



# The averaging of numerosities: A psychometric investigation of the mental line

Naama Katzin<sup>1</sup> · David Rosenbaum<sup>1</sup> · Marius Usher<sup>1</sup>

Accepted: 2 September 2020  
© The Psychonomic Society, Inc. 2020

## Abstract

Humans and animals are capable of estimating and discriminating nonsymbolic numerosities via mental representation of magnitudes—the approximate number system (ANS). There are two models of the ANS system, which are similar in their prediction in numerosity discrimination tasks. The log-Gaussian model, which assumes numerosities are represented on a compressed logarithmic scale, and the scalar variability model, which assumes numerosities are represented on a linear scale. In the first experiment of this paper, we contrasted these models using averaging of numerosities. We examined whether participants generate a compressed mean (i.e., geometric mean) or a linear mean when averaging two numerosities. Our results demonstrated that half of the participants are linear and half are compressed; however, in general, the compression is milder than a logarithmic compression. In Experiments 2 and 3, we examined averaging of numerosities in sequences larger than two. We found that averaging precision increases with sequence length. These results are in line with previous findings, suggesting a mechanism in which the estimate is generated by population averaging of the responses each stimulus generates on the numerosity representation.

**Keywords** Decision making · Summary statistics · Population coding · Mental number line · Approximate numerical system (ANS) · Numerosity representation

Research over the past two decades has converged on the idea that humans (including infants) and animals have at their disposal a set of analog numerosity (or magnitude) representations, which allows them to estimate and discriminate the numerosity of large sets of items (e.g., dots in a visual display or rapid sequences of sound clicks) without counting (Barth, Kanwisher, & Spelke, 2003; Dehaene, Dehaene-Lambertz, & Cohen, 1998; Feigenson, Dehaene, & Spelke, 2004; Gallistel & Gelman, 2000; Katzin, Salti, & Henik, 2018; Leibovich & Henik, 2014; Nieder, Freedman, & Miller, 2002; Nieder & Miller, 2003; Piazza, Pinel, Le Bihan, & Dehaene, 2007; Whalen, Gallistel, & Gelman, 1999). These numerosity representations, also labeled as the approximate number system (ANS), are akin to a “number line,” which is thought to be

mediated by broadly tuned numerosity detectors in the parietal cortex (Nieder et al., 2002; Nieder & Miller, 2003). The ANS representations account for data showing that humans (including infants) and animals are characterized by a Weber fraction in numerosity discrimination and estimation tasks (Barth et al., 2003; Cordes, Gelman, Gallistel, & Whalen, 2001; Whalen et al., 1999). Moreover, it has been suggested that the ANS representations are, at least partially, involved in the processing of symbolic numbers, as indicated by well-known distance and magnitude effects (Dehaene, Dupoux, & Mehler, 1990; Moyer & Landauer, 1967).

There are at present two versions of the ANS systems, which are roughly equivalent in terms of their prediction in numerosity discrimination tasks. The first is the log-Gaussian model, which assumes that the location of the number representation on the mental-line continuum is logarithmically compressed with a fixed variability (Dehaene, 2007; Feigenson et al., 2004). The second is the scalar variability model, which assumes that the representation of numerosities is linear (noncompressed) on the mental line, with regards to both its mean and its variability (Gallistel & Gelman, 2000), resulting in similar Weber-type predictions (Dehaene, 2007). Some extra support for the log-Gaussian model comes from

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.3758/s13414-020-02140-w>) contains supplementary material, which is available to authorized users.

---

✉ Naama Katzin  
[naamakatzin@gmail.com](mailto:naamakatzin@gmail.com)

<sup>1</sup> Tel-Aviv University, Tel Aviv, Israel

the mapping of numbers onto space (or a number line), which typically indicate compression (10 and 20 are more separated than 80 and 90 on the number line), especially for young children (Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003) or for adults under attentional load (Anobile, Cicchini, & Burr, 2012). However, the scalar variability model can also account for this compression under the assumption that participants rely on the central tendency principle (a type of regression to the mean), which affects larger numbers more than small numbers due to their increased encoding variability (Anobile et al., 2012).

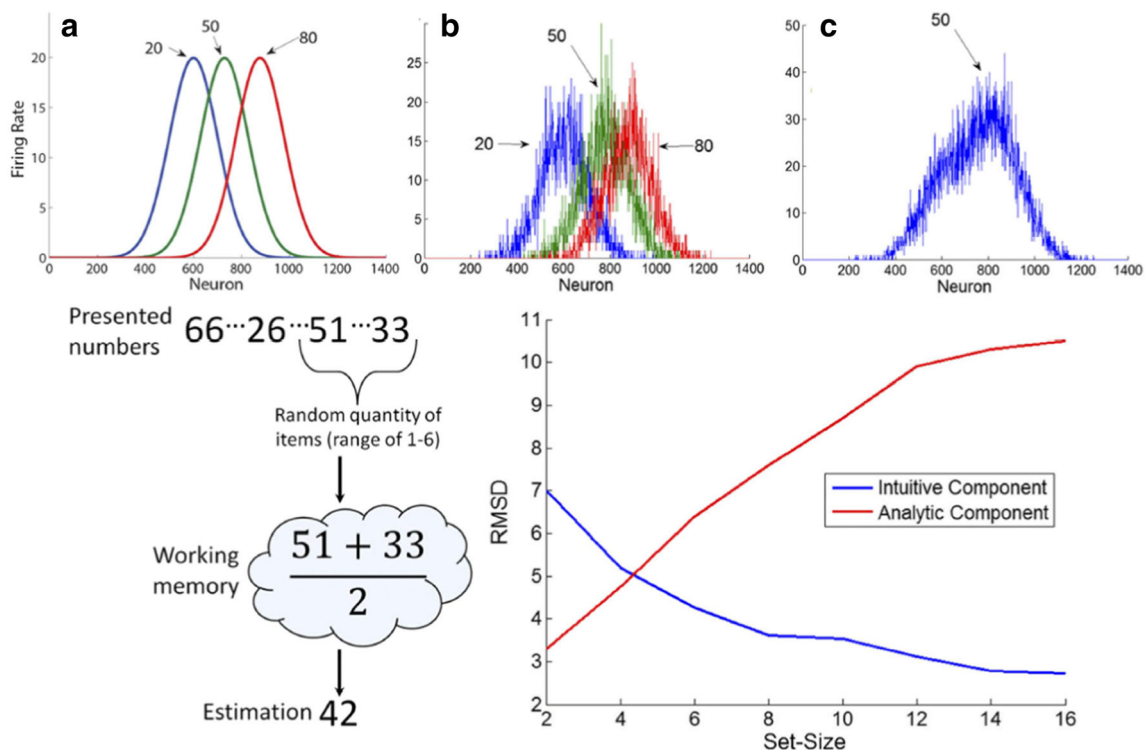
While most of the numerosity research has relied on estimation, discrimination, or comparison (same–different) tasks, some research has also targeted arithmetic operations, like addition and subtraction (Barth et al., 2006; Barth, La Mont, Lipton, & Spelke, 2005; Cordes, Gallistel, Gelman, & Latham, 2007; McCrink, Dehaene, & Dehaene-Lambertz, 2007; Pica, Lemer, Izard, & Dehaene, 2004). One area that was less explored within the ANS domain, however, is averaging. While, analytically, averaging can be viewed as being equivalent to addition (followed by division by the number of terms), there are reasons to believe that this is not the way that participants estimate the average of rapid sequences of (symbolic) numbers (Brezis, Bronfman, & Usher, 2015, 2018; Malmi & Samson, 1983; Mitrani-Rosenbaum, Glickman, & Usher, 2020), as they can provide accurate and rapid estimations of the average, even when they do not know the number of elements, or when elements in a specific range are to be discarded after the sequence presentation (Malmi & Samson, 1983). Rather, the evidence indicates that the estimation mechanism corresponds to a frequency-based estimation (the estimation of the center of mass of a noisy frequency distribution of the numbers), which is somewhat similar to the one suggested by the ANS representation system (see Brezis, Bronfman, Jacoby, Lavidor, & Usher, 2016; Brezis et al., 2015, 2018). In particular, Brezis et al. (2015, 2018) have proposed an ANS type of population code model, which accounts for a characteristic signature of the population code: Precision improves with the length of the sequence (see Fig. 1, blue line). While this is a straightforward prediction of population averaging (encoding noise in the representation of each number is averaged out), it contradicts the predictions of an *analytical* (and working-memory capacity limited) model, which computes via a sequential rule-based algorithm, predicting a decreasing precision with the number of terms (see Fig. 1, red line; Brezis et al., 2015). The latter is illustrated by the red line, which shows the prediction of the analytical model, which symbolically computes the average based on three random samples from a sequence of  $n$  numbers (the larger the  $n$ , the less the samples can approximate the true average).

The aim of this study is to probe the estimation of sequence-average with numerosity stimuli—sets of dots. This is important for several reasons. First the estimation of the average is critical for common life activities, like decision-

making, in which one has to estimate the utility of alternatives that vary across time or attributes (Betsch, Kaufmann, Lindow, Plessner, & Hoffmann, 2006; Brusovansky, Glickman, & Usher, 2018; Brusovansky, Vanunu, & Usher, 2017; Pleskac, Yu, Hopwood, & Liu, 2019; Roe, Busemeyer, & Townsend, 2001; Spitzer, Waschke, & Summerfield, 2017; Tsetsos, Chater, & Usher, 2012; Usher & McClelland, 2004; Vanunu, Pachur, & Usher, 2018; Zeigenfuse, Pleskac, & Liu, 2014). Second, recent research has indicated an impressive ability of human subjects in estimating summary statistics (in particular the average) of perceptual properties of sets of elements, such as size, orientation, and even emotional expression (Ariely, 2001; Chong & Treisman, 2005; Dakin, 2001; Haberman, Harp & Whitney, 2009; Haberman & Whitney, 2011; Khayat & Hochstein, 2018; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Robitaille & Harris, 2011). To our knowledge, there is less research on the averaging of (nonsymbolic) numerosities. While there is research on averaging of numerical (symbolic) numbers (Brezis et al., 2015, 2018; Spitzer et al., 2017; Vandormael, Castañón, Balaguer, Li, & Summerfield, 2017), testing averaging of nonsymbolic numbers has the added bonus of excluding symbolic computations, and thus exclusively targeting the ANS system.

In testing the averaging of numerosities, we wish to focus on two central questions: (i) Can we find evidence for systematic biases, which would confirm/disconfirm the presence of a compression mechanism in the number-line representation? (i.e., will some participants show a bias towards a geometric mean, as possibly suggested by a log-Gaussian model; see next section). (ii) Does the precision of the estimate increase (decrease) with the length of the sequence? An increased precision with sequence length would indicate that the ANS system can operate not only for single (or pairs of) stimuli but also for multiple ones, and that it can contribute to the mechanism for the formation of preferences over sequences of numerical values or payoffs (Brusovansky et al., 2018; Vanunu et al., 2018; Zeigenfuse et al., 2014).

Towards this aim, we carried out three experiments. The first experiment examined the averaging of pairs of numerosity stimuli. Here, we wanted to establish whether people can perform the task, by indication their estimate on a continuous mental line, and we examined potential compressive biases (in all the experiments, we quantified individual differences). In our second and third experiments, we examined sequences that vary in length from two to eight stimuli, and we focused on the estimation precision as a function of sequence length. The two experiments vary regarding the manipulation of the sequence length (randomized in Experiment 2 and blocked in Experiment 3), and regarding the response mode (on a continuous scale in Experiment 2, and based on comparison with a probe in Experiment 3). To anticipate our results, we find that whereas almost all participants were able to make good



**Fig. 1** The population code model of numerical averaging. Upper panels show the ANS-based averaging that operates on a set of analog broadly tuned numerosity detectors (**a**). Each stimulus generates a noise response on this representation (**b**), illustrated here for a sequence with three stimuli (20, 50, and 80). The responses are summed (**c**). The average is the center of mass of the distribution (see Brezis et al., 2015, for an explicit

mechanism), resulting in a precision that improves with sequence length (blue line in lower panel). An alternative symbolic-based computation results in precision that decreases with sequence length (red line). Reproduced with permission from Brezis et al. (2015). (Color figure online)

estimations, there are compressive biases in about half of them, but (except in one participant) those were milder than logarithmic. Critically, we find that precision improves with the length of the sequence, as predicted by the population code mechanism operating on ANS representations.

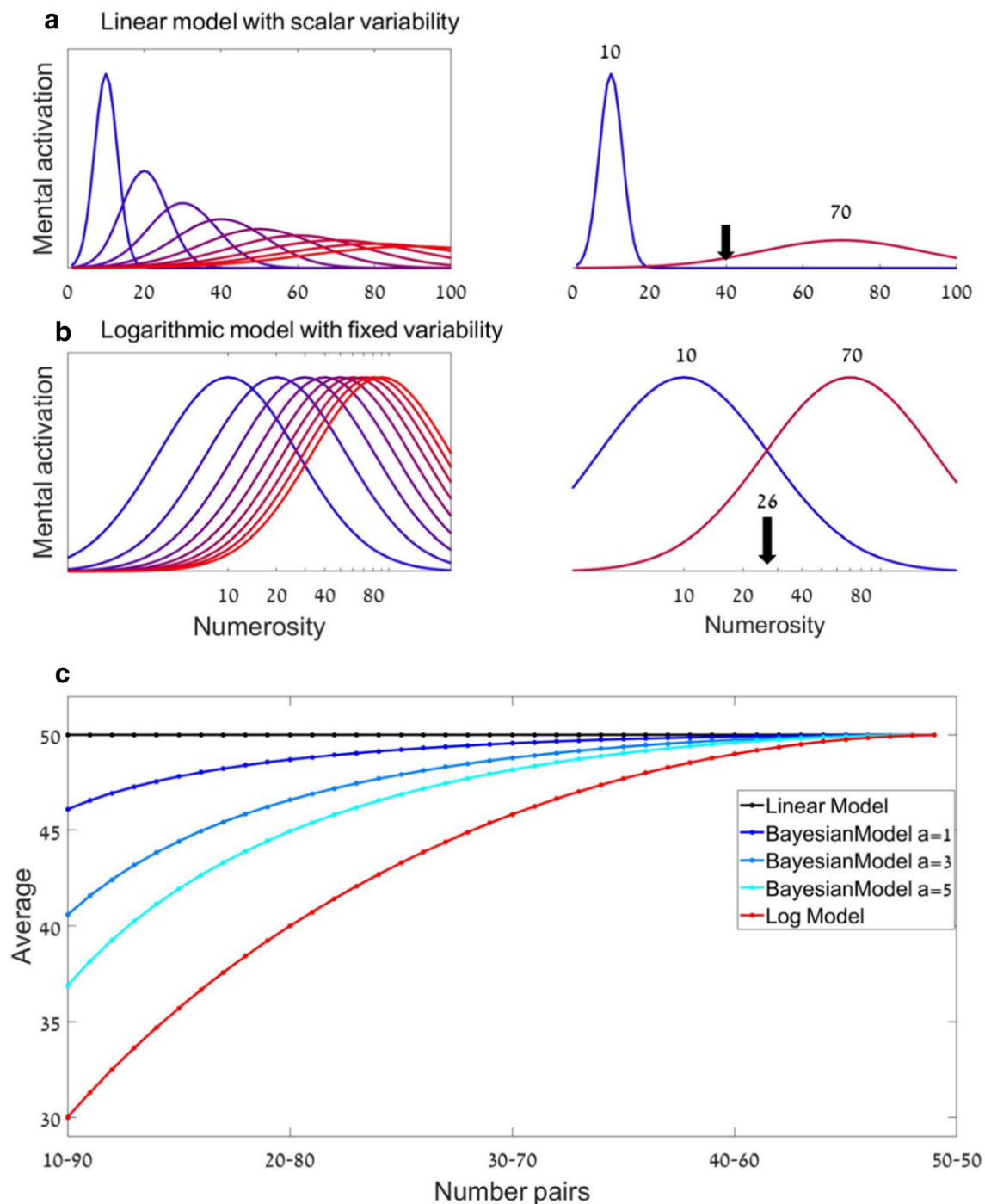
## Experiment 1

### Computational predictions and design motivation

Whereas the log-Gaussian and the scalar-variability models make similar predictions in discrimination tasks, they can potentially be distinguished in averaging. Figure 2 illustrates how this can happen. The left panels illustrate two extreme models (linear scalar variability vs. logarithmic compression; see also Feigenson et al., 2004), and the right panels show the responses following a pair of 10–70 numerosity stimuli. If the average is estimated by a population averaging over the same numerosity representation, one may expect, for the logarithmic (but not the linear model), an underestimation of the average that increases with the difference between them (see values pointed by the arrows in the right panel), which is analogous to the way risk aversion is generated as a result of a compressive utility

function (see Figs. 1–2 in Birnbaum, 2008). This is illustrated in Fig. 2a–b, with a simple sequence of two numerosities: 10, 70. By computing the center of mass (on the same ANS representation), we obtain the arithmetic mean (40) if the numbers are represented according to the scalar variability model, and the geometric mean (26.4) if the numbers are encoded based on the log-Gaussian model (see Fig. 2c, black and red lines, respectively, for an illustration over the range of 10–90). Interestingly, an intermediate degree of compression is obtained if the averaging in the scalar variability model is weighted by the variance of the number representation, as suggested by a combination of the scalar-variability model with a Bayesian framework that includes a prior (Anobile et al., 2012; see Appendix for computational details).

To test the sensitivity of the average estimate to the difference between the numerosities, we designed the stimuli to systematically manipulate the difference between the pairs of stimuli. This is illustrated in Fig. 3a, which illustrates the predictions for extreme ANS representation models (linear [black] and a logarithmic [red]; in the latter, we assumed that the estimation would correspond to the geometric mean) for pairs of numbers in the range 5–85 (each bin has a width of 10). As one can see, the difference is small close to the diagonal (when the numbers are similar), but increases with the



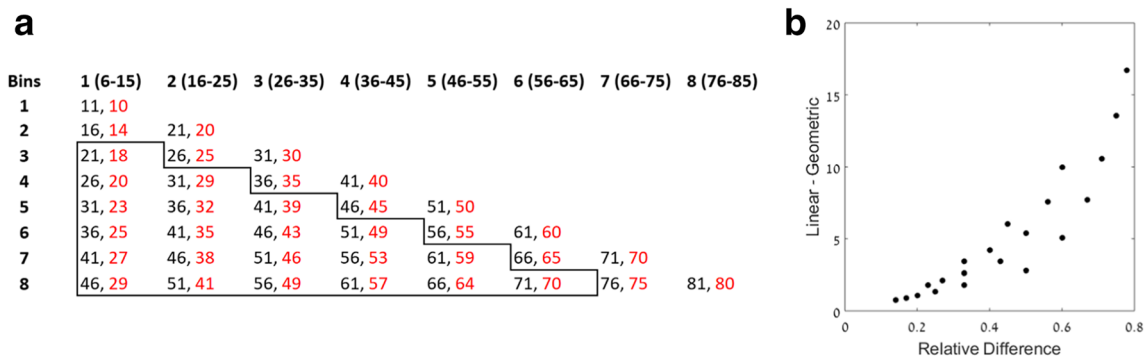
**Fig. 2** **a–b** The left panels illustrate the representation assumptions of two ANS models: (a) the scalar variability model and (b) the log-gaussian model. In both cases, we plot the neural response for a sequence of numbers 10–90 (in steps of 10); the red–blue continuum corresponds to the number magnitude. The right panels illustrate the averaging of numerosities 10 and 70 in each of the models. In both models, the average is the midpoint on the ANS, in the linear model it is the arithmetic average (40), and in the log-Gaussian model it is the geometric average (26.4). **c**

Averaging estimation for pairs of values (x-axis) that are symmetric around the midrange (50), for the scalar-variability model in which the average is the center of mass (black line), the log-Gaussian model (red line). The blue lines correspond to a Bayesian version of the scalar-variability model, in which each value is weighted in inverse proportion to its standard deviation, and a prior for the estimate is assumed to be a Gaussian centered at the middle of the range (50) with standard deviation labeled by  $a$  (see legend). (Color figure online)

difference of the pair (larger difference for bins 1–8; bottom left corner). In Fig. 3b, we illustrate how the difference

between the linear and the geometric average depends on the relative difference.





**Fig. 3** **a** The linear (black) and geometric (red) average of the various numerical bin pairs. The bins close to the diagonals were not sampled, as the difference between the averages is small (making the task too easy). **b** The difference between the linear and the geometric average depends on a

single variable, the relative difference,  $RD = \Delta(\text{bins}) / \text{Sum}(\text{bins})$ . The Spearman correlation between the RD and the delta of the averages is high ( $r_s = .96$ ,  $p < .001$ ). (Color figure online)

## Method

**Participants** Fifteen undergraduates from Tel-Aviv University ( $M_{\text{age}} = 22.4$  years,  $SD = 1.2$ ) participated in the experiment. Participants had normal or corrected-to-normal vision. Participants were awarded with course credit for their participation. All procedures and experimental protocols were approved by the ethics committee of the Psychology department of Tel Aviv University (Application 1-0000317). All experiments were carried out in accordance with the approved guidelines.

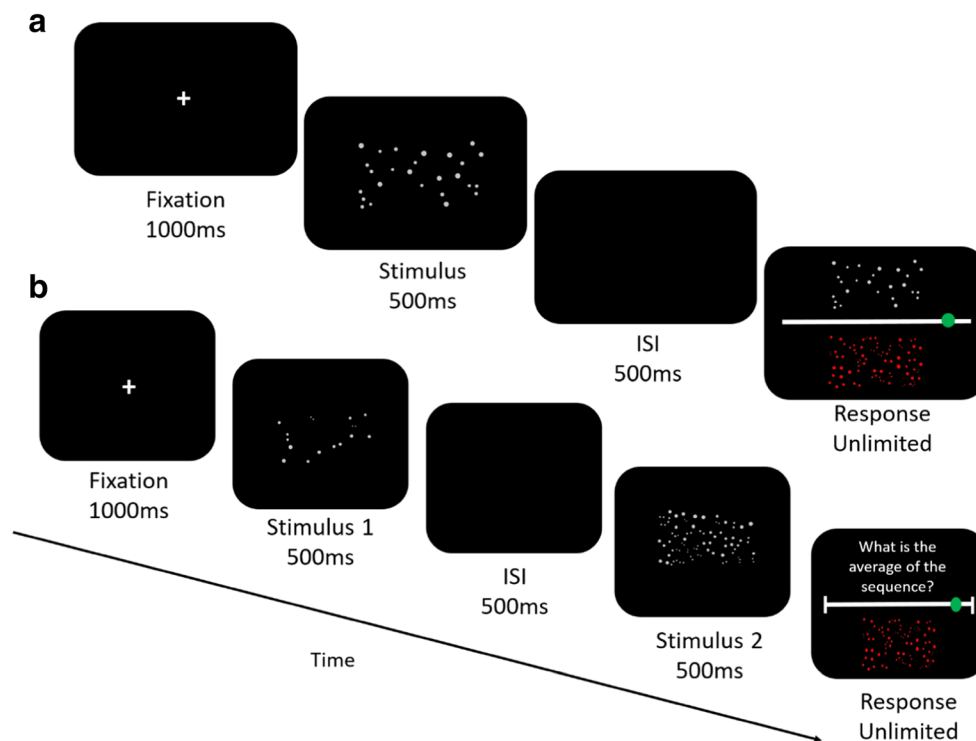
**Apparatus and stimuli** Stimuli consisted of dots randomly scattered on the screen. The dots diameter varied from 25 to 45 pixels. Minimum distance between dots was 25 pixels. The color of the dots was light grey for the sequence arrays (RGB: 201, 201, 201) and red (RGB: 255, 0, 0) for the scale arrays. The dots appeared on a black background.

**Training procedure and design** The experiment was built in OpenSesame. Before running the experiment, the participants received training with responding to a single numerosity stimulus using a continuous response scale. In order to assist them in doing so, the location of the mouse on the scale dynamically generated numerosity stimuli (in a different color from the one they estimated), which the subject could match to their mental representation of the stimulus (see Fig. 4a). In the training procedure, participants practiced the nonsymbolic number scale. Each trial began with a green fixation cross presented at the center of the screen for 1,000 ms. Next, the target, a cloud of white dots, was presented for 500 ms at the center of the screen. The numerosity of the target was between 5 and 90, in jumps of 5 (i.e., 5, 10, 15, 20, 25, ... 90), 18 target numerosities in all. After the target, a blank black screen appeared for 500 ms. Next, a response screen appeared, in which the target was presented at the top part of the screen, and at the center the nonsymbolic number scale. When participants pointed to the scale, beneath it, a red dot cloud appeared, with the numerosity of that location

on the scale. The scale ranged from 5 (left edge) to 90 (right edge) for 10 participants, and 5 (left edge) to 100 (right edge) for 5 participants.<sup>1</sup> In half of the trials, the starting point of the mouse was on the left edge, and in the other half, on the right edge. Participants were instructed to move the mouse until they find a red dot cloud that had the same (or as similar as possible) numerosity as the white dot cloud (participants were allowed and encouraged to move the mouse until they were satisfied of the match). Once participants clicked with the mouse, the trial ended and a new trial began. Each target numerosity appeared four times, 72 trials in all (see Fig. 4a).

**Averaging experiment procedure and design** As illustrated in Fig. 4b, each trial began with a white fixation cross presented at the center of the screen for 1,000 ms. After the fixation, two white dot clouds appeared, one after the other, each for 500 ms, with a blank interstimulus interval screen for 500 ms between them. After the stimuli, the response screen appeared. The response screen included a nonsymbolic number scale, which was the same as in the training procedure. Participants were instructed to move the mouse on the response scale until the red dot cloud matched the average numerosity of the stimuli. The numerosities of the stimuli were sampled from eight bins of 10 between 6 and 85 (i.e., Bin 1 = 6–15; Bin 2 = 16–25, ... , Bin 8 = 76–85; see Fig. 3a). In each trial, two bins with a distance ( $\Delta$ ) of at least two were sampled, corresponding to the area that is inside the encircled perimeter in Fig. 3a. For example, stimuli could include bins (1, 3), (1, 4), (2, 4); 21 combinations of bins in all. The experiment started with a practice block of 10 trials. Then, there were five experimental blocks with two repetitions of each bin combination, 210 trials in all, 42 trials per block.

<sup>1</sup> The maximum stimulus was 90. We wanted to ensure in those five participants that the compression is not due the maximum value on the scale being too close to the maximum stimulus. This difference can only affect the training; in the main experiment, the averages varied in the range of [10–75] and thus are distant from the minimum or maximum scale value (5/90).

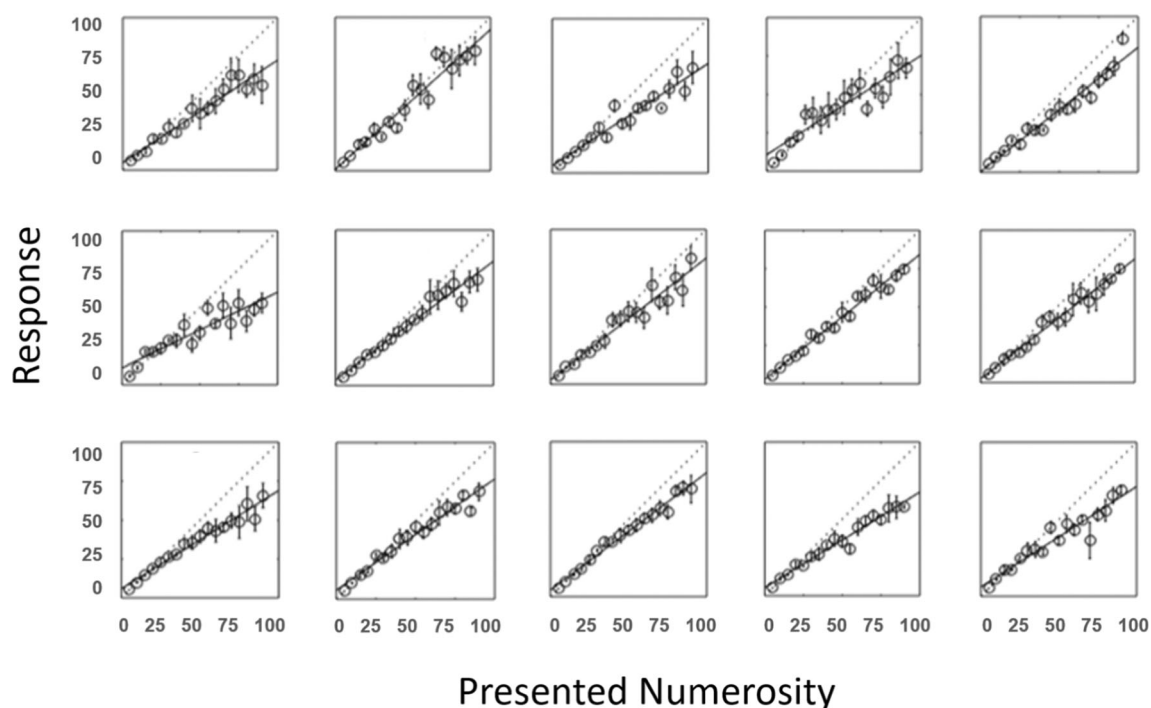


**Fig. 4** Examples of a trial in the training procedure (a) and in the averaging experiment (b)

## Results

**Training data analysis** For each participant, we plotted the response (averaging across trials with the same stimulus) as a function of the stimulus numerosity. In Fig. 5 we show the

response of each participant (averaged over trials) as a function of the stimulus numerosity. As one can see, the participants are able to use the continuous scale to indicate their impression of the stimulus numerosity (Pearson correlations for each participant between the presented numerosity and the



**Fig. 5** Correlation between presented numerosity and participant's response for all participants (participants are ordered based on their sensitivity to the difference between the numerosity; see Fig. 7)

participant's response was high.  $r = .97$   $SD = .02$ ). The fitted linear slope was on average  $b = .71$  ( $SD = .1$ ).

The purpose of the training task was twofold. First, we wanted participants to become familiar with the nonsymbolic number scale. Second, it enabled us to calibrate participants' responses. Accordingly, we performed a regression analysis with the presented numerosity as the dependent variable, and participant's response as the independent variable. We examined which fit was better: linear ( $y = b \times x + a$ ) or power ( $y = b \times x^\alpha$ ) fit. Both AIC and BIC parameters were lower for the power fit: AIC,  $t(14) = -5.55$ ,  $p < .001$ ; BIC  $t(14) = -5.55$ ,  $p < .001$ ; see Table S1 in the Supplementary Materials), with a compression exponent (average  $\alpha = .82$ ,  $SD = .09$ . The lowest  $\alpha$  was .68, and four participants had an  $\alpha$  larger than .9). This indicates that despite the presence of the red stimulus, there was a small tendency to underestimate the numerosities.<sup>2</sup> Based on this calibration, we can transform each response,  $y$ , into the experienced stimuli ( $x$ ) by inverting the  $y(x)$  function. The analysis in the main averaging experiment were carried out both with and without this calibration.

**Averaging experiment data analysis** Analyses were performed both with participants' raw responses and with their responses scaled according to the fit found in the training procedure. Results were similar for both, so here we report the results with the raw responses.

To see how the estimates vary along the  $x_1$ – $x_2$  continuum, we carried out a regression, in which we predicted the response based on three predictors: (i) the arithmetic average  $(x_1 + x_2) / 2$ , (ii) the difference  $|x_1 - x_2|$  (this corresponds to  $10 \times$  the  $\Delta$  of the bins), and (iii) a subject-dependent intercept. The second predictor allows us to test the presence of a compression in the representation of the numerosities. As illustrated in Fig. 3, the linear average is not affected by the  $\Delta$  of the bins. For example, in Fig. 3a, the  $\Delta$  of pairs of bins 8–1, 7–2, 6–3, are 7, 5, and 3, respectively. While the  $\Delta$  for these cells varies, their linear averages are all equal (46). An average based on compressed representations predicts that the estimate decreases with the  $\Delta$  of the bins. For all participants, the linear average was a significant predictor (average  $b = .57$ ,  $ps < .001$ ), as illustrated in Fig. 6.

The delta variable, on the other hand, was only significant for seven participants (average  $b = -.14$ , all  $ps < .05$ ). For the other eight participants, the delta coefficient was not significant (average  $b = -.01$ , all  $ps = ns$ ; see Fig. 7).

Next, we contrasted the linear and the logarithmic representations regarding their expected response biases. If a participant relies on linear (noncompressed) numerosity representations, the deviation between the estimate and the arithmetic average should not correlate with the relative difference;

however, the deviation between the estimate and the geometric average should correlate with the relative difference. The converse should happen if a participant relies on logarithmically compressed representations. For each participant, we compared two correlations: (i) The correlation between the relative difference and the linear average minus the subject's estimate, and (ii) the correlation between the relative difference and subject's estimate minus the geometric average.

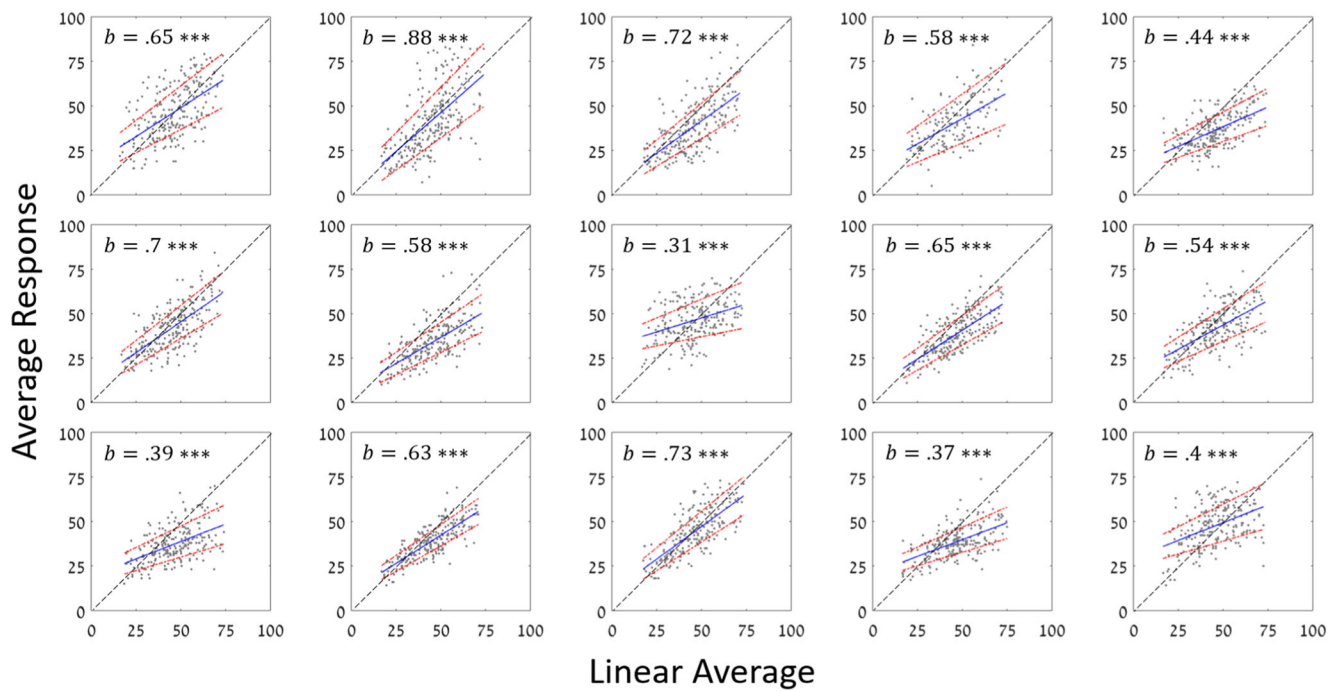
For all participants except one, we found a stronger positive correlation between RD and their response minus the geometric average (mean  $r = .52$ ,  $SD = .12$ , all  $ps < .001$ ), compared with the correlation between their response minus the linear average (mean  $r = -.19$ ,  $SD = .13$ ), suggesting that their responses are more linear than geometric. Only one participant displayed a geometric-average pattern (Subject 2 in Fig. 7, top row, second panel from the left), a more positive correlation with the linear average minus the response ( $r = .22$ ,  $p < .001$ , compared with  $r = .09$ ), suggesting this participant is more geometric than linear. This participant also displayed a compressed pattern in the previous analysis. The rest of the participants that displayed a compressed pattern in the previous analysis were more linear than geometric in this analysis, suggesting that their compression is not as strong as a logarithmic compression.

## Discussion

We examined the ability of participants to estimate the average of two numerosity stimuli by moving a mouse on a continuous response line. To facilitate the participants with the use of the scale, they first received training with single stimuli. In addition, the location of the mouse on the scale dynamically created a numerosity stimuli (in a different color; see Fig. 4), which the participant can compare with their mental estimate. For all the participants, the average estimates increase with the average of the stimuli pair; however, we also observe a (lower than 1) slope indicating the presence of regression to the mean (see Fig. 6). Since the task is not easy, the presence of regression to the mean is a normative way to deal with uncertainty (Anobile et al., 2012; Jazayeri & Shadlen, 2010). The results demonstrate that averaging is an operation that participants are able to carry out with a pair of numerosity stimuli.

The central question of this study was whether there are systematic deviations from the linear average, which are induced by the compression of the ANS representation. To examine this, we examined the dependency of the estimate on the difference between the two numerosities. For about half of the participants, such a dependency was found: The estimate decreased with the difference when the average was controlled for (akin to the phenomenon of risk aversion that would make a person prefer a lottery of 40 with  $p = .5$ , 60 with  $p = .5$ , to one of 10 with  $p = .5$ , 90 with  $p = .5$ . For the other half of the participants, the estimates were quite flat (with the difference)

<sup>2</sup> The five participants who had the response scale extended (from 90 to 100) did not vary in terms of compression from the others, average  $\alpha = .88$ .

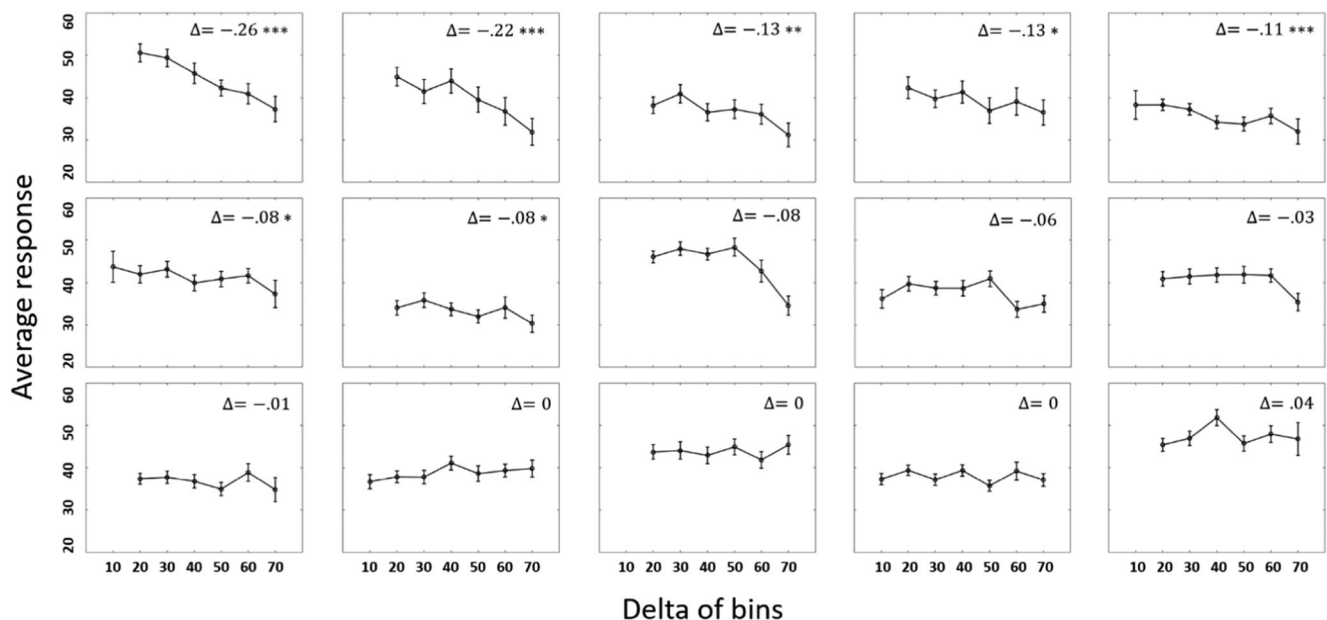


**Fig. 6** Participants' response as a function of the linear average of the stimuli presented. The blue line is the regression line, the red lines are the confidence interval, and the dashed black line is the identity line. Values

in each panel correspond to the regression coefficient of the linear average.  $*p < .05$ .  $**p < .01$ .  $***p < .001$ ; participants are ordered as in Figs. 5 and 7. (Color figure online)

supporting noncompressed numerosity representations (those results were obtained using the raw data, but the results are similar if we use transformed values based on the training calibrations). When contrasting between the linear and logarithmic compression, in particular, we found that only one participant for which the estimates were closer to the

geometric (than linear) average. This indicates that the compression that we have in the other subjects is milder than logarithmic. While we focused here on a binary contrast between the log-Gaussian and the linear (scalar variability) representation of numerosities, this binary (compression/no-compression) contrast is a simplification. As we have shown in



**Fig. 7** Participants' response by the delta of bins; vertical lines are within-subject standard error. Participants are ordered by delta coefficient. Values in each panel correspond to the regression coefficient of  $\Delta(x)$ .  $*p < .05$ .  $**p < .01$ .  $***p < .001$



Fig. 2c (blue lines), a milder compression can be obtained if, as previously suggested by Anobile et al. (2012), for the case of the number-line estimation of single numerosity stimuli, the participants (in our case with two stimuli) weight up the values of the two samples and the prior, based on their relative representational uncertainty. As the uncertainty is larger for the higher numerosity, it results in a milder compression effect (see Fig. 2c, blue lines) whose magnitude depends on the prior-variance parameter. Thus, differences in how the representational uncertainty increases with numerosity can account for the compression variability in our task.

## Experiment 2

The aim of the next two experiments is twofold. First, we wanted to expand the task from pairs of stimuli to longer sequences: two, four, or eight. Second, we aimed to test the ANS population-coding prediction that the precision should improve with sequence length. In these experiments, we did not manipulate the variance of the sequences, so our focus is not on compression biases, but rather on how the precision of the estimate varies with the length of the sequence. We do examine, however, another type of bias: temporal biases (do people give more weight to recent or earlier stimuli?).

### Method

**Participants** Twenty undergraduates from Tel-Aviv University ( $M_{\text{age}} = 23.15$ ,  $SD = 2.35$ ) participated in the experiment. Participants had normal or corrected-to-normal vision. Participants were awarded with course credit for their participation. All procedures and experimental protocols were approved by the ethics committee of the Psychology department of Tel Aviv University (Application 1-0000317). All

experiments were carried out in accordance with the approved guidelines.

**Apparatus and stimuli** Stimuli consisted of dots randomly scattered on the screen. The dots diameter varied from 25 to 45 pixels. Minimum distance between dots was 25 pixels. The color of the dots was light grey for the sequence arrays (RGB: 201, 201, 201) and red for the scale (RGB: 255, 0, 0) for the probes. The dots appeared on a black background.

**Training procedure and design** The experiment was built in MATLAB R2015. This procedure was similar to the training procedure in Experiment 1, except that the starting point of the mouse was the middle of the scale.

**Averaging experiment procedure and design** Each trial began with a white fixation cross presented at the center of the screen for 250 ms. After the fixation, a sequence of two, four, or eight dot clouds appeared one after the other, each for 500 ms. Once the sequence terminated, the participants were instructed to estimate the sequence's mean value by choosing a red dot cloud on the nonsymbolic number scale that represents the mean (see Fig. 8). Sequence length was randomized. The sequences were randomly drawn from three uniform distributions with a range of 30 and means 35, 50, and 65. Participants underwent 306 experimental trials divided into six blocks. Accordingly, there were 102 trials for each sequence length (two, four, eight), and within each sequence length 34 trials for each average (35, 50, 65).

### Results

**Training procedure** To verify that participants were correctly performing the task, we calculated Pearson correlations for each participant between the presented numerosity and the

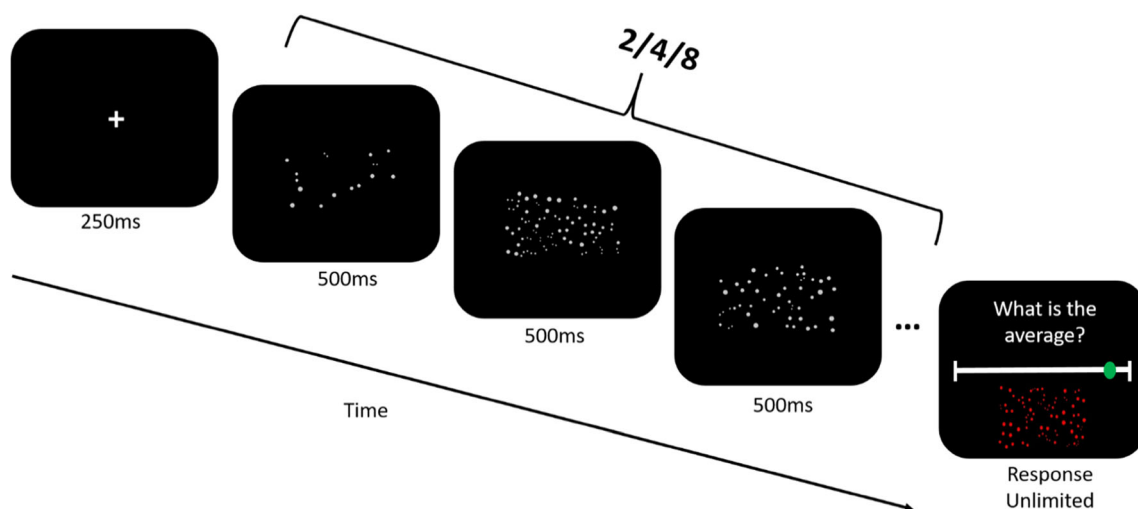


Fig. 8 An example of a trial in Experiment 2

participant's response. The average correlation was very high ( $r = .97$ ,  $SD = .03$ ). Like Experiment 1, we performed a regression analysis, with the presented numerosity as the dependent variable and participant's response as the independent variable. We examined which fit was better: linear ( $y = b \times x + a$ ) or power ( $y = b \times x^\alpha$ ). Both AIC and BIC parameters were lower for the power fit: AIC,  $t(19) = -3.37$ ,  $p < .01$ ; BIC,  $t(19) = -3.37$ ,  $p < .01$ ; see Table S3 in the Supplementary Materials). The average  $\alpha$  was 0.96 ( $SD = .11$ ).

**Averaging experiment** All analyses were performed both with participants' raw responses, and with their responses scaled according to the fit found in the training procedure. Results were similar for both, so here we report the results with the raw responses.

**Averaging precision** We quantified participants averaging precision using two measures. First, we calculated the Pearson correlation between the real and estimated averages of the sequences. The average Pearson correlation was high ( $r = .66$ ,  $SD = 0.12$ ) and was significantly higher than zero (all  $ps < .001$ ; see Fig. 9a for an individual participant). Second, we computed the root mean square deviation (RMSD) between the real averages and the participants' responses (note that higher values of RMSD imply lower accuracy). The RMSD was significantly lower than the simulated RMSD generated by randomly shuffling participant's responses across trials ( $RMSD = 11.92$ ; shuffled  $RMSD = 22.18$ ),  $t(19) = -21.78$ ,  $p < .001$ .

A repeated-measures ANOVA, with RMSD as the dependent variable and sequence length as the independent variable, was carried out. As illustrated in Fig. 9b, the main effect of sequence length was significant,  $F(2, 38) = 48.28$ ,  $p < .001$ ,  $\eta_p^2 = .71$ . Further analysis revealed a significant linear trend,  $t(38) = -9.83$ ,  $p < .001$ . A repeated-measures ANOVA, with

RTs as the dependent variable and sequence length as the independent variable, was also carried out. The main effect of sequence length was not significant,  $F(2, 38) < 1$ ,  $p = ns$ , indicating that this improvement is not the result of a speed-accuracy trade-off.

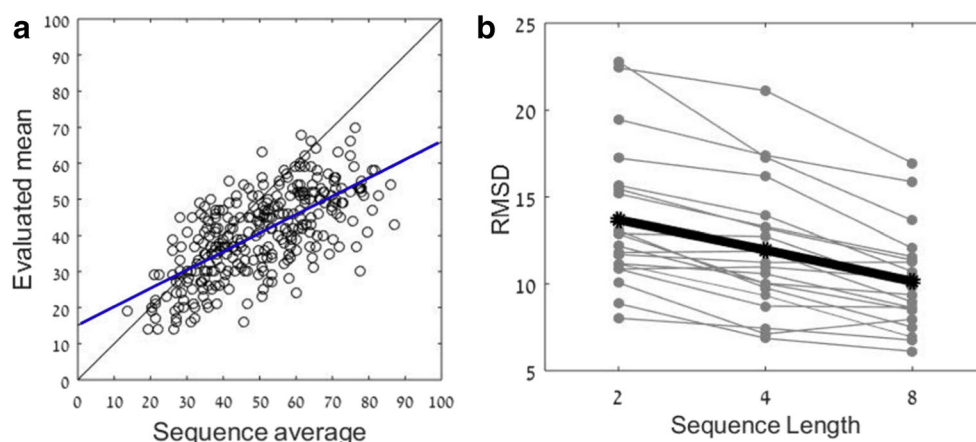
**Regression to the mean and temporal biases** To estimate these biases, we predicted the estimates based on two models. We compared a model without a temporal bias and a model with a temporal bias. In both models, we included the possibility of regression to the mean (corresponding to a slope parameter,  $b$ ) and a subject dependent intercept,  $a$ . In the non-temporal bias model, we carried out a linear regression, with participant's response as the dependent variable and the real average of the sequence as the independent variable:

$$\hat{y} = a + b (\text{sequence average}). \quad (1)$$

For the temporal bias model, we regressed the participant's response on the leaky integrated value with a leak parameter ( $\lambda > 1$  indicates a primacy effect,  $\lambda < 1$  indicates a recency effect, and  $\lambda = 1$  means no temporal bias):

$$\hat{y} = a + b \sum_{i=1}^n x_i * \lambda^{n-i}. \quad (2)$$

As shown in Table 1, the participants varied in their temporal bias. For five participants, a model without a temporal bias is better in both AIC and BIC measures. Indeed, the average lambda of these participants in the temporal bias model is 1.03. For six participants, a temporal bias model is better in both AIC and BIC measures, and their average lambda is 0.73. The rest of the participants (10 out of 15) have an average lambda of 0.83. In all these participants, the leak improves the log-likelihood of the fit. In nine of them, the complexity costs given by



**Fig. 9** **a** A typical individual participant's scatter plot of evaluated mean versus sequence average,  $r = .67$ ; see [Supplementary Materials](#) for the data of all participants. **b** RMSD as a function of sequence length. The grey lines mark individual participants; the black line is the average

**Table 1** Parameters for linear and temporal bias fit

| Subject        | Linear      | Leak        |             | $\Delta$ AIC  | $\Delta$ BIC  |
|----------------|-------------|-------------|-------------|---------------|---------------|
|                | <i>b</i>    | <i>b</i>    | $\lambda$   | Linear - leak | Linear - leak |
| 1              | 0.58        | 0.58        | 0.99        | -1.9          | -5.7          |
| 2              | 0.28        | 0.28        | 1.02        | -1.8          | -5.7          |
| 3              | 0.35        | 0.35        | 1.16        | -1.4          | -5.1          |
| 4              | 0.65        | 0.66        | 0.96        | -1.6          | -5.3          |
| 5              | 0.54        | 0.54        | 1           | -2            | -5.7          |
| 6              | 0.66        | 0.66        | 0.82        | 1.8           | -1.9          |
| 7              | 0.39        | 0.4         | 0.77        | 3.5           | -0.2          |
| 8              | 0.47        | 0.48        | 0.79        | 3.6           | 0.6           |
| 9              | 0.48        | 0.47        | 0.81        | 0.2           | -3.5          |
| 10             | 0.68        | 0.68        | 0.9         | 0.1           | -3.6          |
| 11             | 0.7         | 0.69        | 0.89        | 0.8           | -2.9          |
| 12             | 0.66        | 0.66        | 0.87        | 2.8           | -0.9          |
| 13             | 0.46        | 0.46        | 0.82        | -0.3          | -4            |
| 14             | 0.73        | 0.72        | 0.83        | 3.1           | -0.6          |
| 15             | 0.54        | 0.53        | 0.82        | 7.4           | 3.7           |
| 16             | 0.34        | 0.34        | 0.68        | 10.5          | 6.7           |
| 17             | 0.69        | 0.69        | 0.62        | 16.6          | 12.9          |
| 18             | 0.46        | 0.45        | 0.79        | 5.2           | 1.4           |
| 19             | 0.51        | 0.51        | 0.72        | 13.5          | 9.7           |
| 20             | 0.6         | 0.59        | 0.74        | 7.2           | 3.4           |
| <b>Average</b> | <b>0.54</b> | <b>0.54</b> | <b>0.85</b> | <b>3.36</b>   | <b>-0.3</b>   |

*Note.* Negative  $\Delta$ AIC and  $\Delta$ BIC values suggest the linear fit is better; positive values suggest the temporal bias fit is better

the AIC and BIC are inconsistent regarding the model selection, while in the last six, both the AIC and the BIC select the leaky integration as the best model.

## Discussion

We have shown that the ability to estimate the average of numerosity sequences extends to longer sequences in the range of two to eight. While this estimation is not perfect (and subject to a regression to the mean bias), one needs to keep in mind that the estimation of numerosity stimuli is noisy even with single stimuli. The important result is that the precision of the estimates improves with the length of the sequence. This could not happen if the participants form the estimates on the basis of few samples, say subject to WM capacity limitations (2-4). Instead, the results are consistent with a population code model, in which the responses to each stimuli is aggregated on the numerosity representation, and the estimate is obtained via a population code that estimates the center of mass (Brezis et al., 2018). In some of the participants, we also found small temporal biases. These subjects gave slightly more weight to more recent stimuli in the sequence.

## Experiment 3

The aim of our final experiment is to replicate the results of Experiment 2, under two important modifications. First the most relevant variable, sequence length, is now blocked rather than randomized. This allows the participants to select the most optimal strategy for each sequence length and reduces task uncertainty. Second, we set out to use here a more conventional method of estimation. Instead of indicating the estimate on a continuous scale, here, the sequence is followed by a probe (in a different color), and the task is to decide whether this probe has a higher (lower) numerosity than the average of the sequence. Another difference (which was not planned) was imposed on us by the COVID-19 restrictions. Due to these restrictions, this experiment was conducted online and not in a laboratory setting.

## Method

**Participants** Twenty three undergraduates from Tel-Aviv University ( $M_{age} = 23.65$ ,  $SD = 1.85$ ) participated in the experiment. The experiment was carried out online. Importantly, the participants were from the same pool as the participants in the two previous experiments—psychology students from Tel Aviv University. Participants had normal or corrected-to-normal vision. Participants were awarded course credit for their participation. All procedures and experimental protocols were approved by the ethics committee of the Psychology department of Tel Aviv University (Application 1-0000317). All experiments were carried out in accordance with the approved guidelines.

**Apparatus and stimuli** The stimuli consisted of dots randomly scattered on the screen. The dots diameter varied from 25 to 45 pixels. Minimum distance between dots was 25 pixels. The color of the dots was either light grey for the sequence array (RGB: 201, 201, 201) and yellow for target arrays (RGB: 255, 255, 0). The dots appeared on a black background.

**Procedure and design** Due to COVID-19 limitations, the experiment was conducted online and not in a laboratory setting. The experiment was built on PsychoPy Builder v3 and ran via Pavlovia. Each trial began with a white fixation cross presented on a black screen for 1,000 ms. The fixation was followed by a sequence of dot clouds. Each presented in the middle of the screen for 500 ms. At the end of the sequence, there was a black screen for 500 ms, followed by the target stimulus. Participants were instructed to decide whether the target stimulus, a yellow dot cloud, was bigger or smaller than the average of the sequence using the keyboard (*f* key for smaller, *j* key for bigger). After each response, participants received feedback, a green *v* for correct answers and red *x* for wrong answers (see Fig. 10). As in Experiment 2, sequences were

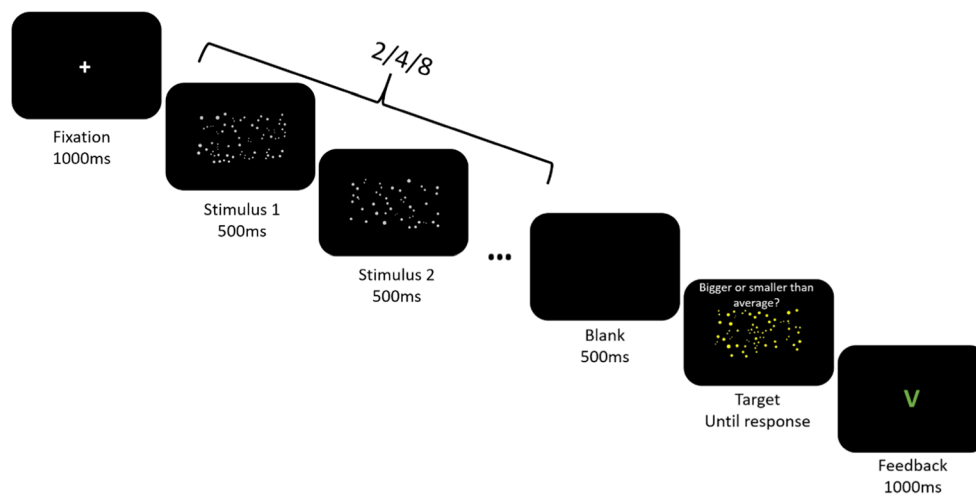


Fig. 10 Example of a trial in Experiment 3

randomly drawn from three uniform distributions with a range of 30 and means of 35, 50, and 65. Unlike Experiment 2, sequence length was blocked, resulting in six blocks, two for each sequence length. The order of the blocks was fixed: 2, 4, 8, 2, 4, 8. Participants were informed about the sequence length before each block. The proportions of the target stimuli to the average of the sequence were 0.71, 0.77, 0.83, 0.91, 1.1, 1.2, 1.3, 1.4 (first four values are the reciprocal of the last four). Each proportion appeared seven times for each average and sequence length, 504 trials in all. The experiment began with a short practice block with 12 trials with a sequence length of two.

## Results

Participants' overall accuracy was 74%,  $SD = .06$  (one participant's performance, while higher than chance, appeared

as a low performance outlier, accuracy = .58. Removing this participant from the analysis does not affect any of the results). We computed a psychometric curve for each participant. The slope of the psychometric curve is another indication of accuracy; a steeper slope represents more accurate responses (see Fig. 11a, for example). We compared the slope of the three sequence lengths:  $n = 2$  (mean = 5.12,  $SD = 2.14$ ),  $n = 4$  (mean = 5.6,  $SD = 1.5$ ), and  $n = 8$  (mean = 6.16,  $SD = 1.54$ ) with a repeated-measures ANOVA. We found a significant main effect,  $F(2, 44) = 6.21$ ,  $p < .005$ ,  $\eta_p^2 = .22$ . Further analysis revealed a significant linear trend,  $t(44) = 3.52$ ,  $p < .005$  (see Fig. 11b). In a similar analysis of RT's of sequence lengths, we did not find a significant effect ( $F(2, 44) < 1$ ).

**Temporal bias** We compared between a model without a temporal bias and a model with a temporal bias of averaging. For

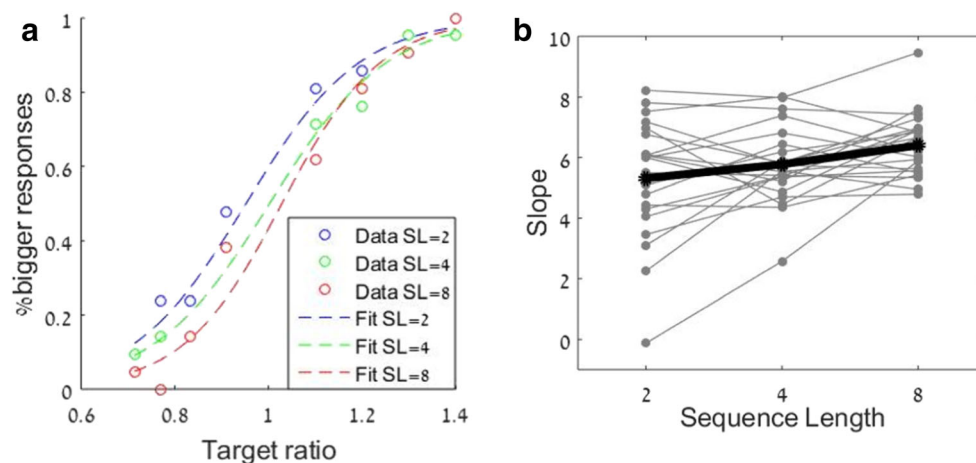


Fig. 11 **a** Psychometric curve of an individual participant, for each sequence length. **b** Slope of psychometric curve per sequence length. The grey lines mark individual participants. The black line is the average

the non-temporal bias model we computed a logistic regression with accuracy as the dependent variable and the delta between the target and the average of the sequence as the independent (predicting) variable.

$$\text{logit}(\hat{y}) = a + b(\text{target} - \bar{x}). \quad (3)$$

For the temporal bias model we computed a logistic regression based on Eq. 4 where:  $\lambda > 1$  indicates a primacy effect,  $\lambda < 1$  indicates a recency effect, and  $\lambda = 1$  means no temporal bias, and  $n$  is the sequence length.

$$\text{logit}(\hat{y}) = a + b((\text{target} - x_1) * \lambda^{n-1} + (\text{target} - x_2) * \lambda^{n-2} + \dots (\text{target} - x_n) * \lambda^{n-n}). \quad (4)$$

For 10 participants the model without a temporal bias was better (based on both AIC/BIC measures). The average  $\lambda$  in

the temporal bias for these participants was 0.89 (SD=.13). For 10 participants the model with temporal bias was better (based on both AIC/BIC measures), their average  $\lambda$  was 0.75 (SD=.04). For 3 participants, the model selection is AIC/BIC ambiguous their average  $\lambda$  was 0.84 (SD=.04; see Table 2).

## General discussion

In three experiments we examined the ability of human participants in a task of averaging of numerosity stimuli (sets of dots). This extends the range of operations on which numerosity representations were used from comparisons, addition or subtractions on pairs of stimuli, to the averaging of multiple stimuli – an operation that is of key importance to decision-making (Brusovansky et al., 2017; Vanunu et al., 2018; Weber, 2010). This task also extends the domain of stimuli on which the extraction of summary statistics was established, from domains such as size, orientation, emotional expression or object category (Ariely, 2001; Chong & Treisman, 2005; Dakin, 2001; Haberman et al., 2009; Haberman & Whitney, 2011; Khayat & Hochstein, 2018, 2019; Parkes et al., 2001; Robitaille & Harris, 2011) to the domain of nonsymbolic numerosities across temporal sequences.

Before discussing our results, we wish to mention a methodological dilemma that we have faced. There is an extensive debate in the numerical cognition literature, on distinguishing between the processing of number information (in dot stimuli sets) from other visual features, such as dot density, dot-size, dots-area (e.g., Gevers, Kadosh, & Gebuis, 2016; Leibovich & Henik, 2013; Leibovich, Katzin, Harel, & Henik, 2016). As this is a geometrically unsolvable problem, it is not possible to vary numerosity alone without co-varying some of the other variables. While we randomized the magnitude of the dots and their locations, in average, in our design the number was (negatively) correlated with dot-density. The problem of correlated features is thought to be more serious for comparison tasks (which of two arrays has more dots) than for estimation tasks (mapping onto a number line), as in the latter, the subject responds to numbers. While we did use a number-line type of response, we also set this up so that the participants could see, for each position of the mouse along the response line, a numerosity display corresponding to this specific location on the scale<sup>3</sup>. It may thus be possible to suggest that what the participants were doing is to move the mouse and, at each location on the scale, compare the display from memory with

**Table 2** Parameters for a model with and without a temporal bias

| Subject | Linear | Leak |           | $\Delta\text{AIC}$ | $\Delta\text{BIC}$ |
|---------|--------|------|-----------|--------------------|--------------------|
|         | $b$    | $b$  | $\lambda$ |                    |                    |
| 1       | 0.08   | 0.08 | 1.08      | −1.8               | −5.1               |
| 2       | 0.11   | 0.11 | 1.01      | −2.1               | −5.9               |
| 3       | 0.11   | 0.11 | 0.87      | 0.4                | −4.2               |
| 4       | 0.11   | 0.11 | 0.9       | −1.3               | −4.7               |
| 5       | 0.1    | 0.1  | 0.9       | −1.1               | −5.3               |
| 6       | 0.12   | 0.12 | 1.04      | −2.1               | −5.9               |
| 7       | 0.08   | 0.08 | 0.9       | −0.7               | −6.4               |
| 8       | 0.11   | 0.11 | 0.87      | 0.1                | −3.9               |
| 9       | 0.04   | 0.04 | 0.82      | −1.2               | −5.4               |
| 10*     | 0.02   | 0.02 | 0.59      | 0.3                | −5.2               |
| 11      | 0.1    | 0.1  | 0.79      | 4.5                | −0.9               |
| 12      | 0.17   | 0.17 | 0.88      | 1.1                | −3.3               |
| 13      | 0.14   | 0.14 | 0.84      | 3.4                | −1.5               |
| 14      | 0.11   | 0.12 | 0.76      | 7.6                | 3.9                |
| 15      | 0.11   | 0.11 | 0.76      | 6.2                | 2.3                |
| 16      | 0.12   | 0.12 | 0.75      | 9.3                | 5.1                |
| 17      | 0.1    | 0.1  | 0.69      | 9.7                | 6.3                |
| 18      | 0.12   | 0.12 | 0.78      | 4.6                | 1.4                |
| 19      | 0.1    | 0.1  | 0.66      | 10.2               | 5.6                |
| 20      | 0.11   | 0.12 | 0.78      | 6.9                | 3.5                |
| 21      | 0.12   | 0.12 | 0.76      | 6.3                | 1.6                |
| 22      | 0.11   | 0.11 | 0.78      | 4.2                | 0.8                |
| 23      | 0.11   | 0.11 | 0.76      | 4.3                | 0.2                |
| Average | −0.1   | 0.11 | 0.82      | 2.9                | −1.1               |

*Note.* Negative  $\Delta\text{AIC}$  and  $\Delta\text{BIC}$  values suggest the fit without a temporal bias is better; positive values suggest the temporal bias fit is better. \*The outlier subject with an insignificant slope

<sup>3</sup> We used this method, instead of presenting an empty scale with min/max displays at the two ends (Anobile et al., 2012), in order to prevent a ratio strategy. We did not want the subjects to compare the numerosity display formed at the end of the sequence with the displays at the ends of the scale, as such a ratio operation, on its own, leads to distortions (Hollands & Dyre, 2000). Rather, we wanted them to select the location on the scale where the stimulus looks as similar (with regard to numerosity) to the “average” estimate formed at the end of the sequence.



the one on the scale, based not on the number of dots, but rather on their density.

We believe this is unlikely for two reasons. First, unlike in the typical number line tasks, here the participants' main challenge was to average across multiple numerosity displays and our main focus was the averaging process (rather than its mapping to the response scale). As each display varies on multiple visual features, the task of generating an average based on visual features is not any easier than that of generating an average numerosity (especially, as this is what the participants were instructed to pay attention to). Based on the idea that numerosity displays automatically activate ANS representations (Nieder et al., 2002; Nieder & Miller, 2003; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004), we reasoned that following the presentation of multiple displays, the ANS representation will contain the composite of each numerosity presented, from which a population code should generate an estimate of the average (as a center of mass; Brezis et al., 2018; see Fig. 1a–c). We believe it is less likely, that one could automatically generate a similar averaging response for a visual feature such as dot-density (unless it is derived from numerosity). If for example, the dot-density (say, the average distance between the dots) was the direct visual feature that a subject monitored, the compression curves should have been inversed. Since there is more precision at small compared to long distances, one may expect (based on a Bayesian model; Anobile et al., 2012), contrary to our results, that large numbers (small distances) will be less affected by the prior (and thus more linear) compared with the small numbers (large distances). Furthermore, the results we obtained here, mimic those obtained in tasks of numerical averaging of symbolic numbers (Brezis et al., 2018). Nevertheless, future studies should exclude the possibilities that subjects could track dot-densities (instead of numerosity in our tasks) by including trials in which the density is kept constant when the numerosity changes.

We started (Experiment 1) with short sequences of two stimuli and we used responses on a continuous scale (after participants received training with the use of the scale). We find that all the participants show responses that are highly correlated with the sequence-mean (Fig. 6), but we also see some biases. First, all subjects show a regression to the mean effect, which is normative for the case in which the subject faces uncertainty (due to, for example, encoding or attentional variability) and the stimuli are distributed on a specified range (Anobile et al., 2012; Hollingworth, 1910; Jazayeri & Shadlen, 2010). Second, we find that about half of the participants show a compressive bias, which makes them underestimate the averages of pairs as a function of their distance (i.e., they estimate the pair (10, 80) less than the pair (30, 60)), although the instruction was to estimate the "average number of dots". This compressive bias was subject to clear individual differences, with some of the subjects showing no

compressive tendency whatsoever, while others showing a mild compression.

From a binary perspective, this variability (logarithmic vs. linear) in the architecture of the mental number line may seem surprising. The mental number line is considered to be a deeply rooted cognitive construct shared across species (Cantlon & Brannon, 2006). Accordingly, one might expect little individual differences. However, a close look at the literature might suggest otherwise. First, there are conflicting results, some found that the mental number line is linear (e.g., Ebersbach, Luwel, Frick, Onghena, & Verschaffel, 2008), while others reported a geometric mental number line (e.g., Dehaene, 2003) and the amount of compression appears to vary with age (Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003) or with attentional load (Anobile et al., 2012). One promising idea that can account for this variability is that it is based on the extension of the Bayesian central tendency model (Anobile et al., 2012) that is combined with the scalar-variability model with multiple samples. As we have shown in Fig. 2c, such a model can produce a range of compression effects. One attractive idea, which will require future research is that these individual differences would correlate with the risk-aversion tendencies (see Patalano et al., 2020; Peters, Slovic, Västfjäll, & Mertz, 2008; Peters et al., 2006, for similar ideas), since the task of averaging 10 and 90 dots shares much with the one of evaluating the attractiveness of a lottery that offers \$10 or \$90, each with probability .5.

In Experiment 2 and 3 we extended the length of the sequences from two to eight. Here we focused on how the precision of the estimate changes with the sequence-length. In Experiment 2 (using estimates on a continuous scale) we found that the estimates were highly correlated with the presented sequence average (average Pearson- $r = .66$ ; see Fig. 9a, and Supplement) and in Experiment 3 (using a more traditional choice procedure), we found that accuracy increases as a sigmoidal function of the difference between the sequence-average and the target (Fig. 11a). Importantly, however, in both experiments we found that the precision increases with the length of the sequence. This is remarkable, for two reasons. First, one would naively expect that computing an average is easier for two compared with eight items. Second, this improvement indicates a pooling operation across multiple stimuli in the sequence, which exceeds the capacity of the VWM. If for example, the participants can only average over 3–4 stimuli, the precision for  $n=8$  would be lower than that for  $n=4$  (see red line in Fig. 1), which is opposite to what the data shows (Fig. 9b, 11b). Future, research, however, will be needed to obtain an estimate of the bounds (or inefficiency; Solomon, May & Tyler, 2016) that operate in the averaging of such numerosity sequences.

The results are consistent, however, with a mechanism in which the estimate is generated by population averaging of the response each stimuli generates on the numerosity representation.

Similar results were previously obtained for the presentation of rapid sequences (rate of 2 Hz) of two digit numbers (Brezis et al., 2015). In that study, however, the precision was not monotonic with sequence length (it decreased from 4 to 8, and then increased again from 8 to 16; Brezis et al., 2015, Experiment 1). Interestingly, however, the precision became monotonically decreasing (like in our present study) when the items were presented at a faster rate (10 instead of 2 Hz), or when a response deadline was introduced (for the 2 Hz rate). Together with our present results this leads to a simple explanation. The mechanism of numerical-averaging with rapid sequences relies on population pooling over an ANS numerosity representation, for both numerosity and symbolic sequences, unless we present a sequence of 4 symbolic numbers and we do not impose a response deadline, which allows participants the opportunity to symbolically compute the average. The difference between numerosity and symbolic sequences, seems to be that with the former an analytic computation strategy is not available even for  $n=2$  stimuli and thus precision improves monotonically with sequence-length.

Much research in the domain of numerosity research has looked into the brain mechanisms, indicating a parietal network of numerosity detectors (Cohen Kadosh, Cohen Kadosh, Kaas, Henik, & Goebel, 2007; Eger et al., 2009; Fias, Lammertyn, Reynvoet, Dupont, & Orban, 2003; Harvey, Klein, Petridou, & Dumoulin, 2013; Nieder & Miller, 2003; Piazza et al., 2004; Piazza et al., 2007). Most of this research has focused on comparison tasks. In one study, the brain mechanism of numerical averaging was examined via tDCS (Brezis et al., 2016), showing that anodal stimulation in the parietal brain area enhances averaging precision compared with frontal or sham-stimulation. In two more recent studies, it was shown the while processing symbolic numbers, the neural (dis)similarity in patterns of electroencephalogram activity reflected numerical distance (Luyckx, Nili, Spitzer, & Summerfield, 2019; Spitzer et al., 2017). Future research with numerosity stimuli, which de-correlate between numerosity and density is needed to further examine the brain mechanism of numerical averaging. Such research could also examine whether one can decode the response to a sequence-average from the brain response when the subject carries out an averaging task and explore the role of the ANS network in decision-making under risk.

**Acknowledgments** We wish to thank Moshe Glickman for very helpful discussions on the design and the analysis, and Avishai Henik for helpful suggestions that led to the design of Experiment 3. This research was supported by grants to Marius Usher from the United States-Israel Binational Science Foundation (CNCRS 2014612), and from the Israel Science Foundation (1413/17).

**Open practices statement** None of the experiments were preregistered. Materials are available on OSF (<https://osf.io/m7qjy/>).

**Funding** This research was supported by grants to Marius Usher from the United States-Israel Binational Science Foundation (CNCRS 2014612), and from the Israel Science Foundation (1413/17).

## Appendix

### Bayesian version of the scalar-variability model simulation

We simulated the Bayesian version of the scalar-variability model (see Fig. 2c) using the following equation:

$$\text{EstimatedAverage} = \frac{w_1 * x_1 + w_2 * x_2 + w_p * 50}{w_1 + w_2 + w_p}, \quad (5)$$

where  $x_1$  and  $x_2$  are the two samples.  $w_1$ ,  $w_2$  are their decision weights, which are inversely proportional to their standard deviation in the scalar variability model (here, we assumed  $SD(x) = .8 \times x$ ), and  $w_p$  is weight given to the prior, chosen here as the middle of the range value (50). For Fig. 2c, we assumed that  $x_1 = 50 - x$ ,  $x_2 = 50 + x$ , and we varied  $w_p$ , via the  $a$  parameter (see Fig. 2c legend).

## References

- Anobile, G., Cicchini, G. M., & Burr, D. C. (2012). Linear mapping of numbers onto space requires attention. *Cognition*, 122(3), 454–459. <https://doi.org/10.1016/j.cognition.2011.11.006>
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2). <https://doi.org/10.1111/1467-9280.00327>
- Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, 86(3), 201–221. [https://doi.org/10.1016/S0010-0277\(02\)00178-6](https://doi.org/10.1016/S0010-0277(02)00178-6)
- Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E. (2006). Nonsymbolic arithmetic in adults and young children. *Cognition*, 98(3), 199–222. <https://doi.org/10.1016/j.cognition.2004.09.011>
- Barth, H., La Mont, K., Lipton, J., & Spelke, E. S. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39), 14116–14121. <https://doi.org/10.1073/pnas.0505512102>
- Betsch, T., Kaufmann, M., Lindow, F., Plessner, H., & Hoffmann, K. (2006). Different principles of information integration in implicit and explicit attitude formation. *European Journal of Social Psychology*, 36(6), 887–905. <https://doi.org/10.1002/ejsp.328>
- Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, 115(2), 463–501. <https://doi.org/10.1037/0033-295X.115.2.463>
- Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, 42(1), 189–201. <https://doi.org/10.1037/0012-1649.41.6.189>
- 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. [https://doi.org/10.1162/jocn\\_a\\_00991](https://doi.org/10.1162/jocn_a_00991)
- Brezis, N., Bronfman, Z. Z., & Usher, M. (2015). Adaptive spontaneous transitions between two mechanisms of numerical averaging. *Scientific Reports*, 5(1), 1–11. <https://doi.org/10.1038/srep10415>
- Brezis, N., Bronfman, Z. Z., & Usher, M. (2018, February 1). A perceptual-like population-coding mechanism of approximate numerical averaging. *Neural Computation*. [https://doi.org/10.1162/NECO\\_a\\_01037](https://doi.org/10.1162/NECO_a_01037)
- Brusovansky, M., Glickman, M., & Usher, M. (2018). Fast and effective: Intuitive processes in complex decisions. *Psychonomic Bulletin & Review*, 25(4), 1542–1548. <https://doi.org/10.3758/s13423-018-1474-1>
- Brusovansky, M., Vanunu, Y., & Usher, M. (2017). Why we should quit while we're ahead: When do averages matter more than sums? *Decision*. <https://doi.org/10.1037/dec0000087>
- Cantlon, J. F., & Brannon, E. M. (2006). Shared system for ordering small and large numbers in monkeys and humans. *Psychological Science*, 17(5), 401–406. <https://doi.org/10.1111/j.1467-9280.2006.01719.x>
- Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, 45(7), 891–900. <https://doi.org/10.1016/j.visres.2004.10.004>
- Cohen Kadosh, R., Cohen Kadosh, K., Kaas, A., Henik, A., & Goebel, R. (2007). Notation-dependent and -independent representations of numbers in the parietal lobes. *Neuron*, 53(2), 307–314. <https://doi.org/10.1016/j.neuron.2006.12.025>
- Cordes, S., Gallistel, C. R., Gelman, R., & Latham, P. (2007). Nonverbal arithmetic in humans: Light from noise. *Perception & Psychophysics*, 69(7), 1185–1203. <https://doi.org/10.3758/BF03193955>
- Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review*, 8(4), 698–707. <https://doi.org/10.3758/BF03196206>
- Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A*, 18(5), 1016. <https://doi.org/10.1364/josaa.18.001016>
- Dehaene, S. (2003, April 1). The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*. [https://doi.org/10.1016/S1364-6613\(03\)00055-X](https://doi.org/10.1016/S1364-6613(03)00055-X)
- Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation Stanislas Dehaene. In P. Haggard (Ed.), *Sensorimotor foundations of higher cognition* (pp. 527–574). Retrieved from [http://www.unicog.org/publications/Dehaene\\_SymbolsQuantitiesMathematicalTheory\\_ChapterAttPerf2007.pdf](http://www.unicog.org/publications/Dehaene_SymbolsQuantitiesMathematicalTheory_ChapterAttPerf2007.pdf)
- Dehaene, S., Dehaene-Lambertz, G., & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences*, 21(8), 355–361. [https://doi.org/10.1016/S0166-2236\(98\)01263-6](https://doi.org/10.1016/S0166-2236(98)01263-6)
- Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 16(3), 626–641. <https://doi.org/10.1037/0096-1523.16.3.626>
- Ebersbach, M., Luwel, K., Frick, A., Onghena, P., & Verschaffel, L. (2008). The relationship between the shape of the mental number line and familiarity with numbers in 5-to 9-year old children: Evidence for a segmented linear model. *Journal of Experimental Child Psychology*, 99, 1–17. <https://doi.org/10.1016/j.jecp.2007.08.006>
- Eger, E., Michel, V., Thirion, B., Amadon, A., Dehaene, S., & Kleinschmidt, A. (2009). Deciphering cortical number coding from human brain activity patterns. *Current Biology*, 19(19), 1608–1615. <https://doi.org/10.1016/j.cub.2009.08.047>
- Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8(7), 307–314. <https://doi.org/10.1016/j.tics.2004.05.002>
- Fias, W., Lammertyn, J., Reynvoet, B., Dupont, P., & Orban, G. A. (2003). Parietal representation of symbolic and nonsymbolic magnitude. *Journal of Cognitive Neuroscience*, 15(1), 47–56. <https://doi.org/10.1162/089892903321107819>
- Gallistel, C. R., & Gelman, R. (2000). Non verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, 4(2), 59–65.
- Gevers, W., Kadosh, R. C., & Gebuis, T. (2016). Sensory integration theory: An alternative to the approximate number system. *Continuous Issues in Numerical Cognition*, 405–418. <https://doi.org/10.1016/B978-0-12-801637-4.00018-4>
- Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, 18(5), 855–859. <https://doi.org/10.3758/s13423-011-0125-6>
- Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision*, 9(11), 1–1
- Harvey, B. M., Klein, B. P., Petridou, N., & Dumoulin, S. O. (2013). Topographic representation of numerosity in the human parietal cortex. *Science (New York, N.Y.)*, 341(6150), 1123–1126. <https://doi.org/10.1126/science.1239052>
- Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review*, 107(3), 500–524. <https://doi.org/10.1037/0033-295X.107.3.500>
- Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7(17), 461. <https://doi.org/10.2307/2012819>
- Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, 13(8), 1020–1026. <https://doi.org/10.1038/nn.2590>
- Katzin, N., Salti, M., & Henik, A. (2018). Holistic processing of numerical arrays. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(6), 1014–1022. <https://doi.org/10.1037/xlm0000640>
- Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, 18(9), 1–14. <https://doi.org/10.1167/18.9.23>
- Khayat, N., & Hochstein, S. (2019). Relating categorization to set summary statistics perception. *Attention, Perception, & Psychophysics*, 81(8), 2850–2872
- Leibovich, T., & Henik, A. (2013). Magnitude processing in nonsymbolic stimuli. *Frontiers in Psychology*, 4(June), 375. <https://doi.org/10.3389/fpsyg.2013.00375>
- Leibovich, T., & Henik, A. (2014). Comparing performance in discrete and continuous comparison tasks. *Quarterly Journal of Experimental Psychology*, 67(5), 899–917. <https://doi.org/10.1080/17470218.2013.837940>
- Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2016). From ‘sense of number’ to ‘sense of magnitude’—The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*. <https://doi.org/10.1017/S0140525X16000960>
- Luyckx, F., Nili, H., Spitzer, B., & Summerfield, C. (2019). Neural structure mapping in human probabilistic reward learning. *eLife*, 8. <https://doi.org/10.7554/eLife.42816>
- Malmi, R. A., & Samson, D. J. (1983). Intuitive averaging of categorized numerical stimuli. *Journal of Verbal Learning and Verbal Behavior*, 22, 547–559. Retrieved from <https://search.proquest.com/openview/fb96a7452bfc5369bcd63dbda8c9e5f9/1?pq-origsite=gscholar&cbl=1819609>
- McCrink, K., Dehaene, S., & Dehaene-Lambertz, G. (2007). Moving along the number line: Operational momentum in nonsymbolic arithmetic. *Perception & Psychophysics*, 69(8), 1324–1333. <https://doi.org/10.3758/BF03192949>



- Mitrani-Rosenbaum, D., Glickman, M., & Usher, M. (2020). *Extracting summary statistics of rapid numerical sequences*. <https://doi.org/10.31234/osf.io/6scav>
- Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, 215(5109), 1519–1520.
- Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science (New York, N.Y.)*, 297(September), 1708–1711. <https://doi.org/10.1126/science.1072493>
- Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, 37(1), 149–157. [https://doi.org/10.1016/S0896-6273\(02\)01144-3](https://doi.org/10.1016/S0896-6273(02)01144-3)
- 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>
- Patalano, A. L., Zax, A., Williams, K., Mathias, L., Cordes, S., & Barth, H. (2020). Intuitive symbolic magnitude judgments and decision making under risk in adults. *Cognitive Psychology*, 118, 101273. <https://doi.org/10.1016/j.cogpsych.2020.101273>
- Peters, E., Slovic, P., Västfjäll, D., & Mertz, C. K. (2008). *Intuitive numbers guide decisions: Judgment and decision making* (Vol. 3). Retrieved from [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1321907](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1321907)
- Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, 17(5), 407–413. <https://doi.org/10.1111/j.1467-9280.2006.01720.x>
- Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, 44(3), 547–555. <https://doi.org/10.1016/j.neuron.2004.10.014>
- Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A Magnitude Code Common to Numerosities and Number Symbols in Human Intraparietal Cortex. *Neuron*, 53(2), 293–305. <https://doi.org/10.1016/j.neuron.2006.11.022>
- Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science (New York, N.Y.)*, 306(5695), 499–503. <https://doi.org/10.1126/science.1102085>
- Pleskac, T. J., Yu, S., Hopwood, C., & Liu, T. (2019). Mechanisms of deliberation during preferential choice: Perspectives from computational modeling and individual differences. *Decision*, 6(1), 77–107. <https://doi.org/10.1037/dec0000092>
- Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision*, 11(12), 18–18. <https://doi.org/10.1167/11.12.18>
- Roe, R. M., Bussemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108(2), 370–392. <https://doi.org/10.1037/0033-295X.108.2.370>
- Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development*, 75(2), 428–444. <https://doi.org/10.1111/j.1467-8624.2004.00684.x>
- Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: evidence for multiple representations of numerical quantity. *Psychological Science*, 14(3), 237–243. <https://doi.org/10.1111/1467-9280.02438>
- 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. <https://doi.org/10.1038/s41562-017-0145>
- Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24), 9659–9664. <https://doi.org/10.1073/pnas.1119569109>
- Usher, M., & McClelland, J. L. (2004, July). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*. <https://doi.org/10.1037/0033-295X.111.3.757>
- Vandormael, H., Castañón, S. H., Balaguer, J., Li, V., & Summerfield, C. (2017). Robust sampling of decision information during perceptual choice. *Proceedings of the National Academy of Sciences of the United States of America*, 114(10), 2771–2776. <https://doi.org/10.1073/pnas.1613950114>
- Vanunu, Y., Pachur, T., & Usher, M. (2018). Constructing preference from sequential samples: The impact of evaluation format on risk attitudes. *Decision*, 6(3), 223–236. <https://doi.org/10.1037/dec0000098>
- Weber, E. U. (2010). Risk attitude and preference. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1), 79–88. <https://doi.org/10.1002/wcs.5>
- Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, 10(2), 130–137. <https://doi.org/10.1111/1467-9280.00120>
- Zeigenfuse, M. D., Pleskac, T. J., & Liu, T. (2014). Rapid decisions from experience. *Cognition*, 131(2), 181–194. <https://doi.org/10.1016/j.cognition.2013.12.012>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.