Dear Author

Here are the proofs of your article.

- You can submit your corrections **online** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- Please return your proof together with the permission to publish confirmation.
- For **fax** submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the journal title, article number, and your name when sending your response via e-mail, fax or regular mail.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.

Please note

Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL:

`http://dx.doi.org/10.3758/s13414-020-02192-y`

If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information, go to: http://www.springerlink.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us, if you would like to have these documents returned.

The **printed version** will follow in a forthcoming issue.

Metadata of the article that will be visualized in OnlineFirst

1	Article Title	**Ensemble perception: Extracting the average of perceptual versus numerical stimuli**	
2	Article Sub- Title		
3	Article Copyright - Year	**The Psychonomic Society, Inc. 2020 (This will be the copyright line in the final PDF)**	
4	Journal Name	Attention, Perception, & Psychophysics	
5	Corresponding Author	Family Name	**Rosenbaum**
6		Particle	
7		Given Name	**David**
8		Suffix	
9		Organization	Tel Aviv University
10		Division	
11		Address	Tel Aviv-Yafo, Israel
12		e-mail	davidros28@gmail.com
13	Author	Family Name	**Gardelle**
14		Particle	**de**
15		Given Name	**Vincent**
16		Suffix	
17		Organization	Paris School of Economics and CNRS, Centre d'Economie de la Sorbonne
18		Division	
19		Address	Paris, France
20		e-mail	
21	Author	Family Name	**Usher**
22		Particle	
23		Given Name	**Marius**
24		Suffix	
25		Organization	Tel Aviv University
26		Division	
27		Address	Tel Aviv-Yafo, Israel
28		e-mail	
29	Schedule	Received	
30		Revised	

Accepted 25 October 2020

Abstract

Recent research has established that humans can extract an average perceptual feature over briefly presented arrays of visual elements or the average of a rapid temporal sequence of numbers. Here we compared the extraction of the average over briefly presented arrays, for a perceptual feature (orientations) and for numerical values (1–9 digits), using an identical experimental design for the two tasks. We hypothesized that the averaging of numbers, more than of orientations, would be constrained by capacity limitations. Arrays of Gabor elements or digits were simultaneously presented for 300 ms and observers were required to estimate the average on a continuous response scale. In each trial the elements were sampled from normal distributions (of various means) and we varied the set size (4–12). We found that while for orientation the averaging precision remained constant with set size, for numbers it decreased with set size. Using computational modeling we also extracted capacity parameters (the number of elements that are pooled in the average extraction). Despite marked heterogeneity between observers, the capacity for orientations (around eight items) was much larger than for numbers (around four items). The orientation task also had a larger fraction of participants relying on distributed attention to all elements. Our study thus supports the idea that numbers more than perceptual features are subject to capacity or attentional limitations when observers need to evaluate the average over an ensemble of stimuli.

Keywords separated by ' - '

Cognitive neuroscience - Decision making - Math modeling

JrnlID 13414_ArtID 2192_Proof# 1 - 14/11/2020

Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-020-02192-y

Ensemble perception: Extracting the average of perceptual versus numerical stimuli

David Rosenbaum[1] · Vincent de Gardelle[2] · Marius Usher[1]

Accepted: 25 October 2020

Abstract

Recent research has established that humans can extract an average perceptual feature over briefly presented arrays of visual elements or the average of a rapid temporal sequence of numbers. Here we compared the extraction of the average over briefly presented arrays, for a perceptual feature (orientations) and for numerical values (1–9 digits), using an identical experimental design for the two tasks. We hypothesized that the averaging of numbers, more than of orientations, would be constrained by capacity limitations. Arrays of Gabor elements or digits were simultaneously presented for 300 ms and observers were required to estimate the average on a continuous response scale. In each trial the elements were sampled from normal distributions (of various means) and we varied the set size (4–12). We found that while for orientation the averaging precision remained constant with set size, for numbers it decreased with set size. Using computational modeling we also extracted capacity parameters (the number of elements that are pooled in the average extraction). Despite marked heterogeneity between observers, the capacity for orientations (around eight items) was much larger than for numbers (around four items). The orientation task also had a larger fraction of participants relying on distributed attention to all elements. Our study thus supports the idea that numbers more than perceptual features are subject to capacity or attentional limitations when observers need to evaluate the average over an ensemble of stimuli.

Keywords Cognitive neuroscience · Decision making · Math modeling

Introduction

Research over the last two decades indicates that human observers can rapidly extract the average of a perceptual feature over sets of visual objects, even when they cannot discriminate if an individual item in the display was presented (Ariely, 2001; Chong & Treisman, 2003; Chong & Treisman, 2005; Dakin, 2001; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Robitaille & Harris, 2011). For example, humans can evaluate the average size of a set of circles presented simultaneously, with an accuracy that does not decrease as the set contains more elements (Ariely, 2001; Chong & Treisman, 2005) or is presented for a shorter duration (Chong & Treisman, 2003). This averaging ability has been demonstrated even in situations where the discrimination of the presence of individual elements in the array appears at chance (Ariely, 2001). This capacity appears to extend from simple visual attributes – such as size, orientation, and spatial position – to more complex properties such as emotional expression (Haberman & Whitney, 2011). Moreover, the extraction of the average appears to take place automatically or, at least, without "intention," as it occurs in parallel (Chong & Treisman, 2005) and affects judgments of memory, in which the set-average is task-irrelevant (Khayat & Hochstein, 2018).

Another type of stimulus in which ensemble perception has been suggested is symbolic numbers (Brezis, Bronfman, Jacoby, Lavidor, & Usher, 2016; Brezis et al., 2015, 2018; Corbett, Oriet, & Rensink, 2006; Vanunu, Hotaling, & Newell, 2020; Sato & Motoyoshi, 2020; Van Opstal et al., 2011; Vandormael, Herce, Balaguer, Li, & Summerfield (2017); Spietzer et al., 2017). Such stimuli are thought to automatically activate a set of analog numerosity representations (Nieder et al., 2002; Nieder & Miller, 2003), as indicated by well-known distance and magnitude effects (Dehaene, Dupoux, & Mehler, 1990; Moyer & Landauer, 1967) and numerical Stroop effects (Henik & Tzelgov, 1982). Studies of numerical averaging have shown that human observers also have a remarkable ability to identify and average symbolic numbers even under stringent processing constraints. For example, Brezis et al. (2015, Exp. 3) presented observers with a sequence of four to 16 two-digit numbers at a rate of ten

✉ David Rosenbaum
davidros28@gmail.com

[1] Tel Aviv University, Tel Aviv-Yafo, Israel

[2] Paris School of Economics and CNRS, Centre d'Economie de la Sorbonne, Paris, France

items/s, and asked participants to indicate the average on a continuous scale. The results show that the estimation precision (the RMSD) improved with the length of the sequence, indicating that observers did not use only a limited sample of the sequence. These results were accounted for by a population-pooling mechanism (in which encoding noise would average out over items).

In a few other studies it has been shown that observers can extract numerical information from arrays of numerical symbols presented simultaneously. For example, Corbett et al. (2006) have shown that observers are able to discriminate between two circular arrays of six digits (comprising 2s and 5s), presented simultaneously (for as short as 80 ms), the one that had a higher average (more 5s) with an accuracy exceeding 80% (Exp. 1). Critically, this discrimination was faster and more accurate with arrays made of the 2 and 5 symbols than with p and q symbols, and this speedup only took place when the numerical meaning could be used as a basis for the classification task (Exp. 3). Finally, using a dual-task methodology, the authors have shown that this ability requires central attention (Exp. 5). This study thus demonstrates that numerical information is rapidly extracted from arrays of numbers, at least when these arrays are relatively simple. However, this very specific set of stimuli makes it possible for observers to adopt a strategy that might not involve computing of an average over all elements.[1] Situations involving more complex arrays remain to be investigated, as they might help uncover the computational algorithms used by observers to evaluate an average over items.

Two recent studies have taken this approach, using larger arrays of two-digit numbers presented simultaneously (for up to 4–5 s) and asking participants to decide whether the average was smaller or higher than a reference (Vandormael et al., 2017; Vanunu et al., 2020). In these studies, the observers' accuracy improved with presentation time, with the distance of the average from the reference, and with sets involving lower variance. The two studies differed, however, in their conclusion about the algorithm used to carry out the task: whereas Vandormael et al. (2017) found robust-averaging – an algorithm that gives less weight to outliers, Vanunu et al. (2020) found on the contrary that extreme values received equal or higher weights. Although the reasons for this discrepancy are still unclear, at least in both cases participants relied on some items more than on others. This finding relates to the notion of capacity that has been put forward in early cognitive models of attention and working memory, and that has also been part of recent theoretical accounts of ensemble perception.

In the context of extracting a set-average, capacity can be defined as the number of items pooled together in the estimation (Alik et al., 2013; Dakin, 2001; Solomon, May, & Tyler, 2016). Whereas this definition assumes an all-or-none selection of some items and not others, one alternative view involving distributed attention can be considered. In this view, all the elements contribute to the estimation of the average, each element receiving a fraction of the attentional resources available, which becomes smaller when there are more elements in the array (Eriksen & StJames; Baek and Chong, 2020; Chong & Treisman, 2005). As shown by Baek and Chong (2020a), a signature of this model is an improved precision with set size (see also Brezis et al., 2015, for the case of sequential presentation).

The appeal of the notion of capacity or distributed attentional resources is that these notions are domain general, and can be compared across observers and across tasks. Surprisingly, however, and despite the fact that many studies have demonstrated that observers form ensemble representations over various dimensions, how these dimensions compare, for example in terms of the capacity, is not clear. In a recent study, Haberman, Brady, and Alvarez (2015) found that individual differences in performances (mean absolute errors when identifying the average over a set) were correlated between two low-level features such as orientation and color, but uncorrelated when comparing a low-level feature to a higher-level feature such as facial expression. This suggests that ensemble representations for different features might operate with different levels of performance, although capacity or distributed attention was not specifically assessed in this study.

Here, we hypothesize that the capacity with which observers build an average representation might depend on how much attentional and visual working memory resources are involved in extracting and manipulating the task-relevant feature. For instance, we expect that limitations in distributed attention or visual working memory capacity (Cowan, 2001; Luck & Vogel, 1997) will affect the averaging of symbolic numbers (as suggested by Corbett et al., 2006) more than of simple visual properties like orientation, which can be processed pre-attentively (Braun & Sagi, 1991; Treisman & Gelade, 1980) and which engage grouping and the formation of a holistic Gestalt (Hess & Field, 1999; Kovács & Julesz, 1993). While there is some debate on the capacity with which orientation can be averaged in a brief array (Baek & Chong, 2020a; Dakin, 2001; Robitaille & Harris, 2011; Solomon, May, & Tyler, 2016; see review in Baek & Chong, 2020b), we expect that capacity would be reduced for numerical stimuli, which are likely to require more attentional resources due to their higher visual complexity.

The aim of our study was to contrast averaging of numerical and visual oriented elements, within the same observers, and using an identical experimental design (with the same visual display and response procedure for these two dimensions). By manipulating the size of the item-set across trials, we aimed to evaluate how performance changes with set size, and reveal the capacity of the integration process. For both dimensions (numbers vs. orientations) we asked participants

[1] One alternative is that the observers estimate if there are more 2s than 5s, but not by how much (which would allow to decide that the average is higher or lower that 3.5, but not by how much). Alternatively, observers may estimate the average, but this could be based on a VWM capacity sample of about four items.

to report an estimation of the average on a continuous scale, in order to encourage the integration of all items, and minimize the use of non-averaging heuristics that might arise in tasks based on a comparison to a reference. We expect that in the numerical averaging task, participants will be more accurate with smaller arrays. By contrast, in the orientation averaging task we expect either a fixed (or improved) precision with the set size of the array, as a result of averaging the encoding noise. To validate these conclusions, we used computational modeling to fit the data with two models, namely (1) the limited-capacity (subsampling) model (Alik et al., 2013; Solomon, May, & Tyler, 2016) and (2) the distributed attention or 'zoom lens' model (Baek & Chong, 2020a), and extracted the capacity or attention parameters for the two tasks. Finally, we examined the weights given to the mid-range and extreme values and compared them across the tasks (Vandormael et al., 2017; Vanunu et al., 2020).

Experiment

The experiment briefly presented arrays of numbers (digits 1–9) or oriented elements (Gabors) of various set sizes (from four to 12) and required participants to estimate the numerical or orientation average on a continuous scale. We used an estimation on a continuous scale rather than a binary decision relative to a reference, as this minimizes the reliance on some heuristics, such as counting the number of elements higher than the reference, or even the number of extreme (high- vs. low-value) elements. Our main focus is the dependency of the estimation precision on set size in the two tasks.

Methods

Participants

Eighteen healthy adult volunteers with normal or corrected-to-normal vision participated in this study. All volunteers gave written informed consent to participate in this study. All procedures and experimental protocols were approved by the ethics committee of the Psychology Department of Tel Aviv University (Application 743/12). All experiments were carried out in accordance with the approved guidelines. Due to the COVID-situation, testing conditions were restricted. We offered our participants the option to be tested (for an equivalent of $15) in the lab under special safety COVID19 guidelines, or to run the experiment at home (same pay) from their own computer (to do this they needed to have Matlab installed on their computer). Ten participants were tested in the lab and eight were tested at home.

Stimuli

In the lab, displays were generated by an Intel I7 personal computer attached to a 24-in. Asus 248qe monitor with a 144-Hz refresh rate, using 1,920 × 1,080 resolution graphics mode. Due to the Covid19 situation, eight participants were tested at home using their own personal computers, but the experimental code was designed so as to detect the monitor's resolution and present the stimuli with the same relative size. All participants were approximately at a distance of 60 cm from the screen.

The stimulus was an array of four, eight, or 12 elements (Gabor patches or numbers, depending on the task), randomly located on a gray background, within an invisible 5 × 6 grid (each cell was 77 × 96 pixels), with a restriction of no two horizontally adjacent elements and no element in the cells just above and below fixation (see Fig. 1a and b). Numbers were integers between 1 and 9, presented in white in David font size 25. Gabor patches were 200 pixels wide, with a spatial frequency of 0.2 cycles per pixel and standard deviation of 20 pixels. Gabors' orientations varied from 42° to 138° in nine equidistant steps. Stimuli were generated using Psychtoolbox for Matlab.

Trial procedure

Each trial began with the onset of a central fixation dot (1 s) followed by the stimulus array (numbers or Gabors), which remained on the screen for 300 ms. After the offset of the array, participants were instructed to report the numbers' average (number task) or the Gabors' average angle (orientation task) on a semicircular scale (an arc from 30° to 150°), using their mouse. The mouse cursor was always at the middle of the circle (red fixation dot) at the beginning of the scale display and thus it had an equal distance from each point of the scale (see Fig. 1). The scale labels were numbers from 1 to 9 (number task) or oriented lines from 30° to 150° (orientation task). The participants had a 5-s deadline to respond.

Design

Each participant completed both the orientation and the number averaging tasks, in separate blocks of 360 trials each, in an order counterbalanced across participants. Set sizes (four, eight, 12 elements) were randomly interleaved across trials. Numbers were drawn randomly from one of three Gaussian distributions with means of 3.5, 5, or 6.5 and standard deviation of 1.5. The Gabors' orientations were drawn from Gaussian distributions with means of 72°, 90°, and 108° relative to horizontal, and SD of 18°. The positions of numbers 1–9 on the response scale corresponded to orientations from 42° to 138°. Due to a coding error, in four (of the 18) participants the Gaussian distribution of the Gabors were located at 76.6°, 96.6°, and 116.6°, generating a small tilt of the overall distribution. This coding error was corrected in the other

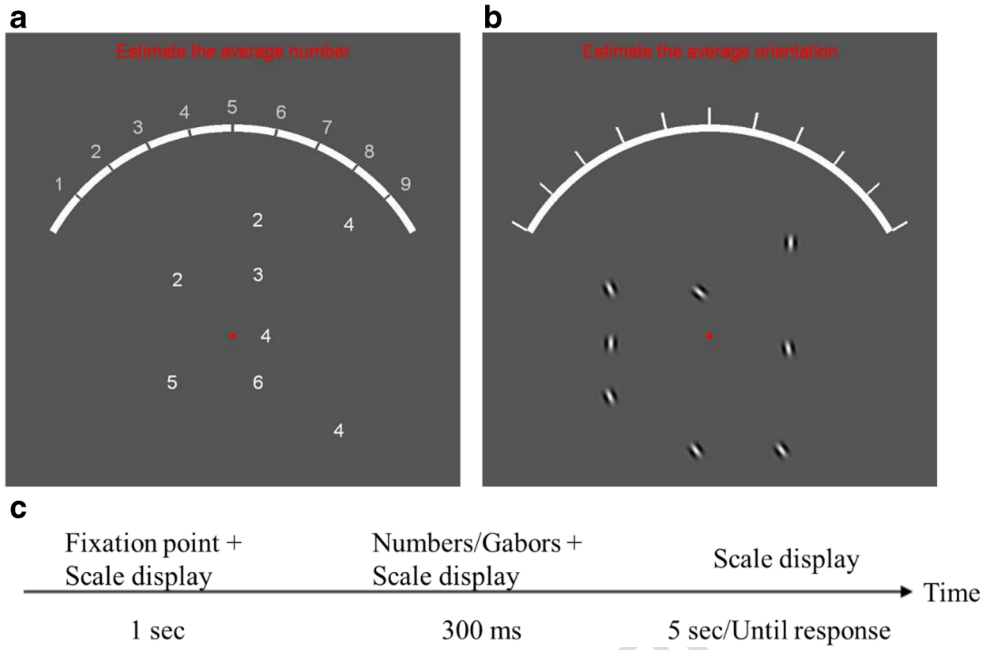

Fig. 1 Representative trial stimuli of each condition (set size 8). (**a**) Numbers condition. (**b**) Gabors condition. (**c**) Timeline diagram of a single trial. Each trial began with a fixation point for 1 s, followed by the array and ended with the response scale display. Trials end when the participant enters a response or after a 5-s deadline

participants. Since responses are made on a continuous scale and the actual deviation can be correctly extracted, all participants were included in the analysis.

Results

For simplicity and normalization between the two tasks we computed the different accuracy measurements in the orientation task after we transformed the orientation angles to numbers of 1–9, based on the mappings above.

Averaging precision

We used two measures to quantify participants' precision in the averaging tasks: First, we looked at the Pearson correlation across trials between the real and estimated averages of the array in each trial (see Fig. 2 for an example participant). The average correlation was high both for the orientation task (average r = .72, SD = 0.1) and for the number task (average r = .80, SD = 0.08). Note that in both tasks we observed regression to the mean, by which responses were biased towards the center of the scale. Second, we computed the root mean square deviation (RMSD) between the real averages and the participants' responses across trials (see Fig. 3a). To obtain a chance-level baseline for this measure, we evaluated the RMSD for randomly shuffled responses across trials, both for the orientation and the number tasks. We found the actual RMSD was significantly lower (more precise) than the shuffled version (orientation task: actual RMSD = 1.00, shuffled RMSD = 1.84, $t(17) = 16.6$, $p < .001$. number task: actual RMSD = 0.86, shuffled RMSD = 1.87, $t(17) = 25.9$, $p < .001$).

In order to test the main effect of set size and its interaction with task, we carried out a two-way repeated-measures ANOVA (set size × task) with RMSD as the dependent variable. There was a significant interaction between the effects of set size and task, $F(2,34) = 6.5$, $p<.01$. A separate ANOVA for each task revealed a significant set size effect for the number task, $F(2,34)=13.8$, $p < .001$, but not for the orientation task; $F(2,34) = 0.88$, $p = .68$. Post hoc comparison using Holm's test in the number task showed that RMSD was significantly lower (more precise) for four items than for eight and 12 items. In sum, in the number task participants were less accurate as set size increased, as opposed to the orientation task in which set size did not influence precision (see Fig. 3a).

Reaction times

To evaluate whether the decrease in performance with set size for numbers might be related to a potential speed-accuracy tradeoff, we also looked at response times (Fig. 3b). We repeated the same two-way repeated-measures ANOVA (set size × task) now with median response time (RT) as the dependent variable. There was a significant interaction between the effects of set size and task on RT, $F(2,34) = 7.7$, $p < .01$. A separate ANOVA for each task revealed a significant set size effect for the number task, $F(2,34) = 10.25$, $p < .001$, but not for the orientation task, $F(2,34) = 2.09$, $p = .13$. Post hoc comparison using Holm's test in the number task showed that

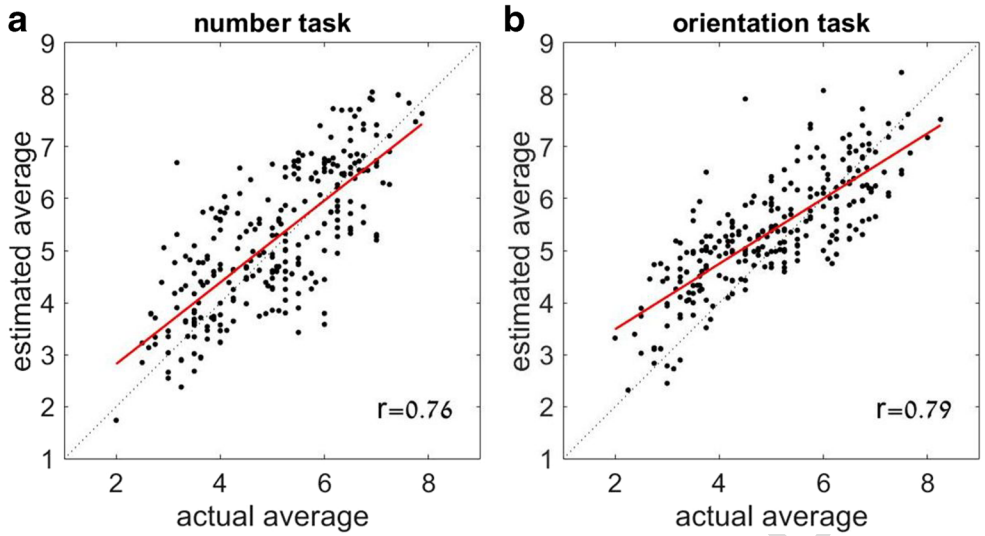

Fig. 2 (**a**) Correlation between the real average and the estimated average of a representative participant in the number task. (**b**) Correlation between the real average and the estimated average of a representative participant in the orientation task. In both panels, each dot corresponds to a single trial, and the red line represents the regression of the estimated average against the actual average across trials

four-items RT was significantly slower than eight- and 12-items RT (see Fig. 3b).

In order to understand if the slowdown in the number-averaging task at the set size of four can account for the improved precision in this condition, we computed for every participant the correlation between absolute errors (RMSD) and RTs across trials, separately for each set size and task. We reasoned that if such a speed-accuracy tradeoff occurred, then longer RTs would be associated with lower errors, resulting in negative correlations between errors and RTs. However, no negative correlations were found at the group level (see Fig. 1 in the Appendix for the distribution of correlation coefficients across participants). In particular, for the numerical averaging task, the correlations were close to zero (for set sizes four, eight, and 12 the mean r values were -.012, -.004, and .042, respectively, with SDs 0.17, 0.14, and 0.10, across participants). For the orientation averaging task we found small (but statistically significant) positive correlations at set sizes four (mean r = .092, SD = 0.15, t(17) = 2.59, p=.019) and set size 12 (mean r = .057, SD = 0.084, t(17) = 2.88, p=.010). To further discard the possibility of a speed-accuracy tradeoff for the set size of four in the number task, we eliminated the 20% slowest trials in that condition, so that the remaining trials had a median RT that was the same as the set size eight condition, and we examined RMSD in this RT-equivalent dataset. As expected from the null correlation between RT and absolute errors, this exclusion of slow trials did not affect the results regarding RMSD. Critically, the interaction between set size and task was maintained ($F(2, 34) = 6.13$, $p = .011$).

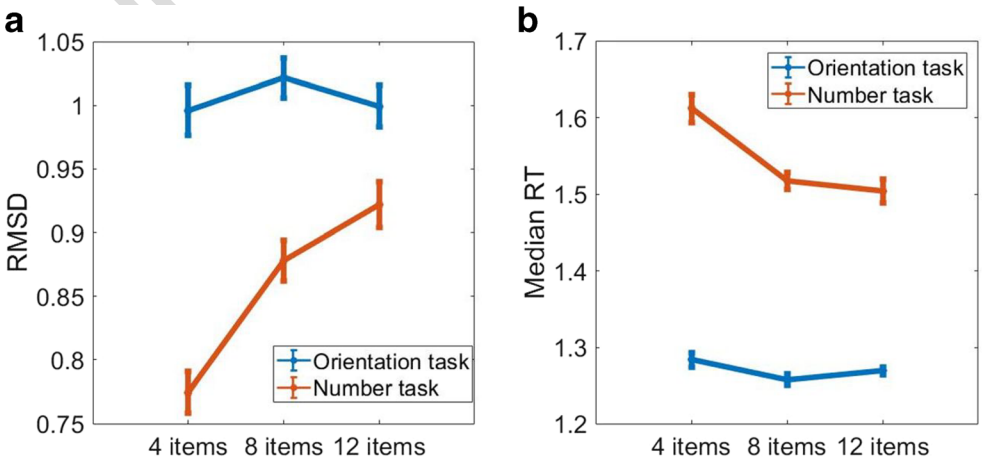

Fig. 3 (**a**) Root mean square deviation as a function of set size. In the orientation condition (blue) participants were not impacted by set size. In contrast, the number condition (red) shows that participants' performance deteriorated as set size increased. (**b**) Median response time (RT) as a function of set size. In the orientation task (blue) there was no difference in RT between the different set sizes. In the number task (red) responses were slower in the four-items condition compared to the eight- and 12-items conditions. In both panels, errors bars represent the mean and its standard error across participants

Summary and discussion

While the participants were able to carry out both tasks relatively well (as indicated by correlations between real and estimated values higher than .70), the precision of their estimation showed a different dependency on the set size of the array in the two tasks. For orientation-averaging, set size did not affect either the precision or the mean RT, suggesting a parallel process (Ariely, 2001; Chong & Treisman, 2005; Robitaille & Harris, 2011). For the numerical averaging on the other hand, both the precision and the RT decreased with set size. One possibility is that for small arrays (four digits), participants could have attempted to carry out the estimation by using a symbolic computation strategy, a strategy that they gave up on with larger arrays (Brezis et al., 2015). The null correlation between RMSD and RT observed in this condition indicates that this extra time did not help the participants to improve their estimation precision.

To conclude, we see that the ability of the participants to average larger arrays of numbers appears more limited, as the precision of the estimation is reduced with the size of the array. This is what would be expected if capacity (i.e., the number of elements the subjects can pool from) was reduced in the numerical task. In the next section we apply computational modeling in order to extract the capacity and attention parameters of the two tasks, and to examine additional biases, such as the weight given to in- or outlying elements (de Gardelle & Summerfield, 2011; Vandormael et al., 2017; Vanunu et al., 2019).

Computational analysis

We applied two computational models to account for the data across all trials and participants, in both tasks. The first model is a version of the limited-capacity (subsampling) model (Alik et al., 2013; Dakin et al., 2001; Solomon, May, & Tyler, 2016). This model assumes that out of N items presented, only M items are pooled up to generate the average-estimate. There are three sources of noise in this estimate. The first one is the sampling noise caused by subsampling (M out of N) elements. The second is an encoding noise, which is averaged out with M. The last component is a late-noise (this may include a motor component), which is not affected by M or N.

$$MeanEstimated = a + b\left(\frac{\sum_{i=1}^{M} x_i + \varepsilon_e}{M}\right) + \varepsilon_m, \quad \varepsilon_e \sim N\left(0, \sigma_e^2\right) \text{ and } \varepsilon_m \sim N\left(0, \sigma_m^2\right) \tag{1}$$

The model is summarized by Eq. 1, where M is the number of sampled items out of the array, x_i is the i^{th} item that was sampled, ε_e is the encoding noise, and ε_m is the motor noise. In this equation, a and b correspond to the intercept and slope parameters by which the internal estimation is mapped onto the external response-scale. Note that $b<1$ would induce a regression to the mean, which appears in the data (Fig. 2), and which is adaptive when observers face uncertainty but have prior knowledge about the distribution of the stimuli (Jazayeri & Shadlen, 2010; Anobile, Cicchini, & Burr, 2012).[2]

The second model is a version of the zoom lens model (Baek & Chong, 2020). The model assumes that while all visual elements contribute to the averaging estimation, they are subject to distributed attentional resources, which can vary from a sharp focus (for small arrays) to a broad one (for larger arrays). The precision of the processing is then in inverse proportion to the area of focus, similar to the zoom lens of a camera. As a result, the model assumes that an increase in set size leads to an increase in encoding noise for each item. There are also three sources of noise in this model. The first two are encoding noise and late noise, similar to the previous model. The third one is the attention parameter (A), which is a noise-reduction factor multiplied to encoding noise.

$$MeanEstimated = a + b\left(\frac{\sum_{i=1}^{n} x_i + \varepsilon_e}{n}\right) + \varepsilon_m, \quad \varepsilon_e \sim N\left(0, \sigma_e^2\right), \ \ \varepsilon_m \sim N\left(0, \sigma_m^2\right)s \tag{2}$$

$$\sigma = \sqrt{\frac{(n-1+A)^2}{n^3}\sigma_e^2 + \sigma_m^2} \tag{3}$$

This model is summarized by Eqs. 2 and 3, where x_i is the i^{th} item that in the array, n is the set size of the array, ε_e is the encoding noise, ε_m is the motor noise, A is the attention parameter, and a and b correspond to the intercept and slope parameters by which the internal estimation is mapped onto the external response-scale.

Since fitting five parameters is computationally challenging (from a model recovery perspective), we carried out the model fits in two steps. First, we conducted a simple regression predicting the trial-by-trial response of each participant from the sequence-average, to determine the a and b parameters for each participant. We then fixed those parameters and we fitted the three noise parameters, $M, \varepsilon_e, \varepsilon_m$, or $A, \varepsilon_e, \varepsilon_m$. For the zoom lens model, the predicted distribution of the estimated mean is Gaussian around the actual average and with a variance determined by Eq. 3. For the sampling model, we resorted to simulations. For each trial we computed the expected distribution of the estimated mean over the array, given the parameters of the model. In both models, from the predicted distribution (in each trial) we obtained the log-likelihood of the response of the observer in that trial. These log-likelihoods were accumulated across trials and the model parameters were optimized to maximize the total log-likelihood (see Tables 1 and 2 in the Appendix for parameters AIC/BIC) Fig. 4.

[2] We also carried out model fitting without the a,b parameters, but the results were less good in terms of AIC/BIC measures, and they provide similar conclusions. So we only report the model comparisons that include intercept and slopes parameters.

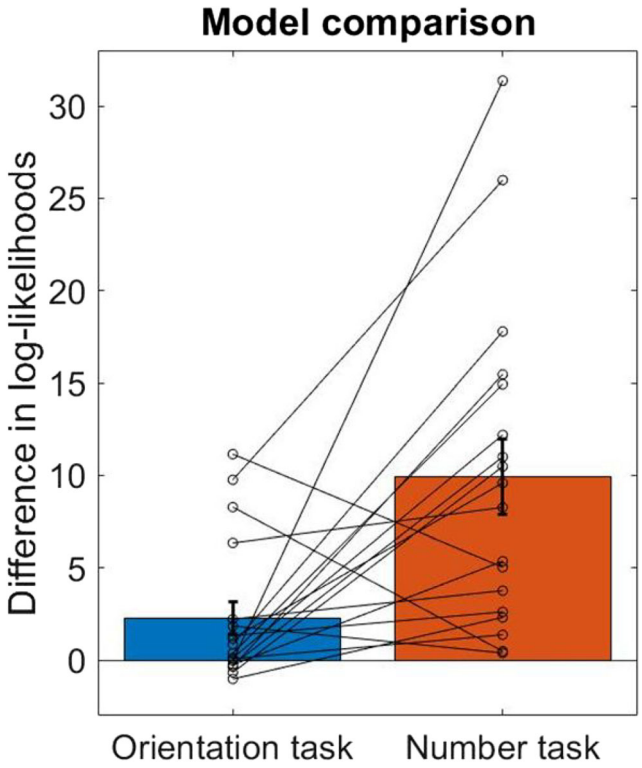

Fig. 4 The difference in log-likelihood between the zoom lens model and the sampling model as a function of task (orientation vs. numbers). Positive values indicate an advantage for the sampling model. Each dot is an individual observer. Error bars correspond to SEM

Model comparison

We compared the capacity/sampling and the zoom lens model to test which of them accounts better to the data in each task. The models have the same number of parameters so we compared directly the log-likelihoods. Figures 5 and 6 shows the difference in log-likelihood (zoom lens minus sampling model, such that positive values are in favor of the sampling model) in each task. As shown in the figure, there are very small differences in the model fits in the orientation task (except for four subjects out of 18), but there are large differences in the number task, where the sampling model fares significantly better (see Tables 1 and 2 in the Appendix for more details).

Interestingly, all the participants for whom the sampling model wins over the zoom lens model are those for whom the fitted value of the capacity parameter, k, was very small (2 or 3; see Tables 1 and 2). Besides, all the participants for whom the capacity parameter was k = 12 in the orientation task (maximum value), were those for whom the zoom lens model won (see Tables 1 and 2). We next focused on how the capacity parameters vary with the task (see Figs. 5 and 6; see Tables 1 and 2 in the Appendix for other parameters). Despite marked variability across individuals, we observed overall a higher capacity in the orientation task (M = 7.3, SD = 3.9) than in the number task (M = 3.6, SD = 1.8). The difference between the two tasks was statistically significant ($t(17) = 3.5$, $p < .005$). This result confirms our hypothesis that when constructing their representation of the average over a set of items, observers integrate more items in the orientation task than in the number task.

Weights of inlying versus outlying elements

Finally, we examined the weights that participants gave to the different elements in the array, depending on their relative rank (among all elements in the array) and depending on the task (i.e., number or orientation). In particular, we compared elements falling in the middle of the sample (hereafter inlying elements) versus elements at the extreme (hereafter outlying elements). For example, for a sequence such as (2, 3, 4 5, 5, 6, 7, 8), we considered that (2, 3, 7, 8) were outlying elements and that (4, 5, 5, 6) were inlying elements. For each task and set size, we then extracted the weights given to outlying and to inlying elements using the following linear regression (Eq. 2):

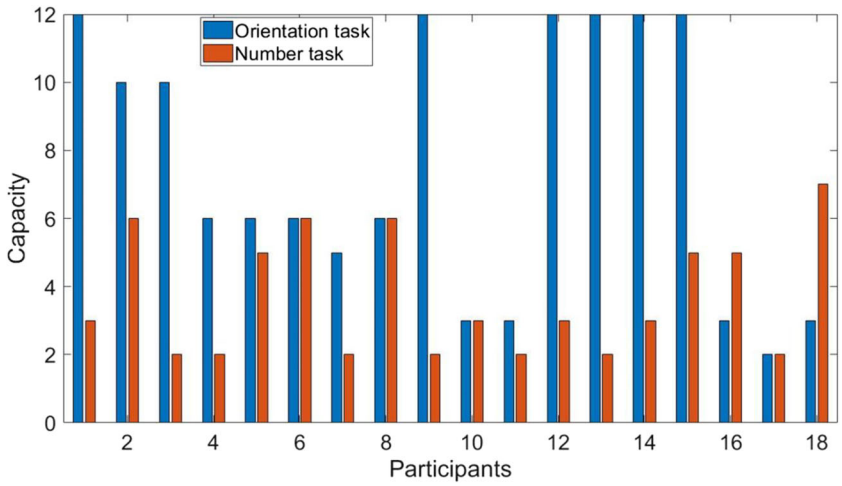

Fig. 5 Capacity-parameter M, for each of the participants in the two tasks

$$Response = \beta_0 + \beta_{in}\left(\frac{2}{n}\sum_{i\in In}X_i\right) + \beta_{out}\left(\frac{2}{n}\sum_{i\in Out}X_i\right)$$

with X_i the ordered samples, and $In = [\![\frac{n}{4}+1, \frac{3n}{4}]\!]$ and $Out = [\![1, \frac{n}{4}]\!] \cup [\![\frac{3n}{4}+1, n]\!]$ the indices for inlying and outlying elements, respectively.

We then examined how these weights varied across conditions (Figs. 5 and 6). A 2 × 3 × 2 ANOVA (task, set size, in/outliers) shows a significant triple interaction ($F(2, 34) = 4.48$, $p = .031$). We thus conducted separate ANOVAs for each task, to examine the effect of set size and element rank. In the number task, there was only a main effect of rank, $F(1, 17) = 11.20$, $p = .004$, in which participants gave more weights to the outlying elements, in a similar manner across all set sizes. By contrast, for the orientation task there was both a main effect of set size, $F(2, 34) = 23.82$, $p < .001$, and an interaction between set size and rank, $F(2,34) = 6.03$, $p = .011$. Further examination of this interaction indicated that inlying elements were down-weighted relative to outlying elements only for the largest sets (size 12: rank effect: $F(1,17) = 10.44$, $p = 0.005$) but not for smaller sets (sizes four and eight: both $p > .05$).

General discussion

We examined and compared the ability of observers to estimate the average number and the average orientation of elements presented simultaneously for a brief (300 ms) duration. Our experimental procedure required observers to make a response on a continuous scale, rather than a binary decision, and our results indicated that in both tasks the observers were able, despite the presence of a regression to the mean component, to make good estimations (see Fig. 2).

The critical difference between the two tasks was the impact of the set size on the precision with which the average was estimated. We expected that the perception of numerical symbols would depend more on attentional and visual working memory resources, compared with the perception of oriented elements, which can generate a more holistic (texture) process (Dakin, 2001; Chong & Treisman, 2005; Robitaille & Haris, 2011). We thus expected to find a higher capacity in the pooling of orientations compared with the pooling of numbers. These predictions were confirmed at the group level, using estimates of capacity based on a sampling model of averaging. In addition, we also compared this model to the (distributed attention) zoom lens model of averaging, which instead of sampling involved distributed attention over all elements (Baek & Chong, 2020b). While in the orientation task the zoom lens and the sampling models were about equal in their fit performance, in the numerical task the sampling model provided a better fit. Consistent with this, the estimation precision decreased with set size only in the numerical task and the extracted capacity parameter M was lower for the numerical task (average M = 3.7), compared to the orientation task (average M = 7.3).

In addition to these group differences, we also observed a large heterogeneity in both tasks. While some participants showed maximal capacity in the sampling model (M values that approached the maximum set size of 12) and RMSD decreasing with set size (as a result of efficient pooling), others showed low capacity (values of M = 2) and RMSD increasing with set size. This type of heterogeneity was previously reported for the orientation averaging (Solomon, May, & Tyler, 2016). One possibility discussed by Solomon et al. (2016) is that the efficiency may be a function of expertize with the task.

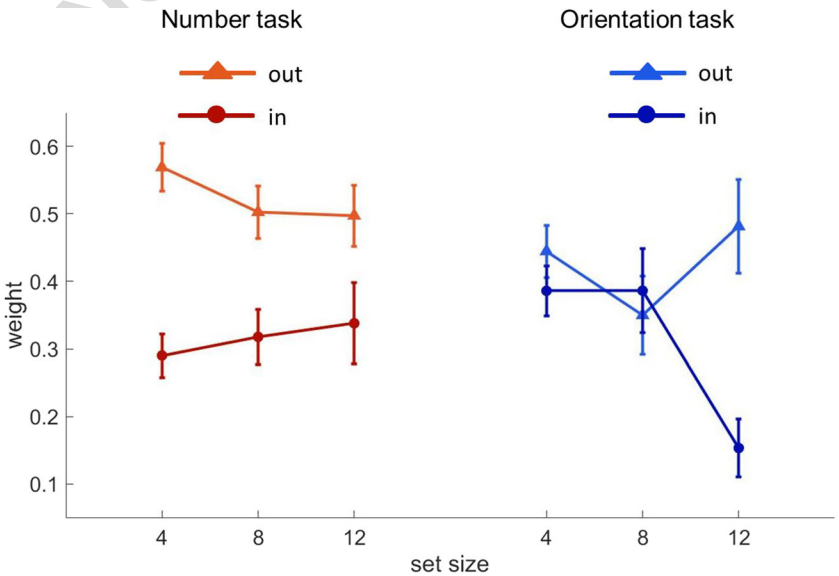

Fig. 6 Regression weights for inlying and outlying elements within each array, separately for the two tasks and the different set sizes. Error bars represents the mean and its standard error across participants

Our finding that orientation averaging is more efficient than averaging of symbolic numbers is consistent with this possibility: the visual system is arguably more expert in extracting orientations from Gabor patches than in extracting the quantity associated with a symbolic number. Could the inter-individual variability in efficiency observed in our data also relate to variations in expertise across participants? Unfortunately, we cannot address this question directly with our protocol, but there was room for variations in expertise across participants, given that the amount of training our participants received before engaging in the main experiment was minimal (360 trials per task). In the case of averaging of symbolic numbers in particular, one could further speculate that familiarity with mathematics (e.g., due to studies, or to work-related or other activities involving mental calculus) may affect the efficiency with which participants compute an average over a set of visually presented numbers. Future studies are needed to further investigate this issue.

Our capacity estimate for the orientation averaging task is somewhat higher that reported by Solomon et al. (2016) as well as in some other studies (see, e.g., Table 1 in Solomon et al., 2016). While as discussed above there was marked heterogeneity in both studies, there are two aspects in the experimental procedure that could account for potential differences. First, while Solomon et al. (2016) used stimuli presented on a circular array, in our experiment they were presented in a texture type display, and random spatial positions, which may enhance texture/grouping processes. Other studies that used texture displays have also indicated a capacity that exceeds the VWM of three to four items (Dakin, 2001; Robitaille & Harris, 2011). Second, we used a continuous response instead of a binary choice relative to a reference. Doing this may have eliminated some non-integration strategy to carry out the task, such as counting the elements higher than the reference. Future work might investigate these aspects.

The main focus of our study was the comparison of the capacity of the orientation and numerical averaging tasks. Regarding this comparison, we should acknowledge one potential limitation of our experimental methodology, in that we did not equate the visual characteristics of the stimuli between the number task and the orientation task. It is possible that the orientation stimuli may have benefited from a greater precision in terms of visual encoding than the number stimuli. Indeed, our number stimuli involved higher spatial frequency content (sharp edges), which may have been degraded towards the periphery of the stimulus display. Fortunately, our computational modeling allowed us to estimate encoding noise for both the number task and the orientation task, and it appears that irrespective of the model considered (sampling vs. zoom lens), this early noise was actually higher for the orientation task than in the number task (see Tables 1, 2, 3 and 4 in the Appendix Material), which we argue alleviates the concern. Further research may, however, better address this issue, by measuring the precision of the representation of single items, in addition to the averaging task.

The lower capacity in the numerical averaging task indicates that for most participants the estimation is based on sampling only a few of the elements. Based on previous work (de Gardelle & Summerfield, 2011; Vandormael et al., 2017; Vanunu et al., 2020), we sought to investigate which elements received more weight. The inlying/outlying analysis shown in Figs. 5 and 6 indicates that those elements are more likely to be extreme elements. Note that when a limited number of samples (say, two) can be used for the averaging process, the precision of the estimation is higher when the extreme ones are selected, compared with a random selection. Thus, if these extreme elements are easier to detect, relying on them could be an adaptive strategy. This interpretation is consistent with the fact that in the orientation task, the weight of the extreme samples exceeds the weight of the midrange samples, only at the largest set size (when the set size exceeds the capacity of the orientation-averaging estimation). While these results stand in contrast to those of Vandormael et al. (2017), who reported robust averaging (lower weights for extreme elements), they are consistent with those reported by Vanunu et al. (2020). We should note that these two studies used long presentation durations (several seconds in both cases) and a binary comparison with a reference, whereas our task involved brief displays and required an estimation on a continuous scale.

The results for the numerical averaging also stand in contrast to those reported in Brezis et al. (2015, 2016, 2018), in which the precision improved with set size, indicating pooling across all (or almost all elements, from four to 16). The critical difference, however, is that while in the present study, the elements are briefly displayed simultaneously, in Brezis et al. they were sequentially presented, resulting in less attentional resource competition between the encoding of the elements. This suggests a framework in which while the estimation mechanism is parallel (e.g., a neural population-coding model in Brezis et al., 2016, 2018), the encoding of the items has some serial (capacity limited) component that is lower for symbols compared to oriented lines.

Finally, in addition to capacity, we also examined response times for the two tasks. One interesting aspect was that RTs markedly increased when participants had to average four numbers, in comparison to eight or 12 numbers. Such an increase for four elements was specific to the number task, and did not occur in the orientation task. Thus, it might indicate that participants approached the number averaging task differently with four items compared to eight or 12 items, for instance by trying to calculate the average rather than by relying on an intuitive estimation. We note, however, that these longer response times did not lead to better responses. Whether this change in strategy was deliberate or not and whether it may reflect an adaptive strategy or not, however, remains to be addressed. Future studies may investigate, in particular,

whether participants have a good insight or not about their own cognitive processes in the averaging task.

Appendix

Model fitting

Optimization procedure. The free parameters of the sampling and zoom lens models were fitted to the data of each participant separately, using maximum likelihood estimation. We carried out the model fits in two steps. First, we carried out a simple regression predicting the trial by trial estimate of each subject from the sequence-average, to determine the *a* and *b* parameters for each subject. We then fixed those parameters and we constructed an n-dimensional grid (n is the number of free parameters for each model), with the four noise parameters (in total for the two models), M, ε_e, ε_m,A. M ranging from 1 to 12 with increments of 1, A ranging from 0 to 1 with increments of .0101, ε_e ranging from 0 to 1.9 with increments of 0.126 for the numbers task and ranging from 0 to 3 with increments of 0.2 for the orientation task and ε_m ranging from 0 to 1 with increments of 0.06 for the numbers task and ranging from 0 to 2 with increments of 0.13 for the orientation task. This grid was searched exhaustively, and for each set of parameters, θ_j, the likelihood was calculated based on a Gaussian probability distribution function:

$$L(\theta_j) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x_i-\mu_i}{\sigma}\right)^2}$$

where N is the number of trials, x_i is the subject's estimated average in each trial, μ_i is the predicted average by the model excluding noise, and σ is the standard deviation such that $\sigma = \sqrt{\sigma_e^2 + \sigma_m^2 + \sigma_M^2}$ for the efficiency model and $\sigma = \sqrt{(n-1+A)\frac{2}{n^3\sigma_e^2+\sigma_m^2}}$ for the Zoom lens model. We also carried out model fitting without the a,b parameters and compared the two fits.

Model selection. In order to evaluate the quantitative fits of the models, we used two methods: (1) Akaike Information Criterion (AIC; Akaike, 1973), and (2) Bayesian Information Criterion (BIC; Schwarz, 1978; Raftery, 1995). These selection criteria implement a trade-off between model goodness of fit and complexity by penalizing additional free parameters according to the following formulas:

$AIC = -2 \cdot LL + 2 \cdot k$

$BIC = -2 \cdot LL + k \cdot log(N)$

where *LL* is the log-likelihood for the best fitting parameters, *k* is the number of free parameters and *N* is the number of trials. AIC/BIC differences exceeding 10 are considered decisive evidence in favor of the model with the lower numerical values (Burnham & Anderson, 2002; Raftery 1995; see Tables 1 and 2 for parameters and BIC/AIC values).

Fig. 7

Table 1 Comparison between the log-likelihood of the sampling model compared to the zoom lens model in the orientation task. The other columns show the parameters' value in each fit

Subject	Sampling model						Zoom lens model			
	LL	a	b	M	Ɛe	Ɛm	LL	A	Ɛe	Ɛm
1	422.9	1.19	0.84	12	2.2	0.80	**422.8**	0.31	3.0	0.6
2	**263.9**	2.24	0.62	10	1.6	0.13	265.2	0.61	1.6	0.3
3	**442.6**	1.28	0.75	6	2.6	0.40	442.7	0.30	2.2	1.0
4	**375.0**	1.71	0.71	7	1.6	0.67	375.3	0.93	1.5	0.8
5	**350.4**	1.2	0.78	6	1.2	0.27	352.6	0.00	0.7	0.6
6	**361.2**	-0.03	0.98	6	1.4	0.00	361.3	0.00	1.6	0.4
7	**537.8**	-0.02	1.1	5	2.2	0.00	538.6	0.00	1.8	0.9
8	**287.8**	2.22	0.55	6	1	0.13	289.6	0.00	0.7	0.5
9	271.5	2.74	0.46	12	0	0.53	**271.1**	0.00	0.3	0.5
10	**446.5**	-0.3	1.07	3	1	0.13	456.3	0.00	0.9	0.8
11	**538.6**	2.32	0.56	3	1.6	0.00	539.8	0.00	0.1	1.1
12	563.4	-0.8	1.16	12	1.2	1.07	**563.2**	0.02	2.2	0.9
13	438.5	1.82	0.62	12	1.2	0.67	**438.2**	0.01	2.0	0.5
14	260.0	2.03	0.62	12	0.8	0.40	**259.0**	0.00	1.5	0.2
15	390.6	3.23	0.38	12	1.8	0.27	**389.9**	0.16	2.2	0.0
16	**433.9**	2.28	0.57	3	1	0.00	440.2	0.00	0.7	0.8
17	**544.2**	1.57	0.65	2	1	0.00	555.4	0.00	0.5	1.1
18	**491.8**	1.09	0.77	3	1.2	0.27	500.1	0.00	0.1	1.0

Table 2 Comparison between the log-likelihood of the sampling model compared to the zoom lens model in the numbers task

Subject	Sampling model						Zoom lens model			
	LL	A	b	M	Ɛe	Ɛm	LL	A	Ɛe	Ɛm
1	**239.7**	0.3	0.9	3	0.1	0.2	271.0	0.00	0.1	0.7
2	**337.8**	1.3	0.8	6	1.8	0.1	340.5	0.62	0.8	0.8
3	**332.5**	-0.1	1.0	2	0.3	0.3	344.7	0.00	0.1	0.9
4	**325.3**	0.6	0.9	2	0.1	0.3	340.3	0.00	0.8	0.8
5	**427.8**	1.0	0.8	5	0.5	0.7	431.6	0.00	0.1	0.8
6	**343.2**	1.1	0.8	6	1.3	0.0	344.5	0.07	0.6	0.6
7	**422.4**	0.9	0.8	2	0.4	0.1	440.2	0.00	0.3	0.8
8	**361.6**	1.3	0.7	6	1.3	0.2	362.0	0.21	0.8	0.6
9	**494.0**	1.0	0.8	2	0.6	0.3	509.5	0.00	0.1	1.0
10	**276.0**	0.3	1.0	3	0.1	0.1	301.9	0.00	0.1	0.6
11	**523.2**	0.5	1.0	2	0.3	0.7	532.8	0.41	1.2	1.0
12	**443.6**	0.6	0.9	3	0.8	0.4	448.9	0.00	0.6	0.8
13	**438.8**	1.0	0.8	2	0.4	0.1	449.8	0.00	0.7	0.8
14	**305.5**	1.1	0.8	3	0.4	0.1	307.8	0.00	0.8	0.5
15	**305.9**	1.1	0.9	5	0.9	0.2	316.3	0.00	0.1	0.6
16	**399.0**	0.9	0.8	5	0.5	0.6	407.3	0.00	0.7	0.7
17	**562.0**	1.6	0.7	2	0.8	0.6	567.0	0.00	0.1	1.2
18	**475.6**	1.7	0.6	7	1.1	0.7	476.1	1.00	0.3	0.9

The other columns show the parameters' value in each fit

t3.1 **Table 3** Comparison between the AIC and BIC parameters for the sampling model with the mapping parameters compared to without the mapping parameters in the orientation task. The other columns show the parameters' value in each fit

t3.2 Subject	Model with mapping parameters							Model without mapping parameters				
t3.3	BIC	AIC	a	b	M	Ɛe	Ɛm	BIC	AIC	M	Ɛe	Ɛm
t3.4 1	**875**	**856**	1.19	0.84	12	2.2	0.80	902	890	12	2.6	0.80
t3.5 2	**557**	**538**	2.24	0.62	10	1.6	0.13	735	724	7	1.4	0.67
t3.6 3	**915**	**895**	1.28	0.75	6	2.6	0.40	919	907	2	1.2	0.53
t3.7 4	**779**	**760**	1.71	0.71	7	1.6	0.67	820	809	6	1	0.93
t3.8 5	**730**	**711**	1.2	0.78	6	1.2	0.27	784	773	5	1.2	0.27
t3.9 6	752	**732**	-0.03	0.98	6	1.4	0.00	**750**	739	6	1.4	0.00
t3.10 7	**1,105**	**1,086**	-0.02	1.1	5	2.2	0.00	1166	1155	2	1	0.53
t3.11 8	**605**	**586**	2.22	0.55	6	1	0.13	851	839	6	1.6	0.13
t3.12 9	**572**	**553**	2.74	0.46	12	0	0.53	926	914	6	0	0.80
t3.13 10	922	**903**	-0.3	1.07	3	1	0.13	**920**	908	3	0.8	0.40
t3.14 11	**1,107**	**1,087**	2.32	0.56	3	1.6	0.00	1175	1163	3	1.6	0.53
t3.15 12	1,156	**1,137**	-0.8	1.16	12	1.2	1.07	**1,155**	1,143	12	1.8	0.93
t3.16 13	**906**	**887**	1.82	0.62	12	1.2	0.67	1003	991	11	0.4	0.93
t3.17 14	**549**	**530**	2.03	0.62	12	0.8	0.40	796	784	6	1.2	0.40
t3.18 15	**811**	**791**	3.23	0.38	12	1.8	0.27	1,084	1,073	12	0	1.07
t3.19 16	**897**	**878**	2.28	0.57	3	1	0.00	1,028	1,017	2	0.4	0.53
t3.20 17	**1,118**	**1,098**	1.57	0.65	2	1	0.00	1,161	1,149	2	1	0.40
t3.21 18	**1,013**	**994**	1.09	0.77	3	1.2	0.27	1,038	1,026	3	1.2	0.40

t4.1 **Table 4** Comparison between the AIC and BIC parameters for the sampling model with the mapping parameters compared to without the mapping parameters in the numbers task

t4.2 Subject	Model with mapping parameters							Model without mapping parameters				
t4.3	BIC	AIC	a	b	M	Ɛe	Ɛm	BIC	AIC	M	Ɛe	Ɛm
t4.4 1	**509**	**489**	0.30	0.92	3	0.13	0.20	506	494	3	0.38	0.00
t4.5 2	**705**	**686**	1.25	0.78	6	1.77	0.13	730	718	5	1.65	0.33
t4.6 3	**694**	**675**	-0.11	1.03	2	0.25	0.27	684	672	2	0.25	0.27
t4.7 4	**680**	**661**	0.55	0.94	2	0.13	0.27	681	669	2	0.38	0.07
t4.8 5	**885**	**866**	0.95	0.80	5	0.51	0.67	911	899	5	1.52	0.27
t4.9 6	**716**	**696**	1.13	0.79	6	1.27	0.00	775	763	5	1.27	0.13
t4.10 7	**874**	**855**	0.94	0.83	2	0.38	0.07	868	857	2	0.13	0.00
t4.11 8	**753**	**733**	1.34	0.69	6	1.27	0.20	862	851	11	1.01	0.67
t4.12 9	**1,018**	**998**	1.03	0.82	2	0.63	0.27	1,031	1019	2	0.25	0.53
t4.13 10	**581**	**562**	0.31	0.95	3	0.13	0.07	572	560	3	0.00	0.00
t4.14 11	**1,076**	**1,056**	0.54	0.99	2	0.25	0.67	1,112	1100	1	0.00	0.07
t4.15 12	**917**	**897**	0.60	0.89	3	0.76	0.40	917	905	3	0.89	0.33
t4.16 13	**907**	**888**	1.02	0.84	2	0.38	0.13	921	909	2	0.00	0.27
t4.17 14	**640**	**621**	1.08	0.83	3	0.38	0.13	713	701	3	0.00	0.33
t4.18 15	**641**	**622**	1.06	0.85	5	0.89	0.20	751	740	3	0.25	0.33
t4.19 16	**827**	**808**	0.88	0.78	5	0.51	0.60	872	861	3	0.38	0.53
t4.20 17	**1,153**	**1,134**	1.62	0.65	2	0.76	0.60	1,192	1180	2	1.27	0.13
t4.21 18	**981**	**961**	1.72	0.63	7	1.14	0.73	1,053	1042	5	0.51	0.93

The other columns show the parameters' value in each fit

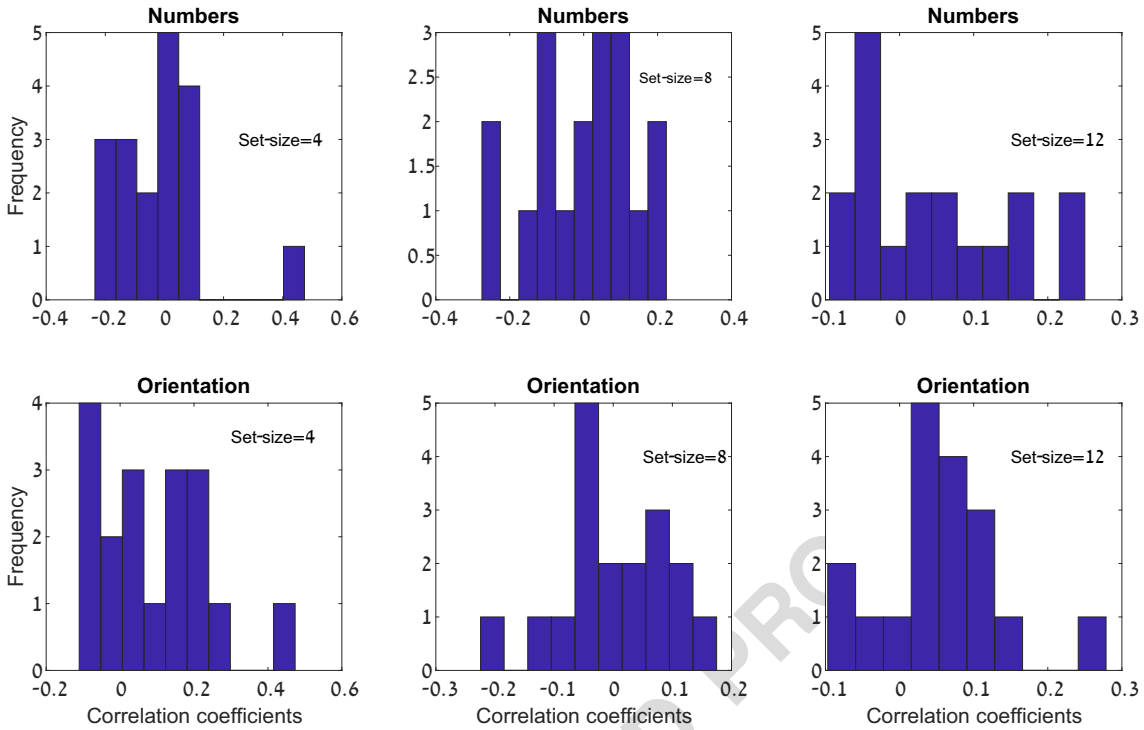

Fig. 7 Distribution of correlation coefficients across participants between root means square deviation (RMSD) and response time (RT)

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In BN, Petrov & F. Csaki (Eds.),Proceedings of the 2nd International Symposium on Information Theory (pp. 267-281). Budapest: Akademiai Kiado.

Anobile, G., Cicchini, G. M., & Burr, D. C. (2012). Linear mapping of numbers onto space requires attention. Cognition, 122(3), 454-459.

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. Psychological Science, 12(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Baek, J., & Chong, S. C. (2020). Distributed attention model of perceptual averaging. Attention, Perception, & Psychophysics, 82(1), 63-79.

Braun, J., & Sagi, D. (1991). Texture-based tasks are little affected by second tasks requiring peripheral or central attentive fixation. Perception, 20(4), 483-500.

Brezis, N., Bronfman, Z. Z., Jacoby, N., Lavidor, M., & Usher, M. (2016). Transcranial direct current stimulation over the parietal cortex improves approximate numerical averaging. Journal of Cognitive Neuroscience, 28(11), 1700-1713.

Brezis, N., Bronfman, Z. Z., & Usher, M. (2015). Adaptive Spontaneous Transitions between Two Mechanisms of Numerical Averaging. Nature Publishing Group, 1–11. https://doi.org/10.1038/srep10415

Brezis, N., Bronfman, Z. Z., & Usher, M. (2018). A perceptual-like population-coding mechanism of approximate numerical averaging. Neural Computation, 30(2), 428-446.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. Vision research, 43(4), 393-404.

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. Perception and Psychophysics, 67(1), 1–13. https://doi.org/10.3758/BF03195009

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and brain sciences, 24(1), 87-114.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. JOSA A, 18(5), 1016-1026.

De Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. Proceedings of the National Academy of Sciences of the United States of America, 108(32), 13341–13346. 10.1073/pnas.1104517108

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is Numerical Comparison Digital? Analogical and Symbolic Effects in Two-Digit Number Comparison. Journal of Experimental Psychology, 16(3), 626–641.

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.

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. Psychonomic Bulletin and Review, 18(5), 855–859. https://doi.org/10.3758/s13423-011-0125-6

Hess, R. F., & Field, D. J. (1999). Integration of contours: New insights. Trends in Cognitive Sciences, 3(12), 480–486.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. Nature neuroscience, 13(8), 1020.

Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. Journal of Vision, 18(9), 23-23

Kovács, I., & Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure–ground segmentation. Proceedings of the National Academy of Sciences of the United States of America, 90(16), 7495–7497

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390(6657), 279-281.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. Nature, 215(5109), 1519-1520.

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.

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. Journal of Vision, 11(12), 1–8. https://doi.org/10.1167/11.12.18

Sato, H., & Motoyoshi, I. (2020). Distinct strategies for estimating the temporal average of numerical and perceptual information. Vision Research, 174, 41-49.

Solomon, J. A., May, K. A., & Tyler, C. W. (2016). Inefficiency of orientation averaging: Evidence for hybrid serial/parallel temporal integration. Journal of vision, 16(1), 13-13.

Spitzer, B., Waschke, L., & Summerfield, C. (2017). Selective overweighting of larger magnitudes during noisy numerical comparison. Nature Human Behaviour, 1(8), 1-8.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. Cognitive psychology, 12(1), 97-136.

Vandormael, H., Herce, S., Balaguer, J., Li, V., & Summerfield, C. (2017). Robust sampling of decision information during perceptual choice. https://doi.org/10.1073/pnas.1613950114

Van Opstal, F., & Verguts, T. (2011). The origins of the numerical distance effect: the same–different task. Journal of Cognitive Psychology, 23(1), 112-120.

Vanunu, Y., Hotaling, M., & Newell, R. (2020). Elucidating the differential impact of extreme-outcomes in perceptual and preferential choice. Cognitive Psychology.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES.

Q1. Please check if the author group is presented correctly.
Q2. Ref. "Corbett, Oriet, & Rensink, 2006" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q3. Ref. "Nieder et al., 2002; Nieder & Miller, 2003" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q4. Ref. "Henik & Tzelgov, 1982" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q5. Ref. "Corbett et al. (2006)" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q6. Ref. "Alik et al., 2013" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q7. Ref. "Eriksen & StJames" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q8. The citation “Baek & Chong, 2020 a,b” has been changed to “Baek and Chong, 2020” to match the author name/date in the reference list. Please check if the change is fine in this occurrence and modify the subsequent occurrences, if necessary.
Q9. Ref. "Corbett et al., 2006" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q10. The citation “Arieli, 2001” has been changed to “Ariely, 2001” to match the author name/date in the reference list. Please check if the change is fine in this occurrence and modify the subsequent occurrences, if necessary.
Q11. Ref. "Vanunu et al., 2019" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q12. Ref. "Alik et al., 2013" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q13. Ref. "Dakin et al., 2001" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q14. Missing citation for Figure 4 was inserted here. Please check if appropriate. Otherwise, please provide citation for Figure 4. Note that the order of main citations of figures/tables in the text must be sequential.
Q15. Figure5 here hasbeen changed to Figures 5 and 6 . Kindly check if appropriate.
Q16. Ref. "Baek & Chong, 2020" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q17. Ref. "Schwarz, 1978" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.
Q18. Ref. "Raftery, 1995" is cited in the body but its bibliographic information is missing. Kindly provide its bibliographic information in the list.