size and the mean orientation yields a cost in performance that is not found in the division of attention between two mean sizes. Moreover, the cost is even more dramatic if the mean size is to be estimated in one set and the mean speed (or orientation) is to be estimated in another set. Huang [57] showed that a cost of dividing attention between the orientation summary (prevalence of vertical items over horizontal ones or vice versa) of one ensemble and the color summary (prevalence of reddish items over greenish ones or vice versa) of another is comparable with the cost of dividing attention between features of different objects. An idea that the observed limits of parallel statistical processing has more to do with limits in encoding several sets than statistics per se is supported by other researchers [58,59]. This concept complicates conclusions about the single vs. multiple processors. We will address these complications in our experiments.

Numerosity and the mean

In previous section, we reviewed the studies probing the functional links between ensemble summaries of the same type, mostly two means (or prevalence [57], which can be considered

as just an alternative measure of the central tendency). However, for a more comprehensive view of the architecture underlying ensemble summary statistics, it is necessary to address the link between summaries of different types, e.g., numerosity and the mean.

Although Haberman et al. [48] demonstrated that even two summaries of the same type can be computed independently (e.g., mean orientation and mean facial expression), it does not necessarily imply that two summaries of different types are also independent. This doubt comes from regular mathematical statistics. You can measure two or more attributes in a sample and calculate mean values from each of them, so that each such calculation will be independent from others (e.g., you do not need to know the mean size to calculate the mean speed). However, you cannot calculate the mean without knowing the number of observations (numerosity). Similarly, you cannot calculate the variance without knowing the mean and the number. Therefore, according to regular statistics, summaries of the same type can be calculated independently (which is also the case in ensemble perception [48]), but summaries of different types cannot. Is that also true for ensemble summary statistics? It is easy to see that answering this question can not only shed light not only on between-domain links, it can also inform us about the nature of computations in the core of statistical representations in vision (although we admit that this is an indirect way to study these computations).

Recently, Lee, Baek, and Chong [60] addressed the link between these different types of summaries. Their participants had to estimate the average size, the number, or the total area of a set of circles. The authors then estimated how well each of the parameters could be predicted by the remaining two parameters. They found that the total area is predicted by both the mean size and numerosity much better than other summaries are predicted by any of the rest. This finding led the authors to conclude that the mean size and numerosity can be processed independently, and the total area is computed as the mean multiplied by numerosity.

The study by Lee et al. [60] is an example of the correlational approach to the functional organization of ensemble perception. However, the authors tested the perception of numerosity, mean, and total area in separate blocks, so their paradigm required participants to pay attention to one summary at a time. In our opinion, however, parallelism should also be tested for a more complete view of the visual statistical system at work. Given the independence between numerosity and the mean size obtained in the separate tests [60], we can ask how these two systems are coordinated when attempting to extract different summaries from the same stimulus. Is there a processing bottleneck limiting their concurrent efficiency? Or, in contrast, can they work in parallel, providing lossless computation of both numerosity and the mean? We will address this issue in our experiments.

Our study

As noted above, Lee and colleagues [60] made the first step to shape the functional relationships between the mean and numerosity. Their findings are based on the correlational design. Here, we combine the correlational approach with that aimed at testing parallelism. It is based on the precue-postcue paradigm. Earlier, Emmanouil and Treisman [56] and Huang [57] used this paradigm to test whether two same-type summaries (two means or two prevalence statistics in different sensory domains) can be estimated together without loss in performance. We use this paradigm to test coordination between two different-type summaries. Combining both tests–correlation and parallelism–in one study has some important advantages. First, it allows deeper probing of the functional structure of ensemble summary statistics, because two different aspects of that structure are tested rather than one. Second, measuring each summary under two different conditions–precue and postcue–provides additional inputs for evaluating in the correlational matrix. Instead of looking at one correlation between two separately

measured rates (as in typical correlational studies, e.g., [46, 48, 60]), we can look at precue-postcue correlations within each summary and compare two between-summary correlations observed under different cue conditions. The precue-postcue correlations are informative because they can show whether our tests are reliable and whether the tested abilities are robust across tasks. In turn, comparing the between-summary correlations can show whether the whole tested structure is reliable and stable. In other words, it can show whether an attempt to process two summaries at once changes the functional links between these two domains.

Based on this approach, we conducted three experiments testing correlation and parallelism in the perception of numerosity and mean size. In Experiment 1, we tested how these two summaries are computed within a single set. Experiments 2 and 3 aimed to compare the efficiency of concurrent computation of two of the same summaries (mean-mean or numerosity-numerosity) with the efficiency of concurrent computation of two different summaries (mean-numerosity). Since it is impossible to assign two mean sizes or two numerosities to the same set, in both Experiment 2 and 3, we used two sets with only one relevant parameter assigned to each set. Together with Experiment 1, this allowed dissociation of the cost caused by attending to two summaries from that caused by attending to two sets.

Experiment 1

Method

Participants. For determining the sample size, we used the statistical tool G*Power 3.0.10 [61]. We set required statistical power at .8, a Type I error at .05, and an expected Cohen's d at .6. The expected effect size was set based on the previous research of cue effects on numerosity and mean size perception [5,56,59]. These values led us to a minimum sample size of nineteen participants. Considering a possibility of technical problems or poor performance in some participants, we recruited twenty-three participants. Psychology students of the Higher School of Economics (HSE) took part in the experiment for extra course credit (14 females, ages ranging between 18 and 22 years). All participants reported having normal or corrected-to-normal vision and no neurological problems. The participants were unaware of the purpose of the experiment. Before beginning the experiment, they signed an informed consent form for the procedure, which had been approved by the Institutional Review Board at HSE.

Apparatus and stimuli. Stimulation was developed and presented through PsychoPy [62] for Linux. Stimuli were presented on a standard VGA monitor at a refresh frequency of 75 Hz with 800 × 600-pixel spatial resolution. A homogeneous gray field subtending approximately 19.44˚ × 19.44˚ was used for stimulus presentation. The rest of the screen space was black and not used for presentation. The "working" gray field was divided into 6 × 6 imaginary cells, each having a side length of 3.24˚. These imaginary cells served for item positioning, as described in the next paragraph.

Filled white circles with different diameters were used as items in a sample set. Each circle could be located within one of the 36 imaginary cells, and each cell could be occupied by only one circle or left empty. Each circle could be placed at a random position within a cell with a restriction that its center be located at least one radius + .16˚ away from any cell border along cardinal axes. This restriction meant that no circle ever overlapped with another. On the other hand, sets looked rather random in terms of spatial arrangement.

The number of circles could vary between 7 and 36; random cells were chosen on each trial for placing the circles. The diameters of the circles were randomly drawn from the following list: .86˚, 1.03˚, 1.19˚, 1.35˚, 1.51˚, 1.67˚, and 1.84˚. The mean diameter of the circles in each drawn display ranged between trials from 1.08˚ to 1.67˚. The standard deviation of the diameter in each display ranged from .14˚ to .21˚.

For precue, the Russian equivalents of either the word "NUMBER" or "MEAN" were used. The words were printed in white Arial (letter height 1.6˚) and located in the center of the screen. For postcue, six question marks were presented instead of the words. For numerosity reports, a white cursor appeared below the gray "working" field to type in a number. For mean size reports, a white adjustable probe circle was presented at the center of the screen. The starting size of the probe varied randomly between .43˚ and 2.21˚.

Procedure. Experimental sessions were run in a darkened room. Participants were seated approximately 50 cm from a monitor. They were instructed to estimate and report the number or/and the mean size of a set of briefly presented circles.

Each trial started with a presentation of a cue screen for 500 ms that also served for gaze fixation on the center of a screen. In the precue trials, a word, either "NUMBER" or "AVERAGE," appeared saying which of the summaries were to be reported in the end of the trial (Fig 1). In the postcue trials, six question marks appeared warning that the relevant parameter would be known after sample presentation (Fig 1). The cue screen was followed by a 500-ms blank screen that was, in turn, followed by the presentation of a sample set for another 500 ms. Immediately after the sample, a test screen was presented. The participants had to type in the number of circles in the sample [5] or adjust a probe circle to match the average size of the circles in the sample [38]. The appearance of the probe screen (a cursor for typing or the probe circle) also informed the participants about their task in the postcue trials (Fig 1). For typing the numbers, a numerical pad of a standard keyboard was used (arrow, "DEL", and "BACK-SPACE" keys could be used for editing an entry). To increase or decrease the size of the probe pixel by pixel, right and left arrow keys were used. The time for response was not limited. To confirm their response and quit the trial, participants had to press ENTER. The next trial started upon pressing SPACE, so participants could progress at a comfortable pace and take a rest whenever they wanted.

Design and data analysis. In this experiment, we used a 2 (Cue–precue vs. postcue) × 2 (Task–numerosity vs. mean size estimation) within-subject design. The number of trials per condition was 60 (two repetitions of 30 displays covering the whole range of numbers from 7 to 36). An experimental block, therefore, consisted of 2 × 2 × 60 = 240 trials presented in a random order. The experimental block was preceded by 8 practice trials that were not included in the data analysis.
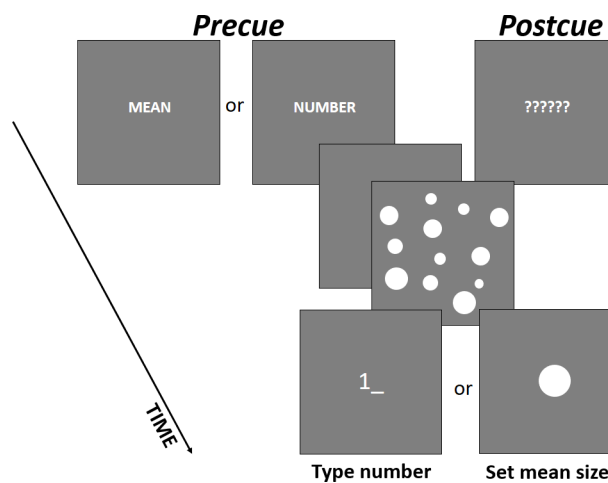

Fig 1. An example of a single trial with time course in Experiment 1. (A) precue and (B) postcue conditions.

For the analysis, a relative error magnitude was calculated on each trial as an absolute deviation of the observer's response from the correct response divided by the correct response. The relative error was considered to characterize the precision of numerosity or mean size judgments.

Two types of analysis were then applied to the relative errors. First, we estimated an effect of Cue on each of the tasks using a standard t-test along with a Bayesian t-test. The latter is considered to be an alternative to classical significance testing [63] and is based on comparing evidence for H_1 against H_0. The H_1-to-H_0 likelihood ratio–the Bayes factor (BF_{10})–is a principal quantitative measure of evidence of one hypothesis over another [64,65]. Bayesian analysis was run using JASP software [66]. The prior effect size distribution was set as a Cauchy with a median of 0 and a width of .707, a default setting recommended when no information about the expected effect size is available in advance [66]. To interpret the Bayes factors, we used Jeffreys' scale [64] with Kass and Raftery's adjustment [65], where $1 < BF_{10} < 3$ is evidence for H_1 that is "not worth more than a bare mention" ([65], p. 777), where $3 < BF_{10} < 20$ is positive evidence for H_1, $20 < BF_{10} < 150$ is strong evidence for H_1, and where $BF_{10} > 150$ is very strong evidence for H1; the inverse ratios define symmetrical evidence ranges for H_0.

Second, a correlation analysis was applied to the data. Averaged relative error values were also calculated for each participant for all four factorial cells. We then estimated correlations between these four data points across participants. There were four correlations of particular interest for us. Correlations between the same type of summaries in different cue conditions ("Numerosity—Precue" with "Numerosity—Postcue", and "Mean size—Precue" with "Mean size—Postcue") were to estimate consistency of the tested abilities across cue conditions. Correlations between different summaries at same-cue conditions ("Numerosity—Precue" with "Mean size—Precue", and "Numerosity—Postcue" with "Mean size—Postcue") were to show if any interrelation exists between numerosity and mean size estimation. Again, two methods were applied for judging whether the correlations exist: (1) the standard significance tests and (2) computing Bayes factors for estimating whether posterior distributions of correlation coefficients more likely have a higher (H_0) or lower (H_1) density around 0 than a homogeneous prior distribution (beta-distribution with a width of 1 [JASP Team]).

Results and discussion

We found no effects of Cue on the relative error (Fig 2A), both in the numerosity task ($t(22)$ = 1.486, p = .151, d = .310, BF_{10} = .573) and in the mean size task ($t(22)$ = 1.540, p = .138, d = .321, BF_{10} = .613). Therefore, Experiment 1 showed no evidence for cost of dividing attention between the two summaries.

Fig 2B and 2C illustrate the informative correlations. We found very high positive precue-postcue correlations within each summary ("Numerosity—Precue" with "Numerosity—Postcue": r = .904, $p < $ .001, BF_{10} = 3.349 × 10^6; "Mean size—Precue" with "Mean size—Postcue": r = .939, $p < $ .001, BF_{10} = 1.566 × 10^8; Fig 2B). These correlations show high consistency between the two conditions in which each summary is measured. From a more general viewpoint, they suggest that the individual differences in both measured abilities–enumeration and averaging– are robust against cue manipulations. In contrast, we found no correlations between different summaries within same-cue conditions ("Numerosity—Precue" with "Mean size—Precue": r = .085, p = .700, BF_{10} = .277; "Numerosity—Postcue" with "Mean size—Postcue": r = .050, p = .820, BF_{10} = .265; Fig 2C). This lack of correlation suggests the absence of a common mechanism for the perception of numerosity and the mean size.

In sum, two results of Experiment 1 are most important in terms of the topic. First, we found no cost of division of attention between the numerosity and the mean size of a set.
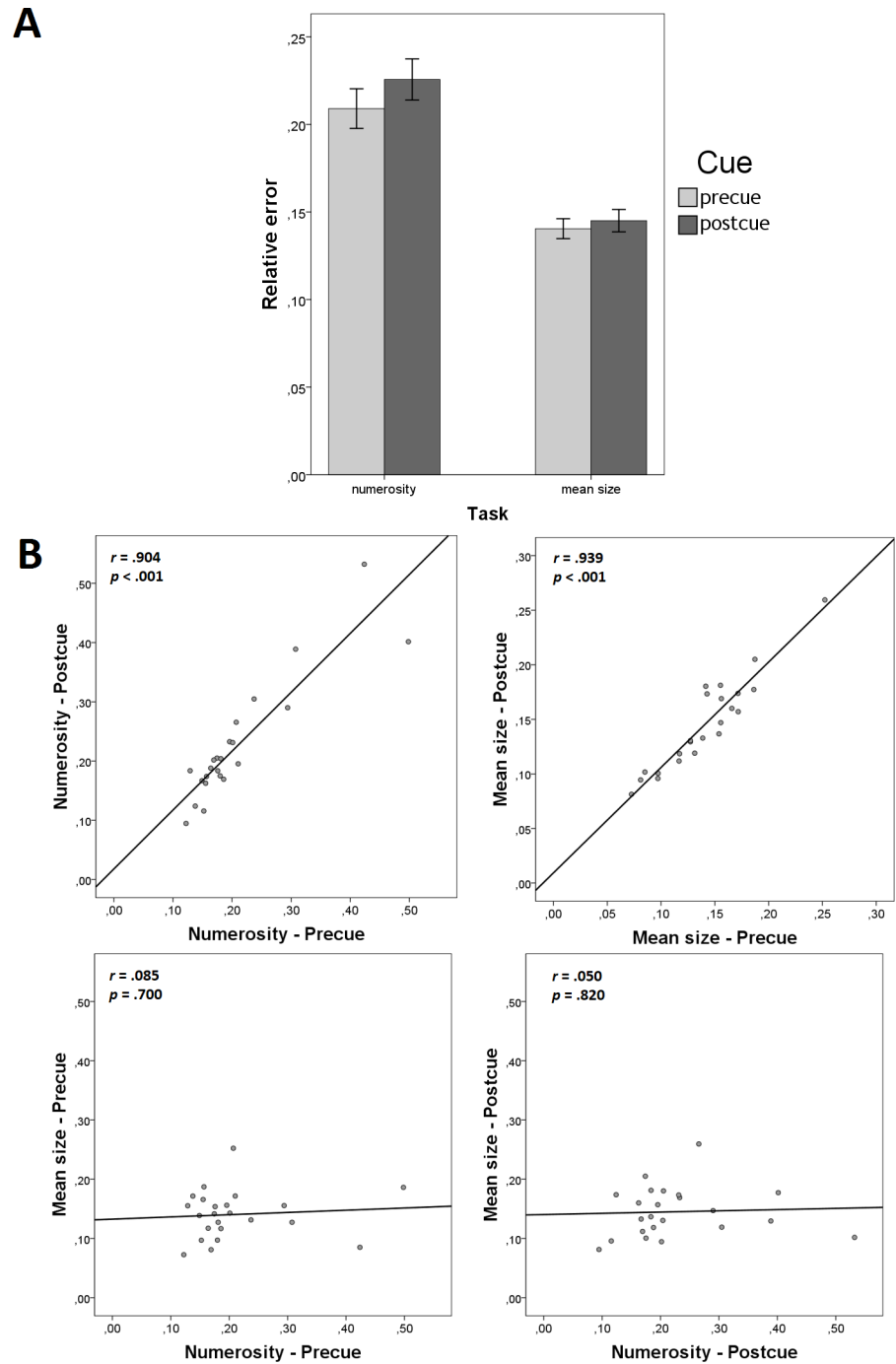
Fig 2. Results of Experiment 1. (A) cue effects on the relative error of numerosity and mean size judgments (error bars denote ±95% CI); and scatterplots showing (B) precue-postcue correlations within summaries and (C) between-summary correlations within different cue conditions.

https://doi.org/10.1371/journal.pone.0185452.g002

Second, we found no correlation between the precision of judging numerosity and the mean. Taken together, these findings suggest that these two tasks can be carried out in parallel and independently of each other, at least when both summaries are estimated within one set of

objects. This is consistent with the idea of non-overlapping well-coordinated mechanisms for computing numerosity and the mean size.

In Experiment 1, observers had to compute two summaries at one time (in the postcue trials). Adopting the idea of the separate computational mechanisms, we assume that each computation is carried independently and share no processing bottleneck. This allows to avoid interference. We can call it *distributed computation*. In the next experiments, we compare this distributed computation (two concurrent computations, each is performed by a separate mechanism) with a situation where a single processor is busy computing two summaries (that we can call *shared computation*). We suppose that shared computation occurs when two same-type summaries are estimated. In Experiment 2, we tested this in a task requiring the distribution of attention between two numerosities or two mean sizes. Obviously, it is impossible to have two same-type summaries within one set, so two sets are always required. As we noted earlier, this causes certain problems in interpretation, because it is hard to dissociate between parallel statistical computations and parallel attending to two sets [56,57]. For this reason, it would be difficult to compare Experiment 2 (same summary in two sets) with Experiment 1 (different summaries in one set). In Experiment 3, we solved this problem by introducing combined conditions, namely, different summaries in different sets.

Experiment 2

Method

Participants.   Nineteen psychology students of the Higher School of Economics took part in the experiment for extra course credit (10 females, ages ranging between 18 and 22 years). All reported having normal or corrected-to-normal vision and no neurological problems. The participants were unaware of the purpose of the experiment. Before beginning the experiment, they signed an informed consent for the procedure approved by the Institutional Review Board at HSE.

Apparatus and stimuli.   Apparatus and stimulation were the same as in Experiment 1, except for a few important differences. Circles in a whole sample set ("the superset") were colored in yellow or blue and intermixed by color in the space of a screen (Fig 3). This led the superset to form two overlapping color subsets. The total number of circles in a set varied from 10 to 36. As can be noted, the lower limit of the total number was higher than in Experiment 1. This was done to ensure that the number of circles within any subset was no fewer than five. We did not use subsets containing under five circles because small sets of objects (1 to ~4) are typically enumerated differently than larger sets [3]. The maximum number of circles in a subset was 23.

The mean size and standard deviation of the circles of the supersets ranged exactly as in Experiment 1. The mean size of each subset differed, on average, by 21% from the mean size of the other subset in a display and by 11% from the overall mean size of all circles. These differences ensured that participants could not reliably report the mean size of a given subset by deriving it from another attended subset or from the entire set. This encouraged the participants to divide their attention between the subsets when it was demanded by the task. As the standard deviations of the supersets were inherited from Experiment 1, keeping the systematic differences between mean sizes in the subsets inevitably reduced the local standard deviations in the subsets. They ranged from .05˚ to .19 (.10˚ on average).

For the precue, the Russian equivalents of either the word "YELLOW" (printed with a yellow font) or "BLUE" (printed with a blue font) were used. For the postcue, six question marks were presented, printed with a neutral white font. The cursor for typing a number and probe circles for adjusting the mean size could also be yellow or blue.
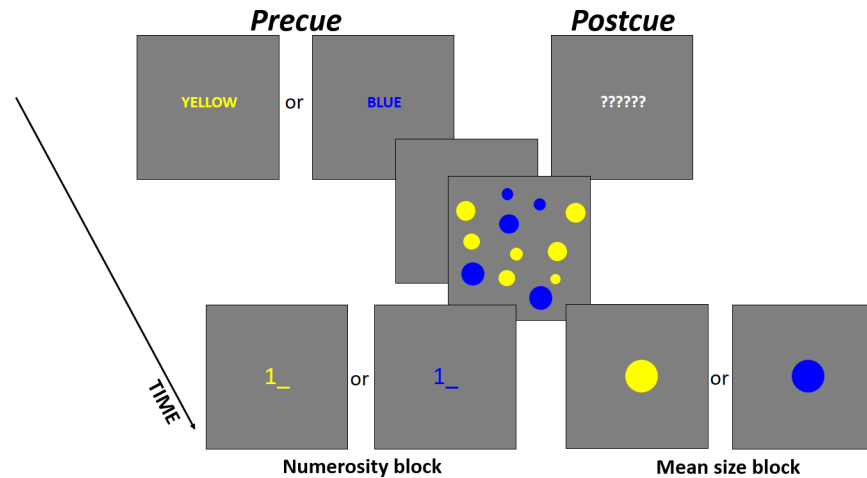
Fig 3. An example of a single trial with time course in Experiment 2. (A) precue and (B) postcue conditions.

Procedure. As in Experiment 1, the participants had to report either the number or the mean size of circles presented in a sample set. The time course of a trial was the same as in Experiment 1. However, there were some important differences. Instead of reporting the relevant summary of the whole set, the participants were instructed to report the summary of a yellow or a blue subset. "Number" trials were presented in a separate block from "mean" trials, so a relevant parameter was always known to the participants. Therefore, their attention was always focused on one summary but could be divided between two subsets.

In the precue condition, a word "YELLOW" or "BLUE" appeared at the beginning of the trial informing the participant about the color of a relevant subset (Fig 3). In the postcue condition, six question marks replaced the cue. 500 ms after the cue, a sample set was presented for 500 ms, followed by the cursor or a probe circle (Fig 3). The cursor or a circle had the color of the relevant subset. That informed the participant about the relevant subset in the postcue trials and served as a reminder in the precue trials.

It is important to note that, in trials where one (the smaller) of the subsets contained fewer than five items, reports were demanded only on a larger subset, both in the numerosity and the mean size task. Nonetheless, the participants were not warned that they would never be asked about the small sets, so they divided their attention between both subsets anyway (unless the relevant subset was precued).

Design and data analysis. In this experiment, we used a 2 (Cue–precue vs. postcue) × 2 (Task–numerosity vs. mean size estimation) within-subject design. The number of trials per condition was 54 (two repetitions of 27 displays covering the whole range of numbers from 10 to 36). The experiment was divided into two blocks, one for the numerosity task and another for the mean size task. The order of block presentation varied across participants. The total number of trials in the two blocks was, therefore, 2 × 2 × 54 = 216 trials presented in a random order within each block (108 trials per block). Each block was preceded by 8 practice trials that were not included in the data analysis.

Data analyses were the same as in Experiment 1.

Results and discussion

We found a strong effect of Cue on the relative error (Fig 4A) in the numerosity task ($t(18) = 7.273$, $p < .001$, $d = 1.669$, $BF_{10} = 19202.171$) and a moderate effect in the mean size task ($t(18) =$
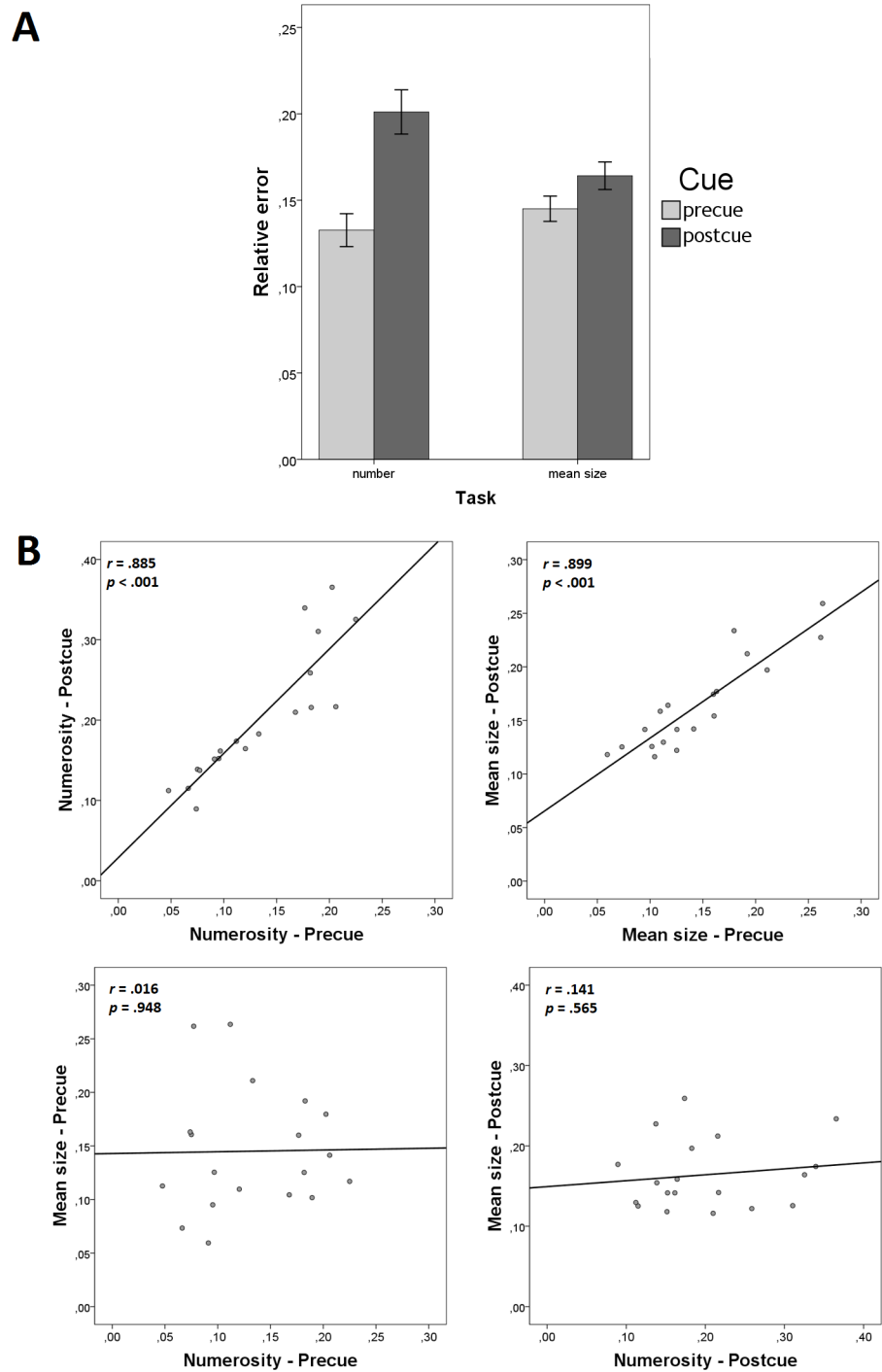
Fig 4. Results of Experiment 2. (A) cue effects on the relative error of numerosity and mean size judgments (error bars denote ±95% CI); and scatterplots showing (B) precue-postcue correlations within summaries and (C) between-summary correlations within different cue conditions.

3.196, $p = .005$, $d = .733$, $BF_{10} = 9.287$). In both cases, the precue led to smaller relative errors compared to the postcue (Fig 4A). We therefore found evidence for the cost of dividing attention between two summaries of the same type when they belong to different sets.

Fig 4B and 4C illustrate the informative correlations. The results were very similar to the results of Experiment 1. Precue-postcue correlations were high within each of the summaries ("Numerosity—Precue" with "Numerosity—Postcue": $r = .885$, $p < .001$, $BF_{10} = 73149$; "Mean size—Precue" with "Mean size—Postcue": $r = .899$, $p < .001$, $BF_{10} = 78566$; Fig 4B) showing high consistency between the measurements across the cue conditions. We also found no evidence for correlations between different summaries within same-cue conditions ("Numerosity —Precue" with "Mean size—Precue": $r = .016$, $p = .938$, $BF_{10} = .284$; "Numerosity—Postcue" with "Mean size—Postcue": $r = .141$, $p = .565$, $BF_{10} = .339$; Fig 4C). These results replicated the absence of correlation that supports the validity of the conclusion from Experiment 1, that there is probably no common mechanism in the processing of numerosity and the mean.

In Experiment 2, the participants did not divide their attention between numerosity and the mean, as each of the tasks was presented as a separate block. Instead, the participants divided their attention between two numerosities or two mean sizes of different subsets. We found a cost of dividing attention under this condition compared to focusing on only one summary. These results somewhat contradict previous work showing that observers are equally good at representing the numerosities [5] and mean sizes ([54], Experiment 3; see also [28]) of two subsets under both the precue and postcue conditions. One difference that could explain the discrepancy between our data and that of Chong and Treisman [54] was the sensitivity of the test. While Chong and Treisman [54] used a two-alternative choice (their participants decided which of two probes had the mean size), our participants performed continuous adjustment, increasing or decreasing the probe in a pixel-by-pixel manner. These continuous estimates could be sufficient to reveal slight but significant differences between the precue and postcue conditions. However, this speculation requires further experiments to be tested. Halberda et al. [5] used a very similar procedure for number estimation to ours and even more difficult presentation conditions (their samples were masked both forward and backward) and found that most of their participants had shown no postcue cost for up to two subsets. However, Poltoratski and Xu ([59], Experiment 1), in their replication of Halberda et al. [5]'s study, found evidence for that cost for two subsets. It appears, therefore, that some differences in stimuli, procedures, or samples of observers can have something to do with the discrepancies in the results. More importantly, most of the mentioned studies explicitly show that the concurrent computation of summary statistics in different sets is prone to severe capacity limitations [28,30,31], which can be probed with the precue [5,59]. It also appears that, in Experiment 2, we found those limitations for both numerosity and the mean when they are tested independently. This finding is a baseline for Experiment 3, where the two summaries are estimated in different subsets concurrently.

Experiment 3

Method

Participants.   Twenty-three psychology students of the Higher School of Economics took part in the experiment for extra course credit (18 females, ages ranging between 18 and 22 years). All reported having normal or corrected-to-normal vision and no neurological problems. The participants were unaware of the purpose of the experiment. Before beginning the experiment, they signed an informed consent for the procedure approved by the Institutional Review Board at HSE.

Apparatus and stimuli.   Apparatus was the same as in the previous experiments. Stimulation was the same as in Experiment 2 concerning sample sets and report screens. Cue screens were the same as in Experiment 1 (words "NUMBER" and "MEAN" as precues, or a string of
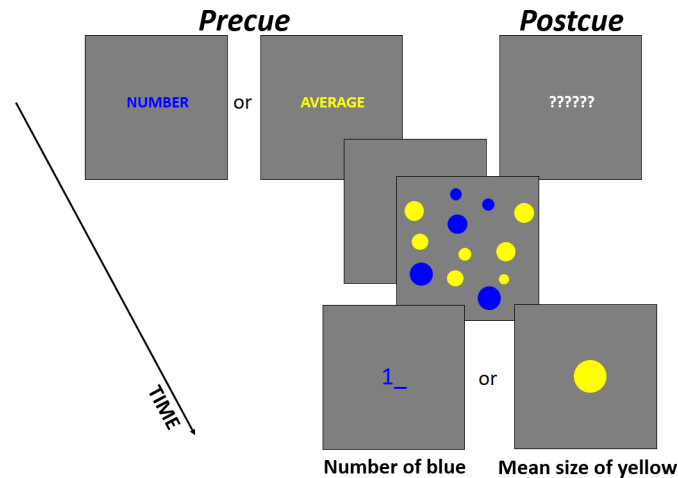
Fig 5. An example of a single trial with time course in Experiment 3. (A) precue and (B) postcue conditions.

question marks as postcues) with an addition that the precue words could be colored in blue or yellow.

Procedure. In Experiment 3, participants were instructed to report the number of one color subset and the mean size of another one. For example, if a relevant subset on a trial was yellow, then the participant had to report its numerosity; if a relevant subset was blue, then the participant had to report its mean size (Fig 5). The participants had, therefore, to encode two different summaries–one from each subset [55]. In precue trials, a word informing about the relevant summary was presented on the cue screen. The precue word was printed in a color corresponding to the color of a subset assigned to the requested summary (Fig 5). In postcue trials, six white question marks appeared on the cue screen. As in the previous experiments, either the cursor or a probe circle appeared at the end of trial. These were also colored in accordance with the relevant subset (Fig 5). The rest details of the procedure were same as in previous experiments.

Design and data analysis. The design and the number of trials per condition were the same as in Experiment 2. There were only two differences. First, numerosity and mean size trials were randomly intermixed rather than blocked. Second, the colors of "numerosity" and "mean" trials varied across participants: One half of the participants had to report the number of yellow and the mean size of blue circles, while another half had to report the number of blue and the mean size of yellow circles. Only one practice block of 8 trials preceded the experimental block.

Data analyses were the same as in the previous two experiments.

Results and discussion

The data of one participant were removed from the analysis as an outlier because of the systematic tendency to overestimate the number of circles by approximately 100% on average. This tendency led to the relative error in the numerosity task far exceeding three standard deviations. Therefore, data from 22 participants were used for further analysis.

We found strong effects of Cue on the relative error (Fig 4A), both in the numerosity task ($t(21) = 4.337$, $p < .001$, $d = .925$, $BF_{10} = 106.3$) and in the mean size task ($t(21) = 4.597$, $p < .001$, $d = .980$, $BF_{10} = 184.4$). In both cases, the precue led to smaller relative errors compared to the postcue (Fig 6A). These results suggest that there was a cost of dividing attention between numerosity and the mean size when these summaries belong to different sets.

Fig 6B and 6C illustrate the informative correlations. Precue-postcue correlations were high within each of the summaries ("Numerosity—Precue" with "Numerosity—Postcue": $r = .670$, $p = .001$, $BF_{10} = 48.163$; "Mean size—Precue" with "Mean size—Postcue": $r = .771$, $p < .001$,
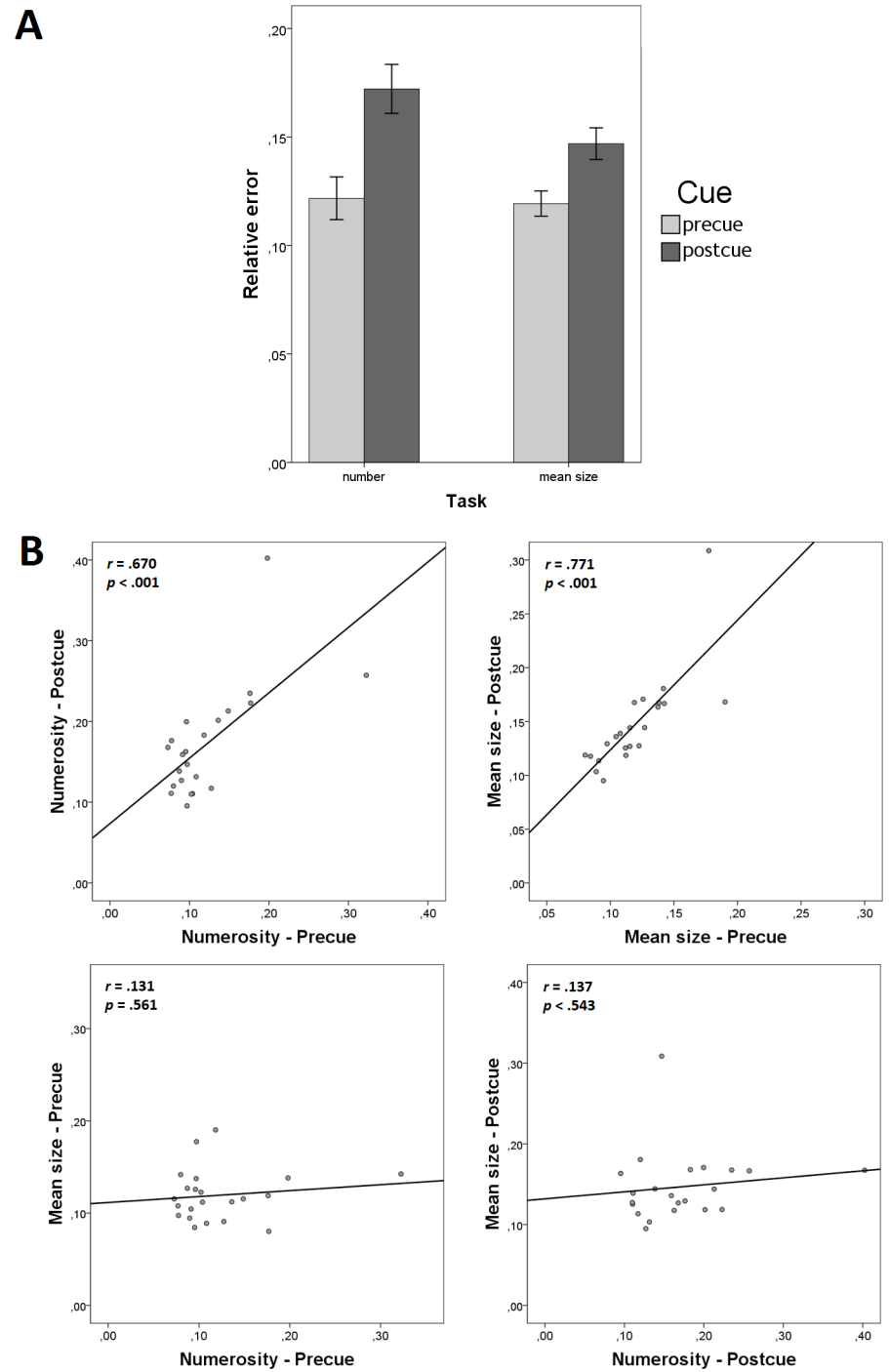
A

B

Fig 6. Results of Experiment 3. (A) cue effects on the relative error of numerosity and mean size judgments (error bars denote ±95% CI); and scatterplots showing (B) precue-postcue correlations within summaries and (C) between-summary correlations within different cue conditions.

BF_{10} = 953.552; Fig 6B) showing high consistency between the measurements across cue conditions. Correlations between different summaries within same cue conditions were non-significant ("Numerosity—Precue" with "Mean size—Precue": r = .131, p = .561, BF_{10} = .320; "Numerosity—Postcue" with "Mean size—Postcue": r = .137, p = .543, BF_{10} = .315; Fig 6C). This pattern replicates the patterns from Experiments 1 and 2, suggesting the absence of a common mechanism in the processing of numerosity and the mean. We conclude that these two summaries are processed independently.

Like in Experiment 1, our participants had to divide their attention between numerosity and the mean size on postcue trials. Like in Experiment 2, they also divided their attention between subsets when estimating those statistics. We found evidence for some cost of divided attention that we did not observe in Experiment 1. However, a similar cost was found in Experiment 2 when only one of the "processors" was always enabled on each block of trials. Therefore, we can conclude that the postcue cost in Experiment 3 was likely to be caused by dividing attention between two subsets rather than two summary statistics.

General discussion

Summarizing the results from Experiments 1–3, we came to three important findings. First, our observers had no difficulties in dividing their attention between the numerosity and the mean size of the same set (Experiment 1). Second, there was no correlation between the precision of numerosity and mean judgments, and this result was highly consistent across all experiments and conditions. Third, the division of attention between same-type (Experiment 2) or different-type (Experiment 3) summaries yielded some costs if these summaries belonged to different sets. We discuss the theoretical implications of these findings below.

Independence and parallelism: Implications for visual statistical processing

In our experiments, we combined a correlational approach [48,60] with the precue-postcue paradigm [5,54,55,57] in order to test correlation and parallelism between two domains of ensemble summary statistics–numerosity and mean size. As we noted above, both these tests focus on different aspects of the functional organization of ensemble perception. Without testing parallelism, the observed uncorrelatedness (independence) only shows the absence of a potential overlap between different domains, but not their coordination in conscious perception when multiple statistics are required for a rich description of a stimulus. Likewise, without testing correlation, it is hard to say whether the observed parallel statistical computations are provided by truly independent domains or by a unitary algorithm that comprises all computations (e.g., computing the average using numerosity information). Experiment 1 was the main and most straightforward test for both correlation and parallelism, given that only one ensemble has always been processed but two different statistical summaries have been extracted. The results of this experiment showed consistency with an idea of two well-separated domains for numerosity and the mean size, as we found no correlation between these tasks and no signs of interference.

In Experiments 2 and 3, we tested the tasks in which the visual system was busy with two summaries of different types, with the tasks in which it was busy with two same-type summaries. This testing scheme allowed us to compare the supposed distributed computation (when coordination between the domains is required) with shared computation (when no coordination is required because only one processor is always used). We also addressed the controversial issue of dividing attention between summaries vs. sets. It turned out that both distributed and shared computations showed the cost of divided attention. The whole triplet of the

experiments eventually provided us with important information about the capacity limitations of ensemble summary statistics. Comparing Experiments 1 and 3, both requiring distributed computation but differing in the number of attended sets, we can see that it is attending to two sets that causes interference. Therefore, when we turn to Experiment 2, requiring the shared computation of two same-type statistics in two sets, we should conclude that there are probably no specific constraints in computational capacities per se, even when they share a single mechanism. This conclusion is supported by the high precue-postcue correlations in same-type summaries in Experiment 2: These correlations show that, if ensemble summary is estimated precisely in one ensemble, it is also estimated precisely in two ensembles, yet with some loss in both. In line with other previous studies [57–59], we conclude that capacity limitations of ensemble summary statistics are associated with the limits in parallel encoding of multiple sets. Beyond this limitation, computational capacity for estimating summary statistics can be rather large, both in distributed and shared computations.

An idea of a single "statistical processor" performing numerous ensemble computations looks intuitively appealing when we borrow terms and concepts from a different discipline–regular mathematical statistics–to describe a part of visual reality. In addition, it seems even more plausible given interesting data in favor of a possible connection between some ensemble statistics (numerosity) and more generic and not purely visual mathematical abilities [67]. However, our results, along with the previously reported data [48,60], go against the idea of such a general processor. Perhaps some overlap (correlation or interference) between ensemble summaries is possible, but so far it has been documented only between a few combinations of basic features–mean size and mean speed, mean size and mean orientation [55], mean orientation and mean color [48]. For more "remote" summaries, such as numerosity and the mean size, there is no evidence for such an overlap.

The parallel and independent character of numerosity and mean size processing leads us to an idea of a more modular (domain-specific) architecture [68] underlying a variety of phenomena referred to as ensemble summary statistics. Modularity implies that a number of highly domain-specific structures work on various ensemble properties, each within its domain specificity, and the modules do not overlap. Our claim that the computation of numerosity and the mean size are provided by separate domain-specific "modules" is supported by neural data. It is shown that numerical functions, including numerosity estimation in multiple objects, are associated with increased activity in the parietal cortex, especially the intraparietal sulcus [69,70], and activity in the prefrontal cortex [70]. On the other hand, Cant and Xu [23] have recently shown that the processing of shape statistics in photorealistic multiple objects is associated with increased activity in the parahippocampal place area and lateral occipital area, known to respond to intrinsic properties of objects and scenes.

How can ensemble summary statistics be calculated?

The second important issue associated with the links between the mean and numerosity is the question whether, like in regular statistics, calculating the mean requires information about the number (as the number of items in a sample is the denominator in the formula to calculating the mean). Applying this idea to our tasks, it could be supposed that, for size averaging, the visual system roughly estimates individual sizes and the number of circles, then sums the sizes, and divides by the number. However, the data from all three of our experiments show that this scenario is unlikely. If two processors, as shown, are independent, then they hardly "know" what calculations the other is doing and what result is being reached. This concept is supported by the absence of significant correlations between the precision of number estimation and that of mean size estimation, which is consistent with the conclusion made by Lee et al. [60].

The finding that the visual system does not act like a regular statistician while calculating the average makes us consider alternative mechanisms that could lie within a core of representing ensemble summary statistics. One approach that could suggest a potential explanation is based on the idea of *sampling*, which resonates particularly in studies of size and color averaging [37,38,40,41]. Sampling implies that the mean can be efficiently estimated based on just few items (2 to 4) that occur to be attended to and loaded into working memory. Therefore, if only a sample of items is averaged, then there is no need for the knowledge of the total number of items in a display: The system averages these few items and approximates the estimate for the rest, regardless of their number. However, a number of studies tend to disprove the limited-capacity sampling mechanism of averaging [30–34,36,71]. This also questions the sample-based explanation of the absence of links between the mean size and numerosity.

The second possible idea that potentially can explain how visual averaging bypasses operating numerosity is based on the concept of *pooling*. It refers to the mechanism of accumulating signals by higher levels of the visual hierarchy from its lower levels in the feedforward processing stream. The idea of pooling as a mechanism of representing summary statistics is widespread [10,25,43,72–74], and for ensemble averaging, it is well explained by Haberman and Whitney [42]. At the lower levels of analysis (for example, at the level of V1), every ensemble element is represented by a population response of feature-selective neurons with narrow receptive fields, that is, every such population is very locally tuned to a small piece of the visual field. On higher levels, receptive fields become larger, accumulating information from several lower receptive fields. Therefore, neurons in those large receptive fields can be selectively activated by a combination of lower-level signals. Presumably, the peak activation will be produced by neurons responsive to the average value of the lower-level signals [24]. That process can explain how the mean ensemble feature can be encoded directly as a sensory property [20]–which is of particular importance for our topic–without taking the number of elements into account. Nevertheless, our data reported in this article do not allow us to directly conclude which of the strategies is actually used for "numerosity-free" computation of the mean size. Further experiments are required to clarify this point.

Conclusions

Numerous demonstrations of rapidly computed ensemble summary statistics raise the question of what performs these computations. Is it a single general structure that can perform a broad range of statistical transformations on data from different sensory domains? Our findings described in this article suggest that different statistical estimates–numerosity and the mean size–extracted from exactly the same stimulus are likely to be provided by independent domain-specific mechanisms, rather than by a single domain-general one. This result also implies that, unlike mathematical statistics, visual averaging does not operate information about the number of individual objects which, in turn, questions whether ensemble perception is based on a set of transformations reproducing regular statistics literally.

Supporting information

S1 File. Full raw data of all participants from Experiments 1–3.
(RAR)

Author Contributions

Conceptualization: Igor S. Utochkin.

Data curation: Konstantin O. Vostrikov.

References

1. Burr D, Ross J. A visual sense of number. Curr Biol. 2008; 18: 425–428. https://doi.org/10.1016/j.cub.2008.02.052 PMID: 18342507

2. Chong SC, Evans KK. Distributed versus focused attention (count versus estimate). WIREs Cogn Sci. 2011; 2: 634–638.

3. Feigenson L, Dehaene S, Spelke E. Core systems of number. Trends Cogn Sci. 2004; 8: 307–314. https://doi.org/10.1016/j.tics.2004.05.002 PMID: 15242690

4. Gallistel CR, Gelman R. Non-verbal numerical cognition: From reals to integers. Trends Cogn Sci. 2000; 4: 59–65. PMID: 10652523

5. Halberda J, Sires SF, Feigenson L. Multiple spatially overlapping sets can be enumerated in parallel. Psych Sci. 2006; 17: 572–576.

6. Ariely D. Seeing sets: Representation by statistical properties. Psych Sci. 2001; 12: 157–162.

7. Chong SC, Treisman AM. Representation of statistical properties. Vis Res. 2003; 43: 393–404. PMID: 12535996

8. Alvarez GA, Oliva A. Spatial ensemble statistics are efficient codes that can be represented with reduced attention. Proc Natl Acad of Sci USA. 2009; 106: 7345–7350.

9. Dakin SC, Watt RJ. The computation of orientation statistics from visual texture. Vis Res. 1997; 37: 3181–3192. PMID: 9463699

10. Parkes L, Lund J, Angelucci A, Solomon JA, Morgan MJ. Compulsory averaging of crowded orientation signals in human vision. Nature Neurosci. 2001; 4: 739–744. https://doi.org/10.1038/89532 PMID: 11426231

11. De Gardelle V, Summerfield C. Robust averaging during perceptual judgment. Proc Natl Acad of Sci USA. 2011; 108: 13341–13346.

12. Maule J, Franklin A. Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. J Vis. 2015; 15(4): 6. https://doi.org/10.1167/15.4.6 PMID: 26114595

13. Watamaniuk SNJ, Duchon A. The human visual-system averages speed information. Vis Res. 1992; 32: 931–941. PMID: 1604862

14. Bauer B. Does Stevens's power law for brightness extend to perceptual brightness averaging? Psychol Record. 2009; 59: 171–186.

15. Haberman J, Whitney D. Rapid extraction of mean emotion and gender from sets of faces. Curr Biol. 2008; 17: R751–R753.

16. Haberman J, Whitney D. Seeing the mean: Ensemble coding for sets of faces. J Exp Psychol Hum Percept Perform. 2009; 35: 718–734. https://doi.org/10.1037/a0013899 PMID: 19485687

17. Morgan M, Chubb C, Solomon JA. A "dipper" function for texture discrimination based on orientation variance. J Vis. 2008; 8(11): 9. https://doi.org/10.1167/8.11.9 PMID: 18831603

18. Norman LJ, Heywood CA, Kentridge RW. Direct encoding of orientation variance in the visual system. J Vis. 2015; 15(4): 3. https://doi.org/10.1167/15.4.3 PMID: 26067349

19. Solomon JA. Visual discrimination of orientation statistics in crowded and uncrowded arrays. J Vis. 2010; 10(14): 19. https://doi.org/10.1167/10.14.19 PMID: 21163954

20. Corbett J, Wurnitsch N, Schwartz A, Whitney D. An aftereffect of adaptation to mean size. Vis Cogn. 2012; 20: 211–231.

21. Alvarez GA, Oliva A. The representation of simple ensemble features outside the focus of attention. Psych Sci. 2008; 19: 392–398.

22. Corbett JE, Oriet C. The whole is indeed more than the sum of its parts: perceptual averaging in the absence of individual item representation. Acta Psychol. 2011; 138: 289–301.

23. Cant JS, Xu Y. Object ensemble processing in human anterior-medial ventral visual cortex. J Neurosci. 2012; 32: 7685–7700. https://doi.org/10.1523/JNEUROSCI.3325-11.2012 PMID: 22649247

24. Treue S, Hol K, Rauber HJ. Seeing multiple directions of motion–physiology and psychophysics. Nature Neurosci. 2000; 3: 270–276. https://doi.org/10.1038/72985 PMID: 10700260

25. Alvarez GA. Representing multiple objects as an ensemble enhances visual cognition. Trends Cogn Sci. 2011; 15: 122–131. https://doi.org/10.1016/j.tics.2011.01.003 PMID: 21292539

26. Brady TF, Alvarez GA. Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. Psych Sci. 2011; 22: 384–392.

27. Brady TF, Konkle T, Alvarez GA. A review of visual memory capacity: Beyond individual items and towards structured representations. J Vis. 2011; 11(5): 4. https://doi.org/10.1167/11.5.4 PMID: 21617025

28. Im HY, Chong SC. Mean size as a unit of visual working memory. Perception. 2014; 43: 663–676. https://doi.org/10.1068/p7719 PMID: 25223110

29. Im HY, Park SJ, Chong SC. Ensemble statistics as units of selection. J Cogn Psychol. 2015; 27: 114–127.

30. Attarha M., Moore CM. The capacity limitations of orientation summary statistics. Atten Percept Psychophys. 2015; 77: 1116–1131. https://doi.org/10.3758/s13414-015-0870-0 PMID: 25810160

31. Attarha M., Moore CM, Vecera SP. Summary statistics of size: Fixed processing capacity for multiple ensembles but unlimited processing capacity for single ensembles. J Exp Psychol Hum Percept Perform. 2014; 40: 1440–1449. https://doi.org/10.1037/a0036206 PMID: 24730736

32. Ariely D. Better than average? When can we say that subsampling of items is better than statistical summary representations? Percept Psychophys. 2008; 70: 1325–1326. https://doi.org/10.3758/PP.70.7.1325 PMID: 18927014

33. Chong SC, Joo SJ, Emmanouil TA, Treisman A. Statistical processing: Not so implausible after all. Percept Psychophys. 2008; 70: 1327–1334. https://doi.org/10.3758/PP.70.7.1327 PMID: 18927015

34. Robitaille N, Harris IM. When more is less: Extraction of summary statistics benefits from larger sets. J Vis. 2011; 11(12): 18. https://doi.org/10.1167/11.12.18 PMID: 22031908

35. Treisman AM. How the deployment of attention determines what we see. Vis Cogn. 2006; 14: 411–443. https://doi.org/10.1080/13506280500195250 PMID: 17387378

36. Utochkin IS, Tiurina NA. Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. Acta Psychol. 2014; 146: 7–18.

37. Allik J, Toom M, Raidvee A, Averin K, Kreegipuu K. An almost general theory of mean size perception. Vis Res. 2012; 83: 25–39.

38. Marchant AP, Simons DJ, De Fockert JW. Ensemble representations: Effects of set size and item heterogeneity on average size perception. Acta Psychol. 2013; 142: 245–250.

39. Maule J, Franklin A. Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. J Opt Soc of America A. 2016; 33: A22–A29.

40. Myczek K, Simons DJ. Better than average: alternatives to statistical summary representations for rapid judgments of average size. Percept Psychophys. 2008; 70: 772–788. PMID: 18613626

41. Simons DJ, Myszek K. Average size perception and the allure of a new mechanism. Percept Psychophys. 2008; 70: 1335–1336.

42. Haberman J, Whitney D. Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In Wolfe J, Robertson L, editors. From Perception to Consciousness: Searching with Anne Treisman. Oxford University Press; 2012. pp. 339–349.

43. Cohen MA, Dennett D, Kanwisher N. What is the bandwidth of perceptual experience? Trends Cogn Sci. 2016; 20: 324–335. https://doi.org/10.1016/j.tics.2016.03.006 PMID: 27105668

44. Whiting BF, Oriet C. Rapid averaging? Not so fast. Psychonomic Bull Rev. 2011; 18: 484–489.

45. Yantis S. Sensation and Perception. New York: Worth Publishers; 2014.

46. Huang L, Mo L, Li Y. Measuring the interrelations among multiple paradigms of visual attention: An individual differences approach. Exp Psychol Hum Percept Perform. 2012; 38: 414–428.

47. Wilmer JB. How to use individual differences to isolate functional organization, biology, and utility of visual functions; with illustrative proposals for stereopsis. Spatial Vis. 2008; 21: 561–579.

48. Haberman J, Brady TF, Alvarez GA. Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. J Exp Psychol General. 2015; 144: 432–446.

49. Kahneman D. Attention and Effort. New Jersey: Prentice Hall; 1973.

50. Sperling G, Melchner MJ. The attention operating characteristic: Examples from visual search. Science. 1978; 202: 315–318. PMID: 694536

51. Wickens CD. Multiple resources and performance prediction. Theor Issue Ergon Sci. 2002; 3: 159–177.

52. Alvarez GA, Horowitz TS, Arsenio HC, DiMase JS, Wolfe JM. Do multielement visual tracking and visual search draw continuously on the same visual attention resources? Exp Psychol Hum Percept Perform. 2005; 31: 643–667

53. Jackson-Nielsen M, Cohen MA, Pitts M. Perception of ensemble statistics requires attention. Cons Cogn. 2017; 48: 149–160.

54. Chong SC, Treisman AM. Statistical processing: Computing average size in perceptual groups. Vis Res. 2005; 45: 891–900. https://doi.org/10.1016/j.visres.2004.10.004 PMID: 15644229

55. Duncan J. Selective attention and the organization of visual information. J Exp Psychol General. 1984; 113: 501–517.

56. Emmanouil TA, Treisman A. Dividing attention across feature dimensions in statistical processing of perceptual groups. Percept Psychophys. 2008; 70: 946–954. PMID: 18717382

57. Huang L. Statistical properties demand as much attention as object features. PLoS ONE. 2015; 10(8): e0131191. https://doi.org/10.1371/journal.pone.0131191 PMID: 26295808

58. Attarha M, Moore CM. The perceptual processing capacity of summary statistics between and within feature dimensions. J Vis. 2015; 15(4): 9. https://doi.org/10.1167/15.4.9 PMID: 26360153

59. Poltoratski S, Xu Y. The association of color memory and the enumeration of multiple spatially overlapping sets. J Vis. 2013; 13(8): 6. https://doi.org/10.1167/13.8.6 PMID: 23838606

60. Lee H, Baek J, Chong SC. Perceived magnitude of visual displays: Area, numerosity, and mean size. J Vis. 2016; 16(3): 12. https://doi.org/10.1167/16.3.12 PMID: 26873776

61. Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Methods. 2009; 41: 1149–1160. https://doi.org/10.3758/BRM.41.4.1149 PMID: 19897823

62. Pierce JW. PsychoPy—psychophysics software in Python. J Neurosci Methods. 2007; 162: 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017 PMID: 17254636

63. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t-tests for accepting and rejecting the null hypothesis. Psychon Bull Rev. 2009; 6: 225–237.

64. Jeffreys H. Theory of Probability (3rd ed.). Oxford, UK: Oxford University Press; 1961.

65. Kass RE, Raftery AE. Bayes factors. J Amer Stat Assoc. 1995; 90: 773–795.

66. JASP Team. JASP (Version 0.8.1.2) [Computer software]. 2017.

67. Halberda J, Mazzocco M, Feigenson L. Individual differences in nonverbal number acuity predict maths achievement. Nature. 2008; 455: 665–668. https://doi.org/10.1038/nature07246 PMID: 18776888

68. Fodor J. The Modularity of Mind. Cambridge, MA: MIT Press; 1983.

69. Dehaene S, Piazza M, Pinel P, Cohen L. Three parietal circuits for number processing. Cogn Neuropsychol. 2001; 20: 487–506.

70. Nieder A, Dehaene S. Representation of number in the brain. Annu Rev Neurosci. 2009; 32: 185–208. https://doi.org/10.1146/annurev.neuro.051508.135550 PMID: 19400715

71. Pavlovskaya M, Soroker N, Bonneh YS, Hochstein S. Computing an average when part of the population is not perceived. J Cogn Neurosci. 2015; 27: 1397–1411. https://doi.org/10.1162/jocn_a_00791 PMID: 25647339

72. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. J Vis. 2009; 9(12): 13. https://doi.org/10.1167/9.12.13 PMID: 20053104

73. Rosenholtz R. Capabilities and limitations of peripheral vision. Annu Rev Vis Sci. 2016; 2: 437–457. https://doi.org/10.1146/annurev-vision-082114-035733 PMID: 28532349

74. Rosenholtz R, Huang J, Raj A, Balas BJ, Ilie L. A summary statistic representation in peripheral vision explains visual search. J Vis. 2012; 12(4): 14. https://doi.org/10.1167/12.4.14 PMID: 22523401