Independent and parallel visual processing of ensemble statistics: Evidence from dual tasks

Vladislav A. Khvostov
National Research University Higher School of Economics, Moscow, Russian Federation

Igor S. Utochkin
National Research University Higher School of Economics, Moscow, Russian Federation

The visual system can represent multiple objects in a compressed form of ensemble summary statistics (such as object numerosity, mean, and feature variance/range). Yet the relationships between the different types of visual statistics remain relatively unclear. Here, we tested whether two summaries (mean and numerosity, or mean and range) are calculated independently from each other and in parallel. Our participants performed dual tasks requiring a report about two summaries in each trial, and single tasks requiring a report about one of the summaries. We estimated trial-by-trial correlations between the precision of reports as well as correlations across observers. Both analyses showed the absence of correlations between different types of ensemble statistics, suggesting their independence. We also found no decrement (except that related to the order of report explained by memory retrieval) in performance in dual compared to single tasks, which suggests that two statistics of one ensemble can be processed in parallel.

Introduction

To overcome the severe restrictions of the processing bottleneck associated with the limited capacity of attention and working memory, the visual system has to compress the big and often redundant amount of visual information received from the environment. One strategy to accomplish such a compressed representation is "to discriminate and to reproduce the statistical moments" related to features and objects in the visual field (Whitney & Leib, 2018, p. 8). This information is often referred to as *ensemble summary statistics* (Alvarez, 2011). The first statistical moment which can be easily computed by the visual system is the *mean*. Observers can extract the information about the average (or central tendency) along a bunch of features from low-level dimensions, such as orientation (Alvarez & Oliva, 2009; Dakin & Watt, 1997; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), hue (de Gardelle & Summerfield, 2011; Maule & Franklin, 2015), brightness (Bauer, 2009), speed (Watamaniuk & Duchon, 1992), spatial position (Alvarez & Oliva, 2008), and size (Ariely, 2001; Chong & Treisman, 2003), to high-level dimensions such as emotional expression, gender, facial identity, gaze direction, and head rotation of a crowd (Florey, Clifford, Dakin, & Mareschal, 2016; Haberman & Whitney, 2007; Sweeny & Whitney, 2014). The second statistical moment is the *variance* (or diversity or range), which can also be computed for low-level (Dakin & Watt, 1997; Morgan, Chubb, & Solomon, 2008; Norman, Heywood, & Kentridge, 2015; Solomon, Morgan, & Chubb, 2011; Suárez-Pinilla, Seth, & Roseboom, 2018) and high-level features (Haberman, Lee, & Whitney, 2015). Another important and well-studied ensemble statistic is the *numerosity* (or sample size, in conventional terms of regular statistics), which is related to the ability to coarsely estimate a number of objects (Burr & Ross, 2008; Chong & Evans, 2011; Halberda, Sires, & Feigenson, 2006). Ensemble perception is not restricted to visual modality: Some studies have shown that people can represent averages in the auditory modality (Albrecht, Scholl, & Chun, 2012; McDermott, Schemitsch, & Simoncelli, 2013; Piazza, Sweeny, Wessel, Silver, & Whitney, 2013). Susceptibility to adaptation aftereffects (Burr & Ross, 2008; Corbett, Wurnitsch, Schwartz, & Whitney, 2012; Norman et al., 2015), the speed extraction of ensemble information (as quickly as 50–200 ms; Chong & Treisman, 2003; Whiting & Oriet, 2011), and the possibility of computing ensemble summary statistics with limited or no conscious access to individual objects (Alvarez & Oliva, 2008; Ariely, 2001; Corbett & Oriet, 2011; Parkes et al., 2001) all support the idea that these statistics can be directly encoded by the visual system along with basic perceptual properties. This seems to be in agreement

Citation: Khvostov, V. A., & Utochkin, I. S. (2019). Independent and parallel visual processing of ensemble statistics: Evidence from dual tasks. *Journal of Vision*, *19*(9):3, 1–18, https://doi.org/10.1167/19.9.3.

https://doi.org/10.1167/19.9.3 Received March 6, 2019; published August 7, 2019 ISSN 1534-7362

with neurophysiological studies in animals (Treue, Hol, & Rauber, 2000), healthy subjects (Cant & Xu, 2012), and individuals with vision conditions (Leib et al., 2012).

The functional architecture of ensemble perception

The described omnipresence of ensemble statistics naturally raises the question of their functional relatedness. Is the full range of ensemble computations performed by a unitary cognitive system ("the general statistician") or are they provided by different mechanisms? How are several ensemble summaries coordinated in gaining access to conscious perception—that is, can they be calculated in parallel and without mutual interference? Both these questions have been previously addressed to some degree using two major approaches to studying various domains in perception and cognition. To answer the first question, an individual-difference approach seems to be useful. Here, testing correlations between performance scores in a set of tasks can shed light on whether there can be a common source of variance for any of these tasks (Huang, Mo, & Li, 2012; Underwood, 1975; Wilmer, 2008). To answer the second question, a parallelism test can be applied, originally coming from studies of divided attention (Alvarez, Horowitz, Arsenio, DiMase, & Wolfe, 2005; Kahneman, 1973; Sperling & Melchner, 1978; Wickens, 2002). While the individual-difference approach and the parallelism test were mostly used in this study, some other approaches could potentially be helpful in addressing these issues (e.g., Rodriguez-Cintron, Wright, Chubb, & Sperling, 2019; Sun, Chubb, Wright, & Sperling, 2016).

Within the individual-difference (or correlational) approach, several tests are run in a group of participants and correlations between these tests are estimated. If some test scores are correlated, it is possible to infer that some overlap between underlying mechanisms can exist. Otherwise, we conclude that there is no common mechanism.

The parallelism test is often based on a precue/postcue paradigm. Observers are presented with several independent targets (e.g., two objects, two sets or strings of objects), one of which is to be subsequently reported. In a precue condition, participants are informed before stimulus presentation which target will be tested at the end of a trial. In a postcue condition, no preliminary information is presented until the response is requested. It is supposed that full attention can be given to one target in the precue condition and that attention should be divided between two or more targets in the postcue condition. If performance stays the same in both precue and postcue trials, it can be concluded that processing of all targets is parallel. If performance degrades under postcue, attention is supposed to be divided imperfectly and the two processes likely competing for the limited-capacity bottleneck.

In the existing literature on ensemble perception, examples of studies addressing both independence and parallelism can be found. For example, Haberman, Brady, and Alvarez (2015) found evidence for independence between averaging processes in "low-level" (color and orientation) and "high-level" (facial expression and identities) domains, although they found some correlations within each of these levels. Emmanouil and Treisman (2008) found that two different average features could not be processed completely in parallel, although the cost of dividing attention was not dramatic. It appears that dividing attention between different summaries (e.g., mean color and mean orientation) is less demanding than dividing attention between different ensembles, even when the same summary is computed (Attarha & Moore, 2015; Attarha, Moore, & Vecera, 2014; Chong & Treisman, 2005; Halberda et al., 2006; Huang, 2015; Poltoratski & Xu, 2013; Utochkin & Vostrikov, 2017).

While most of the aforementioned studies were concerned with parallelism and independence in statistical computations of the same type (several means or several numerosities), less is known about summary statistics of different types. For example, are the number of objects, their mean size, and their size variance calculated by independent computational mechanisms? Can attention be divided among all these summaries in parallel? Only few studies have addressed one or both of these questions. In one such study, Yang, Tokita, and Ishiguchi (2018) demonstrated the lack of correlation between the perception of mean and the perception of variance. This suggests that these two statistics can be computed quite independently, even when applied to the same domain (size or orientation). Lee, Baek, and Chong (2016) asked a similar question about mean size, numerosity, and total area estimated by participants in three separate tasks, trying to predict performance on each of these tasks based on the other two. They found that the total area could be predicted from judgments of mean size and numerosity, but neither mean size nor numerosity could be predicted by the other two summaries. This suggests the relative computational independence of mean and numerosity (although the authors found a small correlation between performance in these two statistics). Utochkin and Vostrikov (2017) tested both parallelism and independence for the perception of mean size and numerosity. In three experiments, they compared the precision of judgments of mean size and numerosity under precue and postcue (parallelism test) and estimated the correlation between these two tasks

(individual-difference approach). In one experiment, participants had to estimate the mean size and the numerosity in one set of circles, which were found to be not affected by the cue manipulation. In another experiment, the observers had to estimate the mean size in one subset and the numerosity in another, and a postcue decrement was found. But the same postcue decrement was found when the participants estimated only one statistic (only mean or only numerosity) in two subsets concurrently. The researchers concluded that the mean and numerosity can be computed in parallel within one ensemble but compete when they belong to different ensembles (see also Duncan, 1984; Huang, 2015). Importantly, they also found an absence of correlation between the tasks that was consistent with the conclusions of Lee et al. about independence.

Our study

Although testing the perception of two statistics in a single precue/postcue task (Utochkin & Vostrikov, 2017) is advantageous over separate tasks (Haberman, Brady, & Alvarez, 2015; Lee et al., 2016; Yang et al., 2018) in terms of its capability to probe parallelism along with independence, an important methodological caveat can be put forward. Specifically, when a correlation between tests is estimated across participants, each data point represents an average score that does not take into account what happens in particular trials. Yet this trial-by-trial picture can be important for the interpretation of the averaged data points. In fact, an absence of cross-observer correlation between the mean size and the numerosity (Utochkin & Vostrikov, 2017) could reflect genuinely parallel allocation of independent resources to both tasks, but it also could reflect a somewhat interdependent allocation of resources fluctuating across trials. For example, if an observer pays more attention to the mean on one half of trials and to the number on another half of trials, his or her average performance would be hard to distinguish from parallel allocation. Although it is a good tool for testing hypotheses about the division of attention between two tasks, the precue/postcue task lacks power to directly probe how attention is divided in particular trials, because each trial measures performance on only one of the tasks (e.g., only mean size or only numerosity).

In the present study, we developed a paradigm allowing us to test performance on two ensemble statistics in each trial instead of testing only one. This manipulation has a long history in the literature on divided attention as a means to probe parallelism (Sperling & Melchner, 1978). Most importantly, it has been used to answer the question whether different target features are represented independently within each single trial (e.g., Bays, Wu, & Husain, 2011; Fougnie & Alvarez, 2011). In terms of our main research focus, the double report would allow us to disentangle *genuinely independent parallel processes* in representing various ensemble summaries from *fluctuating attentional reallocation* to one statistic at the cost of another. If the two representations are indeed independent and served by parallel processes, then we expect low correlation between the precision of representing one statistic and the precision of estimating the other within a trial. Conversely, if attentional reallocation occurs, then better precision at reporting one statistic would entail worse precision at reporting the other—that is, they should correlate negatively. Apart from the absolute precision, systematic biases can be a useful source of information about relationships between the ensemble summaries. For example, if the mean size is derived from the total area divided by the number (as regular statistics defines the mean), then underestimated numerosity should correlate with overestimated mean size, and vice versa. Again, our double-response method is appropriate for testing this prediction due to its power to track trial-by-trial bias fluctuations, whereas averaged scores can fail to detect such correlations in scores collapsed across the fluctuations.

As the double response requires dividing attention between two target features (dual task), it should be matched to baseline conditions that require focusing attention on each of the target features (single tasks). This is the gold standard of testing parallelism in a dual-task paradigm (Sperling & Melchner, 1978). We administered these single-task baseline tests and the dual-task tests in separate blocks.

Finally, while previous studies have focused on testing the links either between mean and numerosity (Lee et al., 2016; Utochkin & Vostrikov, 2017) or between mean and variance (Yang et al., 2018) and used different approaches, here we tested both these links using the same approach.

Experiment 1

Methods

Participants

To determine the number of participants, we used the statistical tool G*Power 3.0.10 (Faul, Erdfelder, Lang, & Buchner, 2007) for a priori power analysis. We set the required statistical power at 0.8, Type I error at 0.05, and an expected effect size at 0.6 (this value was based on the previous research of cue effects on perception of numerosity and mean size; Emmanouil & Treisman, 2008; Halberda et al., 2006; Poltoratski &

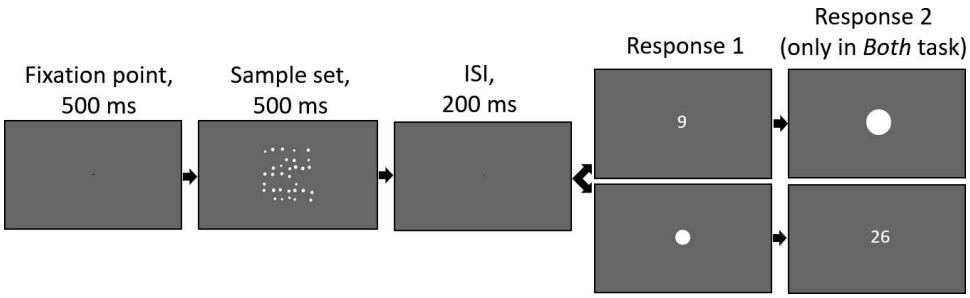

Figure 1. The time course of a typical trial in Experiment 1.

Xu, 2013). Considering a power issue related to correlational analyses, we used the parameters of Utochkin and Vostrikov's (2017) study, which had a similar design. All these parameters led us to a minimum sample size of 19 participants. Considering the possibility of technical problems or poor performance in some participants, we increased the number of recruited observers by one to five participants in each experiment. Twenty-four undergraduate students at the Higher School of Economics participated in Experiment 1 (20 female, four male; mean age = 18.3 years) for extra course credits. All had normal or corrected-to-normal vision and no neurological problems. At the beginning of the experiment, the participants gave written informed consent.

Apparatus and stimuli

Stimuli were developed and presented through PsychoPy for Linux (Pierce, 2007). They were presented on a standard VGA monitor with a refresh frequency of 75 Hz and a spatial resolution of 1,024 × 768 pixels. Viewing distance was approximately 50 cm, and the width and height of the screen were 36.6 and 27.4 cm, respectively. From this viewing distance, one pixel subtended approximately 0.04° of visual angle. A 22.64° × 22.64° square at the center of the screen was used as a working field for presenting stimuli; the rest of the screen space remained empty. The working field was divided into $7 \times 7 = 49$ cells by an imaginary grid (each cell 3.28° on a side). Each cell could be used for positioning a single item of a stimulus set. Within the cell, an item could be randomly jittered within ±0.82° along both the horizontal and vertical directions.

We used white circles as items in a sample set. Individual circle diameters in a trial were randomly chosen from a uniform distribution. The minimum value of the distribution was randomly chosen from the interval from 0.49° to 0.94° with a step of 0.04° (one pixel); the maximum value was the minimum multiplied by 1.6. Thus, some circles in a set could have identical sizes. The mean diameter could vary from 0.57° to 1.31°. The physical diameters were scaled to fit Teghtsoonian's (1965) perceived size scale, which has been shown to be used by some observers for estimating mean size (Chong & Treisman, 2003; Lee et al., 2016). The correct response for a report of mean size, therefore, was defined as an average of individual sizes in the units of Teghtsoonian's scale. The number of circles varied from seven to 36, with each particular number presented three times per block of trials. The mean size and numerosity were assigned independently of each other in every trial.

Procedure

Each trial (Figure 1) started with the presentation of a fixation point for 500 ms, followed by a sample display for 500 ms. At 200 ms after the sample offset, a probe item (a random integer from 1 to 45 or a circle with a random size from 0.25° to 2.05°) appeared. Using this probe item, the participants had to report either the number of presented elements in the sample display (numerosity), the mean size of circles, or both (sequentially one after another). The observers set their answers by rotating a mouse wheel upward or downward to increase or decrease the numeral or the diameter of the circle. Allowed ranges for responses were the same as the aforementioned limits for probe items. As soon as observers set an appropriate value, they had to press the space bar on a keyboard to enter the answer. No feedback was provided about the precision of answers. The next trial started upon a repeated press of the space bar, so observers could progress at a comfortable pace and take a rest whenever they wanted.

There were three tasks in the experiment. In the first task, participants had to report the mean size of all presented circles (Mean task). In the second task, they had to report the number of presented circles (Numerosity task). In the third task, they had to sequentially recall both the mean and the number (Both task). The serial order of mean and number reports was balanced across trials within the Both task and treated as a fixed factor in subsequent analyses. As the manipulation with the serial order of responses duplicated the necessary number of trials, the Both task was presented in two blocks of 90 trials preceded by

eight practice trials. Therefore, there were four blocks ($90 \times 4 = 360$ trials per observer) in total, whose serial order was counterbalanced across participants. Each participant completed a unique sequence of blocks out of 24 possible serial orders.

Design and data analysis

For each of the tested ensemble summaries, we obtained the measures of baseline single-task performance in the Mean or the Numerosity task and dual-task performance in the Both task. The latter measures were split based on the serial order of report (Mean–Numerosity vs. Numerosity–Mean). Therefore, we compared performance in three different conditions for each of the summaries. Our primary measure of interest was the normalized absolute error (Error), which is inversely related to precision. The Error was calculated in each trial as Error = |Participant's Response − Correct Response|/Correct Response. Normalized signed Error (Signed Error) was additionally calculated for the trial-by-trial analysis of biases: Signed Error = (Participant's Response − Correct Response)/Correct Response.

We applied three different analyses of Errors. The first analysis was related only to the Both task, where participants had to report two different statistics in each trial. For each participant, we calculated a trial-by-trial correlation between an error in the mean judgment and an error in the numerosity judgment. Second, we estimated correlations across participants, like previous researchers did (e.g., Haberman, Brady, & Alvarez, 2015; Utochkin & Vostrikov, 2017; Yang et al., 2018). Specifically, we were interested in autocorrelations of each of the summary statistics under three different conditions (e.g., mean size at baseline, mean size when reported first, and mean size when reported second) and in cross-correlations between the mean and numerosity under similar conditions (e.g., mean baseline with numerosity baseline and mean when reported first with numerosity when reported first). The autocorrelations aimed to test the reliability of the measured ensemble representations across conditions. The cross-correlations aimed to provide estimates of independence or interdependence between mean and numerosity on the macro level (based on scores averaged within each observer, like in previous studies; Haberman, Brady, & Alvarez, 2015; Utochkin & Vostrikov, 2017; Yang et al., 2018). Third, in order to estimate parallelism in the dual task we compared performance in this task with corresponding single-task performance using t tests (e.g., mean when reported first vs. mean baseline, mean when reported second vs. mean baseline, and mean when reported first vs. mean when reported second).These analyses were run using standard significance tests and Bayes factors. In the Bayesian statistical inference, the Bayes factor (BF_{10}) is the odds showing the relative likelihood of H_1 compared to H_0 given the data. The Bayes factors were calculated in JASP statistical software (JASP 0.9.0.1; JASP, Amsterdam, the Netherlands). Jeffreys's (1961) scale, with Kass and Raftery's (1995) adjustment, was used to interpret the Bayes factors. For t tests, Bayes factors were calculated using the Bayesian t test; we used a prior width set at $r = 0.707$ (Rouder, Speckman, Sun, Morey, & Iverson, 2009), which is recommended as the default value for this test (JASP; Wagenmakers et al., 2018). For calculating Bayes factors for correlation (Ly, Marsman, & Wagenmakers, 2018), we used a default uniform prior (JASP; Wagenmakers et al., 2018).

In this experiment, we used slightly different displays for reporting the numerosity (a numeral having a symbolic nature) and the mean size (a single circle having a more visual nature). We chose these displays because they have been often used in previous studies for the measurement of numerosity (e.g., Halberda et al., 2006) and mean size (e.g., Utochkin & Vostrikov, 2017). Although the differences in report formats could affect how our observers calibrated their judgment scales for each of the tasks, we did not consider it as a substantial problem, since we did not directly compare these scales (for example, we did not ask whether mean size was estimated more precisely than numerosity). Rather, we compared judgment errors only within each task, asking how these errors change as a function of the division of attention (dual vs. single task) or of another task's performance (correlation analysis).

Results and discussion

The data of one participant were excluded from analysis because the average Error exceeded three group standard deviations in two out of six conditions. Therefore, the data of 23 participants were analyzed. As response ranges were limited by the experimental procedure, it could yield floor and ceiling effects on responses. To diminish these effects, we excluded trials with ceiling and floor responses along at least one dimension (1 and 45 for numerosity; 0.25° and 2.05° for the mean) from analysis. In total, 0.3% trials were excluded from analysis based on this restriction.

Data from this and all the other experiments have been deposited on OSF (https://osf.io/g5rwy/).

Trial-by-trial mean–numerosity correlations within the dual task

We found that, in 20 participants, correlation coefficients between Errors in reporting the mean size and the numerosity did not reach significance (rs <

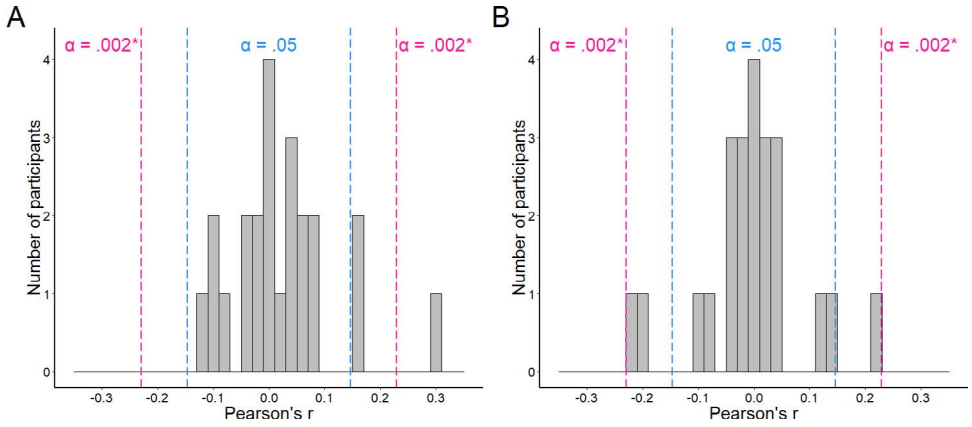

Figure 2. The distribution of correlation coefficients between (A) normalized absolute errors and (B) normalized signed errors in reporting the mean size and numerosity in Experiment 1 (trial-by-trial analyses of responses in Both task separate for each observer). The blue and red lines represent the values of Pearson's *r*, after which correlations become significant at 0.05 and 0.002 (*Bonferroni-corrected for 23 comparisons) levels correspondingly (for df = number of paired observations − 2 = 178).

0.12, ps > 0.108, BF_{10}s < 0.338), and only three participants showed a small or moderate positive correlation (rs > 0.155, ps < 0.033). Bayes factors for two of them provided bare evidence in favor of the absence of correlation ($BF_{10} = 0.769$ and 0.923), and only one participant had a strong value ($BF_{10} = 491.3$). The distribution of correlation coefficients is plotted in Figure 2A; note that concerning false negative results, we compare the correlation coefficient not only with the Bonferroni-corrected value of alpha ($\alpha = 0.002$) but also with the classical one ($\alpha = 0.05$). For Signed Errors (Figure 2B), the results were similar: Most of the observers showed weak and nonsignificant correlations (rs < 0.225, ps > 0.002, BF_{10}s < 0.287), but three of them showed moderate correlations (rs > 0.114, ps < 0.131, BF_{10}s > 2.631).

Note that the trial-by-trial correlation analysis over 180 trials per participants is quite a large sample, potentially inflating the statistical power and the probability of Type I error. Considering this point, the lack of correlations in this case strongly favors unrelatedness between judgments of numerosity and mean size.

Correlations across participants

The conclusion about nonoverlapping mechanisms for computing different ensemble statistics can be also supported by cross-correlation data between average data points obtained from individual participants (Figure 3). We found that the Error in reporting mean size in all experimental conditions (baseline, Mean–Numerosity, and Numerosity–Mean) did not correlate with the Error in reporting numerosity in the corresponding conditions (rs < 0.269, ps > 0.215, Bonferroni-corrected $\alpha = 0.006$, BF_{10}s < 0.532; Figure 3A). At the same time, the autocorrelations between judgments of mean size under different conditions were high (rs > 0.649, ps < 0.001, Bonferroni-corrected $\alpha = 0.006$, BF_{10}s > 51.52; Figure 3B), as were the autocorrelations for numerosity judgments (rs > 0.502, ps < 0.016, Bonferroni-corrected $\alpha = 0.006$, BF_{10}s > 4.308; Figure 3C).

Dual-task versus single-task performance

Pairwise comparisons between the single task (baseline) and the two report orders in the dual task showed that the dual-task estimates of both mean size and numerosity were as precise as their baselines when reported first, ts(22) < 2.504, p > 0.02, Bonferroni-corrected $\alpha = 0.008$, Cohen's ds < 0.523, confidence interval (CI) [–0.261, 0.954]; Bayesian analysis showed strong evidence for this result for mean calculations ($BF_{10} = 0.278$) but bare evidence for better performance at baseline for the numerosity task ($BF_{10} = 2.746$). However, when a tested statistic was reported second, the Errors were much greater compared to the baseline and to the report in the first place, ts(22) > 2.946, p < 0.008, Bonferroni-corrected $\alpha = 0.008$, $BF_{10} > 6.287$, Cohen's ds > 0.614, CI [0.162, 1.514]. We conclude, therefore, that the dual task per se did not impair (or only slightly impaired) performance in any of the tasks (Figure 4 depicts the Errors in the form of attention operating characteristics; Sperling & Melchner, 1978), although the order of report had an effect. It turns out, therefore, that there was no substantial cost of dividing attention between the two summaries, which is in line with parallelism. The effect of the second report order can likely be explained by memory interference at recall rather than by problems that could arise from the division of attention at encoding.

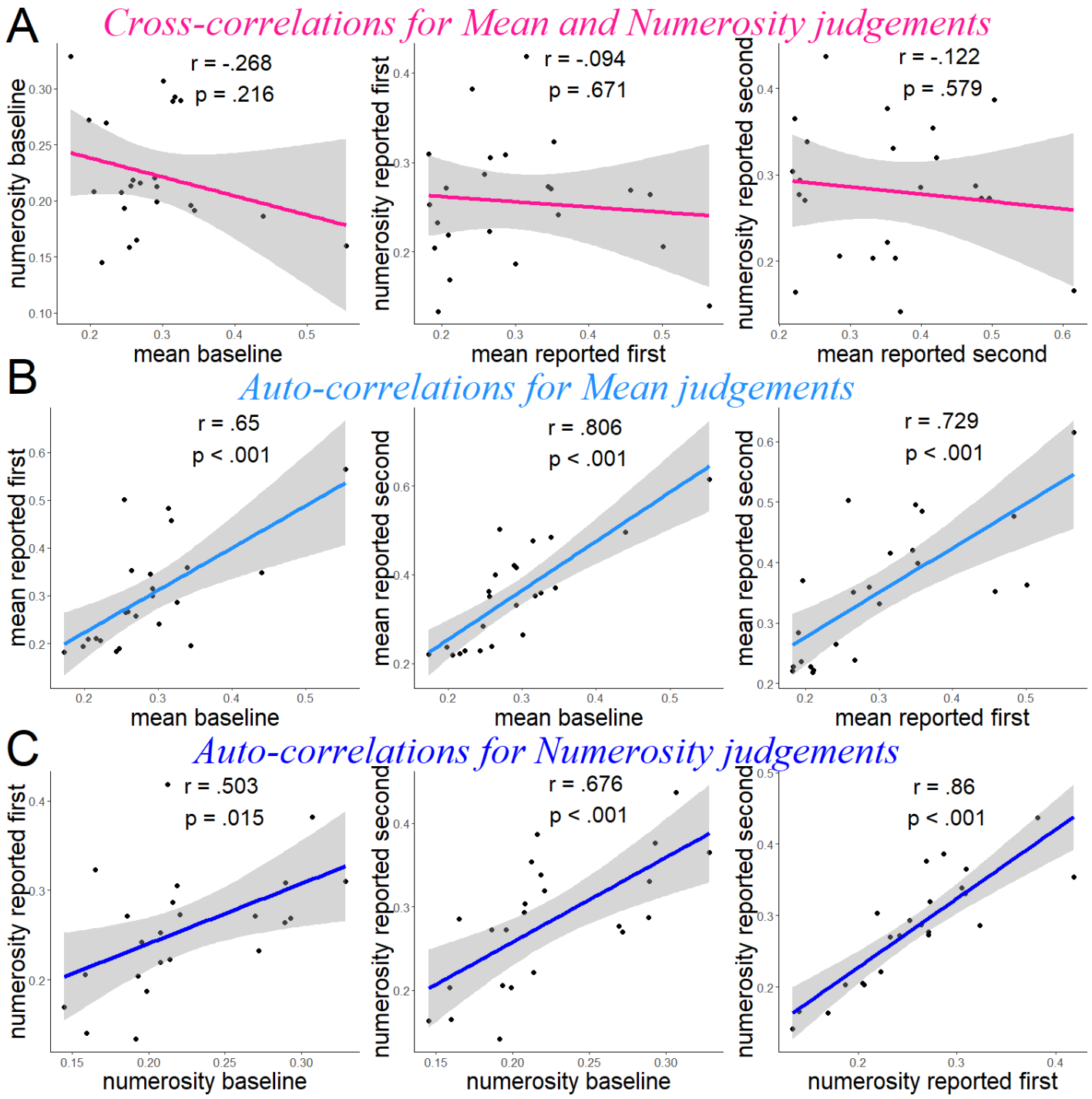

Figure 3. (A) The cross-correlations between reports of different summaries in the same report conditions (baseline, first- and second-order reports in dual task) and (B) the autocorrelations of mean estimation and (C) numerosity estimation under different report conditions (baseline, first- and second-order reports in dual task). The gray regions denote 95% confidence intervals.

Experiment 2A

To test whether our conclusion about parallelism and independence between various ensemble summaries can be generalized beyond mean and numerosity, we introduced the variance (range) perception in Experiment 2A. There is ongoing discussion whether people can estimate the variance of sizes or just use the range information (e.g., Lau & Brady, 2018), so in our study we manipulated the range (but note that range is strongly correlated with variance in our manipulations). Specifically, we tested parallelism and independence between averaging and range estimates.

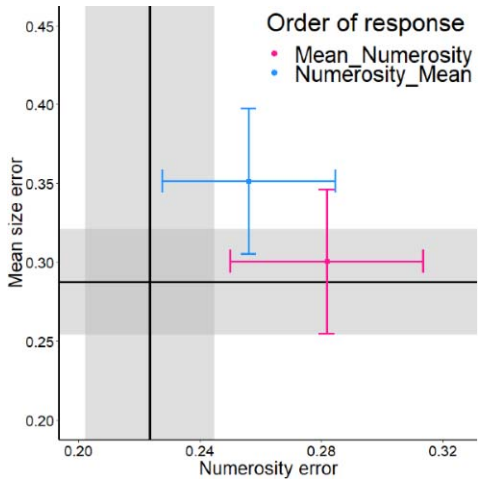

Figure 4. Attention operating characteristics for precision (normalized absolute error) in Experiment 1. Average errors in numerosity estimates are on the *x*-axis, while those in mean size estimates are on the *y*-axis. Black horizontal and vertical lines represent the average error in baseline conditions (Mean and Numerosity tasks, respectively). The red data point denotes the Error in the Both task when participants reported the mean size first; the blue data point denotes the Error in the Both task when participants reported numerosity first. Error bars around the data points and gray shaded areas around the baselines denote ±95% confidence intervals.

However, a few potential issues must be addressed before designing the dual task for these two summaries. Combining mean reports with range reports in a single paradigm can be potentially tricky within a dual task. It is a requirement for such a paradigm that range be manipulated independently from mean at test. If an observer has to report both the mean size and the range of the same sample set, the test set of items for the range report (there is no way to probe range/variance perception of a set other than in another set; Haberman, Lee, & Whitney, 2015, Yang et al., 2018) should have a mean size randomly differing from the true mean size of the sample; otherwise the range test would unambiguously inform the observer about the mean. While some studies have shown that observers can transfer their impression of range/variance from a set with one mean to a set with another mean (e.g., Haberman, Lee, & Whitney, 2015; Solomon et al., 2011), there is an empirical question whether observers' error of reporting range/variance changes with increasing distance between the mean sizes of sample and test sets.

It is also important that manipulations of the mean sizes between sample and test sets entails a potential confound with relative density. If the mean size of all items is changed, then mean spacing between them changes in an opposite direction. It is another empirical question whether this relative density plays a role in range perception. Experiment 2A aimed to answer these two questions. Here our participants performed only a range-adjustment task. We systematically manipulated the difference between mean sizes of a sample set and a test set and the absolute and relative density of these sets in order to see whether the precision of range perception is immune to these manipulations.

Methods

Participants

Sixteen undergraduate students at the Higher School of Economics (11 female, five male; mean age = 19.1 years) participated in the experiment for extra course credits. All had normal or corrected-to-normal vision and no neurological problems. At the beginning of experiment, the participants gave written informed consent.

Apparatus and stimuli

We used the same apparatus as in Experiment 1. Both sample and test screens always consisted of 16 white circles presented within an imaginary square ($4 \times 4 = 16$ cells) with a center at fixation. Each cell contained one circle positioned at the center of the cell with a random jitter of ±10% of a side length. The side length of each cell, in turn, could be *fixed* or *scaled*. In the fixed condition, a side of a cell was always 4.09° for both sample and test sets regardless of their difference in mean size. In the scaled condition, the side length was the mean size multiplied by a factor of 2.5. These two conditions were used to control for potentially complicated interactions between variations in mean size and spatial density that inevitably correlate. In the fixed layout, the absolute density (number of circles per degree of visual angle) was constant but the relative density (ratio between mean circle size and mean between-circles space) changed depending on the mean size. By contrast, the scaled layout provided stability of relative density with variable absolute density.

The mean diameter of a sample set of circles was randomly chosen from the narrow interval between 1.46° and 1.56°, whereas the mean diameter of the test set differed from the sample mean on average by the percentage from the interval [−60%, 60%] with a step equal to 10% along Teghtsoonian's (1965) perceived size scale. The size distribution always consisted of four sizes equally spaced along Teghtsoonian's size scale, with each size assigned to four circles in a set (therefore, it was always a uniform equally spaced distribution). The range was measured as a distance between the biggest and the smallest sizes in units of the mean and could be drawn from the interval between 0.20 and 1.60

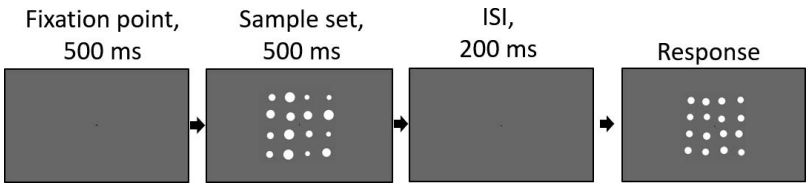

Figure 5. The time course of a typical trial in Experiment 2A.

in a sample and between 0 and 1.80 in a test. For example, in the range of 0.20, the biggest circle was 1.10 of the mean size and the smallest circle was 0.90 of the mean size; in the range of 1.60, the biggest circle was 1.80 of the mean size and the smallest circle was 0.20 of the mean size. In other words, the biggest and the smallest sizes were $\pm 1/2$ of the range away from the mean. Two middle sizes were $\pm 1/6$ of the range away from the mean. Importantly, the same coefficients defined the speed of size changes during range adjustments, thus preserving the equally spaced uniform distribution with a stable mean.

Procedure

The time course of a typical trial in the experiment was similar to that of a typical trial in Experiment 1. The only difference concerned the response (see Figure 5). Participants had to adjust the range of a test set to match the range of a sample set. To do this, they had to rotate a mouse wheel that changed the range of the size distribution (rotating upward increased the range, rotating downward decreased it).

In the instruction, we explained the concept of range as "a general impression of the diversity of sizes" and also as "the degree of contrast between big and small circles." This provided participants a clear understanding of the task.

Design and data analysis

In this experiment, we parametrically varied the difference between mean sizes of a sample set and a test set in a broad range from −60% to 60% (13 points in total including 0%), providing the uniform distribution of all values across trials. The second fixed factor was layout (two conditions: fixed vs. scaled). Each observer was exposed to 2 (layout) × 13 (differences in mean size) × 12 repetitions = 312 trials preceded by 26 practice trials. For data analysis, we merged trials with equal absolute differences in mean size but opposite signs. This yielded 24 trials per data point in each observer. For the 0 difference in mean size, we merged trials from the fixed and scaled layouts, as these two layouts were metameric under this difference in mean size. This also yielded 24 trials per data point (it is easy to note that the same subset trials were used as an input to an analysis of variance [ANOVA] for both layouts under this particular difference in mean size).

For the precision of range estimates, we calculated normalized absolute error using the same formula as in Experiment 1. There is no strong consensus among researchers (e.g., Haberman, Lee, & Whitney, 2015; Suárez-Pinilla et al., 2018) whether the error in range/variance adjustment should be normalized (like we did in Experiment 1, adding the correct answer as a denominator) or not by the reference range/variance (correct answer). However, as it has been shown that the error tends to increase with the reference in a fashion according to Weber's law (Haberman, Lee, & Whitney, 2015), we consider our normalized formula to be justified by the nature of the measured property.

We applied two analyses to our data. First, the correlation between correct responses and observers' absolute responses was estimated to test whether participants could perform the task at all. Second, a 2 (layout: fixed vs. scaled) × 7 (absolute difference in mean size: 10, 20, 30, 40, 50, 60) repeated-measures ANOVA was applied to values of Error to test whether observers transferred their range impressions from the mean size of a sample to the mean size of the test and to estimate the effect of layout. For the Bayesian ANOVA, the width r of a prior Cauchy distribution of effect sizes was set at 0.5, following the default settings recommended by Wagenmakers et al. (2018) and the JASP Team for fixed-effects models.

Results and discussion

To diminish the floor and ceiling effects of the adjustment range, we removed all extreme answers (0 and 1.80 ranges) from analysis, meaning 3% of responses were not taken into analysis. We found that participants could report the range of a sample rather precisely, which is supported by a highly positive correlation between correct and observer responses ($r = 0.624$, $p < 0.001$, $BF_{10} > 10^{25}$; Figure 6A).

The ANOVA model (Figure 6B) showed no significant effect of layout according to frequentist statistics, $F(1, 15) = 16.563$, $p = 0.144$, $\eta^2_p = 0.137$), although Bayesian statistics showed an inconclusive result ($BF_{10} = 1.013$). Also, we found no effect of the absolute difference in mean size, $F(6, 90) = 0.843$, $p = 0.54$, $\eta^2_p = 0.053$, $BF_{10} = 0.042$, nor of the interaction of the two

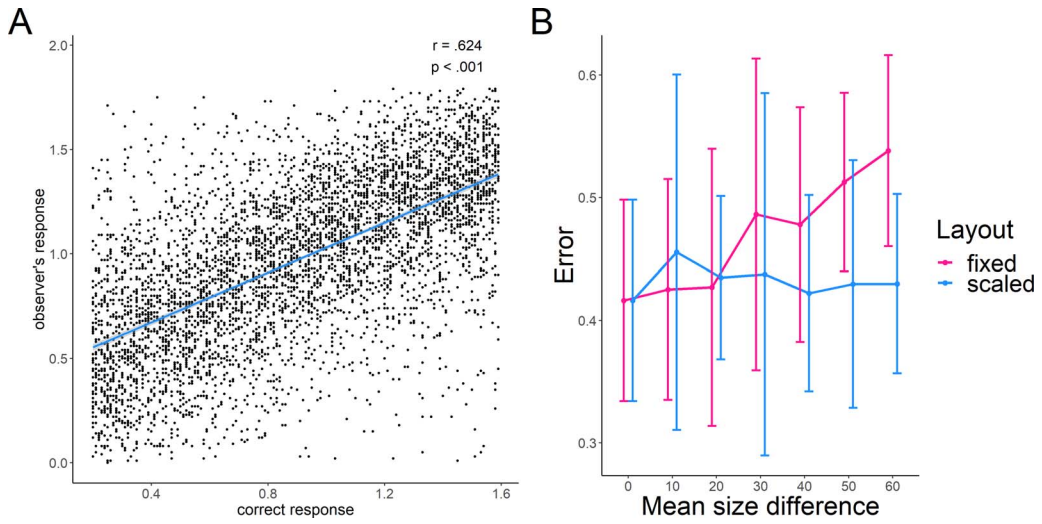

Figure 6. The results of Experiment 2A. (A) Trial-by-trial correlation between correct and observer responses. (B) Error as a function of layout and difference in mean size. Error bars (B) and gray region (A) denote ±95% confidence intervals.

factors, $F(6, 90) = 1.565$, $p = 0.166$, $\eta^2_p = 0.094$, $BF_{12} = 6.286$ (where BF_{12} is a ratio of BF_{10} for a model taking into account only the main effects to BF_{10} for a model taking into account the main effects and interaction).

Overall, our results show that the participants could report range with constant precision regardless of the manipulations with mean difference between the sample and the test and with layout. The results of Experiment 2A suggest that it is possible to adjust range on a probe set not informing observers about the mean size of a sample set. Therefore, we conclude that the mean and range can be estimated independently within a single trial of a dual task.

Experiment 2B

In Experiment 2B, we basically replicated the design of Experiment 1. However, the pair of tested ensemble statistics was different, namely, mean and range.

Methods

Participants

Nineteen undergraduate students at the Higher School of Economics participated in the experiment (18 female, one male; mean age = 21.4 years) for extra course credits. All had normal or corrected-to-normal vision and no neurological problems. At the beginning of experiment, the participants gave written informed consent.

Apparatus, stimuli, and procedure

We used the same apparatus as in Experiments 1 and 2A. However, the numerosity task was replaced by a range task similar to that used in Experiment 2A. Sample sets were made in the same way as in the fixed condition of Experiment 2A. Specifically, there were always 16 circles, each having one of four sizes drawn from a uniform equally spaced distribution along Teghtsoonian's (1965) scale. The circles were located within a 4×4 grid with a fixed cell side of ~4.09°. The mean size of a sample set could be taken from the interval between 0.68° and 1.7°. The range of the sample set could be 0.3, 0.6, 0.9, 1.2, or 1.5 (see the explanation of units in Experiment 2A), which were uniformly distributed across the experiment. The mean size and range were assigned independently from each other in every trial.

The adjustment of the perceived mean size was performed on a single test circle at fixation (like in Experiment 1), whose diameter was randomly drawn from the interval between 0.25° and 2.05° and changed within the same interval by rotating a mouse wheel. The adjustment of perceived size range was performed on a set of 16 circles (like in Experiment 2A) with a fixed mean size of 1.25°. The range could be changed by rotating a mouse wheel that increased or decreased the diversity of the test distribution between 0 and 1.80 with a step of 0.02.

The procedure was the same as in Experiment 1 in terms of timing and events, except that the numerosity task was replaced by the range task from Experiment 2A (see Figure 7). As in Experiment 1, we ran two single tasks for each statistic and a dual task with a varying order of report.

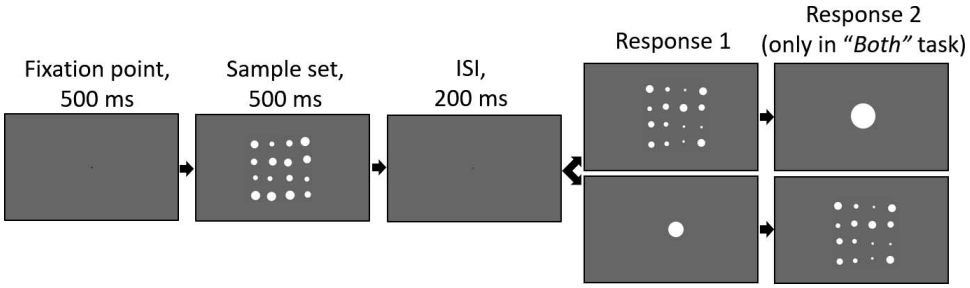

Figure 7. The time course of a typical trial in Experiment 2B.

Design and data analysis

The design, data analysis, and dependent variables were the same as in Experiment 1. The only difference is that instead of the Numerosity task we had a Range task. As in Experiment 1, we argue that the difference in report displays (a single probe item in the mean size task and a set in the range task) could not influence our results (see earlier for these arguments).

Results and discussion

Like in the two previous experiments, we excluded trials with ceiling and floor responses along at least one dimension (0.25° and 2.05° for the mean) from analysis. In total, 3.8% of trials were excluded from analysis based on this restriction.

Trial-by-trial Mean–Range correlations within the dual task

We found that most of the participants (Figure 8A) showed no evidence of trial-by-trial correlation between Error values in reporting the mean size and the range ($rs < 0.143$, $ps > 0.053$, $BF_{10}s < 0.573$), but two showed a small or moderate positive correlation ($rs > 0.144$, $ps < 0.044$, $BF_{10} = 0.68$ and 2.066). We can conclude that there is no interrelation between the precision of mean and range computations, which can suggest their independence. As in Experiment 1, we also tested whether systematic biases (Signed Error) correlated between the mean and range judgments. We did not find strong evidence for such a correlation: All participants showed weak and nonsignificant correlation ($rs < 0.17$, $ps > 0.02$, $BF_{10}s < 1.384$; Figure 8B), except for two who showed small to moderate positive correlations ($rs > 0.15$, $ps < 0.037$, $BF_{10} = 0.779$ and 1.383).

Correlations across participants

The computational independence of mean and range perception can be also supported by auto- and cross-correlations of observers' averaged scores (Figure 9). We found no cross-correlations between reported mean size and range regardless of the task or report order ($rs < 0.36$, $ps > 0.141$, Bonferroni-corrected $\alpha = 0.006$, $BF_{10}s < 0.775$; Figure 9A). At the same time, the

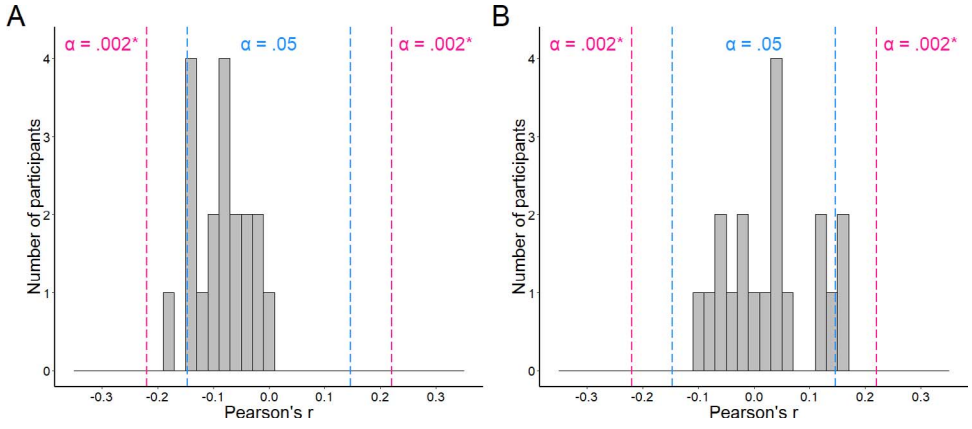

Figure 8. The distribution of correlation coefficients between (A) normalized absolute errors and (B) normalized signed errors in reporting the mean size and numerosity in Experiment 2B (trial-by-trial analyses of responses in the Both task separate for each observer). The blue and red lines represent the values of Pearson's *r*, after which correlations become significant at 0.05 and 0.002 (Bonferroni-corrected for 23 comparison) levels correspondingly (for df = number of paired observations − 2 = 178).

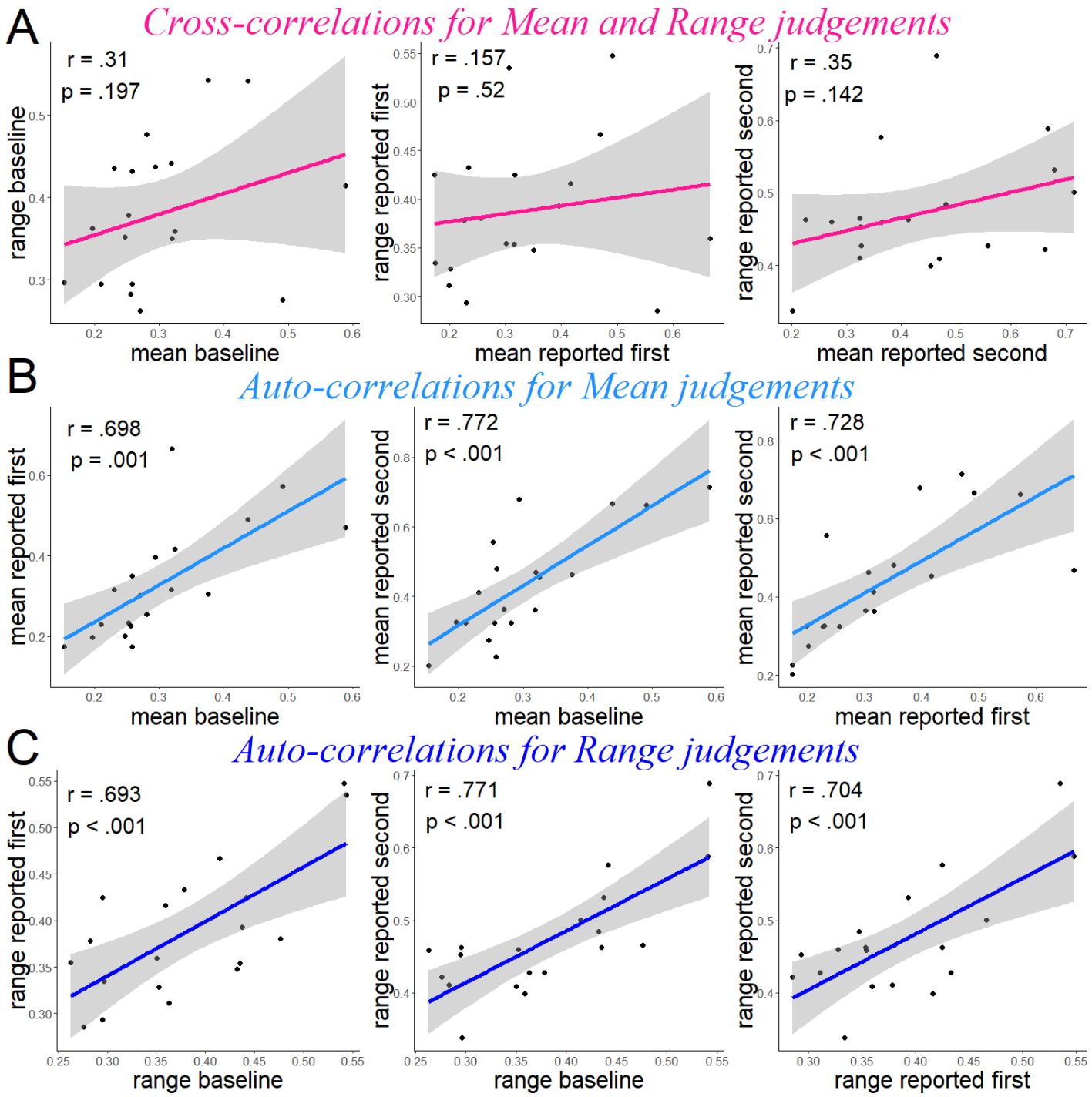

Figure 9. (A) The cross-correlations between reports of different summaries in the same report conditions (baseline, first- and second-order reports in dual task) and the autocorrelations of (B) mean estimation and (C) range estimation under different report conditions (baseline, first- and second-order reports in dual task). The gray regions denote 95% confidence intervals.

autocorrelations were strong for both the mean ($rs > 0.698$, $ps < 0.001$, Bonferroni-corrected $\alpha = 0.006$, $BF_{10}s > 47.326$; Figure 9B) and the range ($rs > 0.692$, $ps < 0.001$, Bonferroni-corrected $\alpha = 0.006$, $BF_{10}s > 43.306$; Figure 9C), suggesting the cross-task consistency of the measurements.

Dual-task versus single-task performance

We found that dual-task performance in both mean and range judgments did not differ from their single-task baselines when these statistics were reported in the first place, $ts(18) < 1.199$, $ps > 0.246$, Bonferroni-corrected $\alpha = 0.008$, $BF_{10}s < 0.444$, Cohen's $ds < 0.514$,

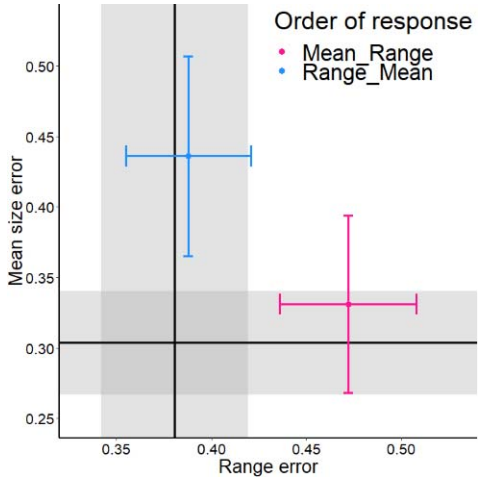

Figure 10. Attention operating characteristics for precision (Error) in Experiment 3. Average errors in range estimates are on the *x*-axis, average errors in mean size estimates are on the *y*-axis. Black horizontal and vertical lines represent the average Error in baseline conditions (Mean and Range tasks, respectively). The red data point denotes the Error in the Both task when participants reported the mean size first; the blue data point denotes the Error in the Both task when participants reported range first. Error bars around the data points and gray shaded areas around the baselines denote ±95% confidence intervals.

CI [–0.304, 0.729]. However, when these summaries were reported in the second place, the responses were less precise than the corresponding baselines and first responses, $ts(18) > 4.133$, $ps < 0.001$, Bonferroni-corrected $\alpha = 0.008$, $BF_{10}s > 55.773$, Cohen's $ds > 0.947$, CI [0.395, 2.3]. This pattern is very similar to that in Experiment 1 and leads to a similar principal conclusion, that attention can be divided between the mean size and range in the dual task without any substantial cost (Figure 10). Therefore, the processing of mean and range are likely parallel. However, the cost associated with the second order of report suggests that there can be memory interactions that can lead to decay or interference for the second response.

General discussion

Summarizing the results of the experiments, we can conclude that the mean size and numerosity (Experiment 1) or mean size and range (Experiment 2B) can be extracted from one ensemble by computationally independent mechanisms; the results of these calculations do not mutually interfere in accessing the limited capacity of conscious processing. This finding supports previous reports of independence (Haberman, Brady & Alvarez, 2015; Yang et al., 2018) and parallelism (Utochkin & Vostrikov, 2017) but at a deeper level of evidence. Specifically, we asked our participants to report both tested statistics in the same trial and were able to track whether a change in one representation (e.g., mean size) was responsive to the change in another (e.g., numerosity or range). This analysis allowed us to dissociate between genuinely independent and parallel processes and those "pretending" to be independent and parallel on the macro level of average scores across participants but being in fact negatively related (for example, due to fluctuating allocation of attention between one or another statistic from trial to trial). We discuss the evidence we obtained in more detail.

Different statistical summaries are computed independently

The claim about independent processing comes mainly from the results of correlational analyses which have the following logic of interpretation. If two different computations are done by independent mechanisms, these calculations should have two different sources of variance, thus predicting zero or very weak correlations between tests measuring these computational processes. In contrast, strong positive correlations would likely indicate some considerable source of variance, probably suggesting commonality in computational mechanisms (like those we naturally observed as autocorrelations of each summary statistic). Haberman, Brady, and Alvarez (2015) have also pointed out that some small correlation might be explained by general factors, such as working-memory capacity or motivation: They obtained a correlation of ~0.2 between ensemble and digital memory-span tasks (but they used much bigger samples of participants). Negative correlations could also reflect some common mechanism for both computations but acting in a different way: Computing one property is carried out at the cost of another, which directly addresses the issue of parallelism. (Note, however, that a strong conclusion about parallelism cannot be made on the basis of correlation alone, and control for the division of attention is necessary).

We obtained unambiguous evidence in favor of independently computed ensemble summaries in Experiments 1 and 2B. The ensemble summaries did not correlate on either the trial-by-trial micro level or the individual-difference macro level. On the micro level, the absence of a positive correlation provides evidence against a common computational mechanism (e.g., a more precise number estimate is not associated with a more precise average estimate, although they are related in regular mathematical statistics), and the absence of a negative correlation makes it unlikely that

one computation is run at the cost of another. Moreover, this conclusion about precision is corroborated by the lack of correlation between the systematic biases (e.g., an overestimated number of items is not associated with an overestimated mean size of those items, which might be predicted based on the regular mathematical method of averaging). On the macro level, we replicated previous demonstrations of uncorrelatedness between mean and numerosity precision (Utochkin & Vostrikov, 2017) and between mean and variance (or range) precision (Yang et al., 2018), showing that observers who are better at estimating one statistic are not necessarily better at estimating another.

An additional argument for independence comes from Experiment 2A, which was initially designed to ensure that mean size and range could be manipulated and tested independently in a dual task and that these manipulations kept the task doable. Our finding that participants did not show any loss in the precision of reported range, despite rather big differences between mean sizes of a sample and a test, supports the idea that range is indeed an independently processed ensemble property that can be transferred across different mean sizes. In a different experimental design, Norman et al. (2015) came to a similar conclusion about the transfer of a variance (range) impression across ensembles with different mean orientations. Also, Tong, Ji, Chen, and Fu (2015, experiment 2) found basically the same for the brightness domain. Their participants had to report which of two sequential displays had higher variance, and the accuracy of the variance report in a condition where the mean brightness was stable between two displays was identical to that in the condition where the mean brightness changed between two displays.

Note that our results do not imply that there is no link between the representations of various statistical summaries. As an example of such a link, a lot of studies have shown that the precision of judgments of mean size usually decreases with an increase in the physical variance of a display (Corbett et al., 2012, experiment 4; Im & Halberda, 2013; Maule & Franklin, 2015; Tong et al., 2015; Utochkin & Tiurina, 2014). It is important, however, that the claims based on the manipulations of physical variance in these studies concern mostly the role of the external, stimulus noise in the representation of the mean. In our study, we were focused on the correlations between errors produced by observers when they estimated each of the summaries. The lack of correlations we observe between the errors could reflect the uncorrelatedness of internal noise sources for each of the judgments. Prospective psychophysical and neuroscientific research can be focused on probing these sources of internal noise to better understand the nature of ensemble representations.

Different statistical summaries are computed in parallel

Our conclusion about parallelism is based on the finding that dual-task performance was as precise as single-task performance, at least when a given summary was reported first, which is the same temporal position as in the single task. This finding corroborates the claim about parallelism of mean and numerosity computations made by Utochkin and Vostrikov (2017) based on the precue/postcue paradigm. The converging evidence we report here is a valuable addition because of slightly different ways to manipulate the division of attention. Specifically, we diminished a potential cost of switching between different statistics from trial to trial that could blur differences between Utochkin and Vostrikov's precue and postcue performance scores. Here, we measured single-task and dual-task performance in separate blocks, encouraging participants to dedicate their attention to each task without switches. Moreover, the obligatory requirement to report both statistics in each trial of the dual task also encouraged our observers to compute both properties. Therefore, with this additional control for task switching and the replicated pattern from Utochkin and Vostrikov, we provided a stronger conclusion about parallel processing of mean and numerosity. The evidence for parallel processing of mean and range is a new conclusion that generalizes this parallelism to a broader range of ensemble properties. It is also in line with earlier reports of no or minor costs of dividing attention between same-type statistics of different visual dimensions within the same ensemble (Emmanouil & Treisman, 2008; Huang, 2015). In a broader perspective, our finding is in accordance with a classical idea that it is easier to divide attention between different properties of a single object than between the properties of different objects (Duncan, 1984).

Implications for the architecture of ensemble perception

These results lead to two theoretical consequences. First, we can view computational processes leading to the extraction of mean size, numerosity, and range as relatively nonoverlapping. Therefore, it is unlikely that there is a general "statistical processor" for different ensemble properties, even if they belong to the same set of objects. Second, none of these computational processes uses the results of the work of another process, unlike in regular statistics. Although it may be tempting to expand this logic of regular statistics (where, for example, mean is calculated based on the number) to the visual domain, our and other previous studies (Lee et al., 2016; Utochkin & Vostrikov, 2017;

Yang et al., 2018) provide evidence against such expansion. The conclusion about independence also provides an important direction of investigation of neural mechanisms beyond ensemble representations. Specifically, the independence may suggest that different neural populations can be involved in representing different ensemble properties. For example, there is neurophysiological evidence that different brain regions can be involved in numerosity perception and in the perception of statistics of internal properties of multiple objects, such as color or shape (Cant & Xu, 2012; Dehaene, Piazza, Pinel, & Cohen, 2003). Although mean and range (variance) can originate from the same early representation (e.g., Khayat & Hochstein, 2018), their computations can be also affected by nonoverlapping sources of later noise (Solomon et al., 2011), which could explain our and previously reported findings (Utochkin & Vostrikov, 2017; Yang et al., 2018) about the lack of correlation between averaging and range (variance) estimation. In future research, these speculations should be addressed thoroughly.

Keywords: ensemble statistics, *numerosity*, *mean*, *range*, *dual task*

Acknowledgments

Research was supported by the Russian Science Foundation (grant number 18-18-00334 to ISU).

Commercial relationships: none.
Corresponding author: Vladislav A. Khvostov.
Email: vkhvostov@hse.ru.
Address: National Research University Higher School of Economics, Moscow, Russian Federation.

References

Albrecht, A. R., Scholl, B. J., & Chun, M. M. (2012). Perceptual averaging by eye and ear: Computing summary statistics from multimodal stimuli. *Attention, Perception, & Psychophysics*, *74*(5), 810–815, https://doi.org/10.3758/s13414-012-0293-0.

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

Alvarez, G. A., Horowitz, T. S., Arsenio, H. C., DiMase, J. S., & Wolfe, J. M. (2005). Do multielement visual tracking and visual search draw continuously on the same visual attention resources? *Journal of Experimental Psychology: Human Perception and Performance*, *31*(4), 643–667, https://doi.org/10.1037/0096-1523.31.4.643.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392–398, https://doi.org/10.1111/j.1467-9280.2008.02098.x.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences, USA*, *106*(18), 7345–7350, https://doi.org/10.1073/pnas.0808981106.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162.

Attarha, M., & Moore, C. M. (2015). The capacity limitations of orientation summary statistics. *Attention, Perception, & Psychophysics*, *77*(4), 1116–1131, https://doi.org/10.3758/s13414-015-0870-0.

Attarha, M., Moore, C. M., & Vecera, S. P. (2014). Summary statistics of size: Fixed processing capacity for multiple ensembles but unlimited processing capacity for single ensembles. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1440–1449, https://doi.org/10.1037/a0036206.

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *The Psychological Record*, *59*(2), 171–185, https://doi.org/10.1007/BF03395657.

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*(6), 1622–1631, https://doi.org/10.1016/j.neuropsychologia.2010.12.023.

Burr, D., & Ross, J. (2008). A visual sense of number. *Current Biology*, *18*(6), 425–428, https://doi.org/10.1016/j.cub.2008.02.052.

Cant, J. S., & Xu, Y. (2012). Object ensemble processing in human anterior-medial ventral visual cortex. *The Journal of Neuroscience*, *32*(22), 7685–7700, https://doi.org/10.1523/JNEUROSCI.3325-11.2012.

Chong, S. C., & Evans, K. K. (2011). Distributed versus focused attention (count vs estimate): Distributed versus focused attention. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 634–638, https://doi.org/10.1002/wcs.136.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *12*, 393–404.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in percep-

tual groups. *Vision Research*, *45*(7), 891–900, https://doi.org/10.1016/j.visres.2004.10.004.

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica*, *138*(2), 289–301, https://doi.org/10.1016/j.actpsy.2011.08.002.

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

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*(22), 3181–3192, https://doi.org/10.1016/S0042-6989(97)00133-8.

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences, USA*, *108*(32), 13341–13346.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*(3–6), 487–506, https://doi.org/10.1080/02643290244000239.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*(4), 501–517, https://doi.org/10.1037/0096-3445.113.4.501.

Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across feature dimensions in statistical processing of perceptual groups. *Perception & Psychophysics*, *70*(6), 946–954, https://doi.org/10.3758/PP.70.6.946.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191, https://doi.org/10.3758/BF03193146.

Florey, J., Clifford, C. W. G., Dakin, S., & Mareschal, I. (2016). Spatial limitations in averaging social cues. *Scientific Reports*, *6*:32210 https://doi.org/10.1038/srep32210.

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12):3, 1–12, https://doi.org/10.1167/11.12.3. [PubMed] [Article]

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432–446, https://doi.org/10.1037/xge0000053.

Haberman, J., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision*, *15*(4):16, 1–11, https://doi.org/10.1167/15.4.16. [PubMed] [Article]

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753, https://doi.org/10.1016/j.cub.2007.06.039.

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, *17*(7), 572–576, https://doi.org/10.1111/j.1467-9280.2006.01746.

Huang, L. (2015). Statistical properties demand as much attention as object features. *PLoS One*, *10*(8), e0131191, https://doi.org/10.1371/journal.pone.0131191.

Huang, L., Mo, L., & Li, Y. (2012). Measuring the interrelations among multiple paradigms of visual attention: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(2), 414–428, https://doi.org/10.1037/a0026314.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception, & Psychophysics*, *75*(2), 278–286, https://doi.org/10.3758/s13414-012-0399-4.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, *18*(9):23, 1–14, https://doi.org/10.1167/18.9.23. [PubMed] [Article]

Lau, J. S.-H., & Brady, T. F. (2018). Ensemble statistics accessed through proxies: Range heuristic and dependence on low-level properties in variability discrimination. *Journal of Vision*, *18*(9):3, 1–18, https://doi.org/10.1167/18.9.3. [PubMed] [Article]

Lee, H., Baek, J., & Chong, S. C. (2016). Perceived magnitude of visual displays: Area, numerosity, and mean size. *Journal of Vision*, *16*(3):12, 1–11, https://doi.org/10.1167/16.3.12. [PubMed] [Article]

Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia*,

50(7), 1698–1707, https://doi.org/10.1016/j.neuropsychologia.2012.03.026.

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Analytic posteriors for Pearson's correlation coefficient: Analytic correlation posterior. *Statistica Neerlandica*, *72*(1), 4–13, https://doi.org/10.1111/stan.12111.

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision*, *15*(4):6, 1–18, https://doi.org/10.1167/15.4.6. [PubMed] [Article]

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, *16*(4), 493–498, https://doi.org/10.1038/nn.3347.

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A 'dipper' function for texture discrimination based on orientation variance. *Journal of Vision*, *8*(11):9, 1–8, https://doi.org/10.1167/8.11.9. [PubMed] [Article]

Norman, L. J., Heywood, C. A., & Kentridge, R. W. (2015). Direct encoding of orientation variance in the visual system. *Journal of Vision*, *15*(4):3, 1–14, https://doi.org/10.1167/15.4.3. [PubMed] [Article]

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739–744, https://doi.org/10.1038/89532.

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science*, *24*(8), 1389–1397, https://doi.org/10.1177/0956797612473759.

Pierce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Poltoratski, S., & Xu, Y. (2013). The association of color memory and the enumeration of multiple spatially overlapping sets. *Journal of Vision*, *13*(8):6, 1–11, https://doi.org/10.1167/13.8.6. [PubMed] [Article]

Rodriguez-Cintron, L. M., Wright, C. E., Chubb, C., & Sperling, G. (2019). How can observers use perceived size? Centroid versus mean-size judgments. *Journal of Vision*, *19*(3):3, 1–14, https://doi.org/10.1167/19.3.3. [PubMed] [Article]

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237, https://doi.org/10.3758/PBR.16.2.225.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, *11*(12):13, 1–11, https://doi.org/10.1167/11.12.13. [PubMed] [Article]

Sperling, G., & Melchner, M. (1978, October 20). The attention operating characteristic: Examples from visual search. *Science*, *202*(4365), 315–318, https://doi.org/10.1126/science.694536.

Suárez-Pinilla, M., Seth, A. K., & Roseboom, W. (2018). Serial dependence in the perception of visual variance. *Journal of Vision*, *18*(7):4, 1–24, https://doi.org/10.1167/18.7.4. [PubMed] [Article]

Sun, P., Chubb, C., Wright, C. E., & Sperling, G. (2016). The centroid paradigm: Quantifying feature-based attention in terms of attention filters. *Attention, Perception, & Psychophysics*, *78*(2), 474–515, https://doi.org/10.3758/s13414-015-0978-2.

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: Ensemble perception of a crowd's gaze. *Psychological Science*, *25*(10), 1903–1913, https://doi.org/10.1177/0956797614544510.

Teghtsoonian, M. (1965). The judgment of size. *The American Journal of Psychology*, *78*(3), 392–402, https://doi.org/10.2307/1420573.

Tong, K., Ji, L., Chen, W., & Fu, X. (2015). Unstable mean context causes sensitivity loss and biased estimation of variability. *Journal of Vision*, *15*(4):15, 1–12, https://doi.org/10.1167/15.4.15. [PubMed] [Article]

Treue, S., Hol, K., & Rauber, H.-J. (2000). Seeing multiple directions of motion—physiology and psychophysics. *Nature Neuroscience*, *3*(3), 270–276, https://doi.org/10.1038/72985.

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist*, *30*(2), 128–134, https://doi.org/10.1037/h0076759.

Utochkin, I. S., & Tiurina, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta Psychologica*, *146*, 7–18, https://doi.org/10.1016/j.actpsy.2013.11.012.

Utochkin, I. S., & Vostrikov, K. O. (2017). The numerosity and mean size of multiple objects are perceived independently and in parallel. *PLoS One*, *12*(9), e0185452, https://doi.org/10.1371/journal.pone.0185452.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76, https://doi.org/10.3758/s13423-017-1323-7.

Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research*, *32*(5), 931–941, https://doi.org/10.1016/0042-6989(92)90036-I.

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

Whitney, D., & Leib, A. Y. (2018). Ensemble Perception. *Annual Review of Psychology*, *69*(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177, https://doi.org/10.1080/14639220210123806.

Wilmer, J. (2008). How to use individual differences to isolate functional organization, biology, and utility of visual functions; with illustrative proposals for stereopsis. *Spatial Vision*, *21*(6), 561–579, https://doi.org/10.1163/156856808786451408.

Yang, Y., Tokita, M., & Ishiguchi, A. (2018). Is there a common summary statistical process for representing the mean and variance? A study using illustrations of familiar items. *i-Perception*, *9*(1), 204166951774729,https://doi.org/10.1177/2041669517747297.