LETTER Communicated by Jordan Rashid

A Perceptual-Like Population-Coding Mechanism of Approximate Numerical Averaging

Noam Brezis
noambrezis@gmail.com
School of Psychology, Tel Aviv University, Tel Aviv 69978, Israel

Zohar Z. Bronfman
zoharbronfman@gmail.com
School of Psychology and Cohn Institute for the History and Philosophy of Science and Ideas, Tel Aviv University, Tel Aviv 69978, Israel

Marius Usher
marius@post.tau.ac.il
School of Psychology and Sagol Institute of Neuroscience, Tel Aviv University, Tel Aviv 69978, Israel

Humans possess a remarkable ability to rapidly form coarse estimations of numerical averages. This ability is important for making decisions that are based on streams of numerical or value-based information, as well as for preference formation. Nonetheless, the mechanism underlying rapid approximate numerical averaging remains unknown, and several competing mechanism may account for it. Here, we tested the hypothesis that approximate numerical averaging relies on perceptual-like processes, based on population coding. Participants were presented with rapid sequences of numerical values (four items per second) and were asked to convey the sequence average. We manipulated the sequences' length, variance, and mean magnitude and found that similar to perceptual averaging, the precision of the estimations improves with the length and deteriorates with higher variance or higher magnitude. To account for the results, we developed a novel biologically plausible population-coding neurocomputational model and showed that it is mathematically equivalent to a population vector. Using both quantitative and qualitative model comparison methods, we compared the population-coding model to several competing models, such as a step-by-step running average (based on leaky integration) and a midrange model. We found that the data support the population-coding model. We conclude that humans' ability to rapidly form estimations of numerical averages has many properties of the perceptual (intuitive) system rather than the arithmetic,

N. B. and Z. B. are co-first authors.

Neural Computation **30, 1–19** (2018)
doi:10.1162/NECO_a_01037

linguistic-based (analytic) system and that population coding is likely to be its underlying mechanism.

1 Introduction

Previous research has indicated that humans act as intuitive statisticians, rapidly forming approximate, or coarse (i.e., not in a rule-based manner), estimations of the mean and variance of sequences of number or numerosity values, at rates as high as two items per second (Beach & Swenson, 1966; Malmi & Samson, 1983; Brezis, Bronfman, & Usher, 2015; Brezis, Bronfman, Jacoby, Lavidor, & Usher, 2016; Rusou, Zakay, & Usher, 2016). This ability is of major importance for daily decisions that are based on streams of information (e.g., a broker who wishes to decide whether to buy a certain stock; Bechara, Damasio, Tranel, & Damasio, 2005; Hertwig & Erev, 2009; Tsetsos, Chater, & Usher, 2012), for intuitive preference formation (Anderson, 1981), or for decisions that are based on information that is not attended (Betsch, Plessner, Schwieren, & Gütig, 2001; Betsch, Kaufmann, Lindow, Plessner, & Hoffmann, 2006; Van Opstal, De Lange, & Dehaene, 2011). The mechanism underlying this ability, however, remains largely unknown.

In previous work (Brezis et al., 2015), we provided preliminary support in favor of a population coding–based account. Motivated by Daniel Kahneman's hypothesis, according to which intuitive decisions are at the interface of perceptual and cognitive processes (Kahneman, 2003), and by Stanislas Dehaene's analog numerical processing theory (Dehaene, 2011), we argued that approximate numerical averaging relies on perceptual-like processes, operating on analog representations, such as those that participate in statistical estimations of the numerosity or size of visual elements (Ariely, 2001; Chong & Treisman, 2003), and we proposed a population coding–based neurocomputational model to account for approximate numerical averaging. In particular, we showed that similar to perceptual processes, yet unlike symbolic or rule-based processes (e.g., the arithmetic computation of an average, as a sum divided by the number of elements), the accuracy of approximate numerical averaging increases as the number of elements increases, since noise associated with each individual item is averaged out as additional items are being integrated.

These results ruled out the possibility that humans deploy an analytic rule-based strategy when computing the average of rapid numerical sequences. However, mechanisms besides population averaging could potentially account for these results as well. One possibility is that in approximate numerical averaging, humans calculate a running average by deploying a serial updating process (Hogarth & Einhorn, 1992) that can be based on either coarse arithmetic (analytic) calculation or on a type of running average (i.e., a leaky integration, whereby a weighted average between the most recent value and the previous estimate is being computed

on each additional sample). Running average–based models have been considered to play a major role in numerical cognition (e.g., Budescu, Weinberg, & Wallsten, 1988; Summerfield & Tsetsos, 2012; Ashby & Rakow, 2014; Wulff & Pachur, 2016), in integration-based perceptual decisions (e.g., Usher & McClelland, 2001; Ossmy et al., 2013), and in preference formation (e.g., Van Overwalle & Labiouse, 2004; Yechiam & Busemeyer, 2005; Tsestos et al., 2012). They are therefore a natural candidate mechanism for approximate numerical averaging. An alternative rule-based but heuristic mechanism for estimating numerical averages, subject to working memory capacity constraints, is to compute the average based on few "well-selected" samples (e.g., the min and the max of the sequence; Myczek & Simons, 2008). Thus, the question of the mechanism underlying approximate numerical averaging remains open.

The aim of our study is thus twofold: to investigate the influence of two key perceptual factors—variance and magnitude—on approximate numerical averaging and to use these factors in order to pit various potential averaging mechanisms one against another by relying on both behavioral-qualitative tests and quantitative model-fitting procedures. Toward this aim, we have used a numerical averaging paradigm (see Figure 1) with a rapid presentation rate of four numbers per second (250 ms per item). We reasoned that if participants would still carry out the task at a reasonable level under such conditions, it will provide further evidence in favor of a holistic (population-based) process compared with a serial step-by step process. Second, we examined and manipulated two additional pivotal factors that affect the accuracy of the perceptual-averaging system: the variance of the elements (higher variance deteriorates accuracy in perceptual tasks; De Gardelle & Summerfield, 2011) and the magnitude of the elements (higher magnitude deteriorates accuracy, assuming Weber-like representations; Shepard, Kilpatric, & Cunningham, 1975). It currently remains unknown whether these two factors affect approximate numerical averaging. Finally, we subjected the data to a model comparison between a population code model, a heuristic model (midrange, which requires only to note and estimate two items, the maximum and the minimum), and a number of step-by-step models (based on leaky integration).

We begin by presenting our experimental design and results. To anticipate, we observe an impressive ability to carry our numerical averaging at a rate of four numbers per second, and find that both the variance and the magnitude of the sequence numbers reduce accuracy. We then present and extend our population code model of numerical averaging to a fully biologically plausible model, showing it is mathematically equivalent to the population vector (Georgopoulos, Schwartz, & Kettner, 1986; Pouget, Dayan, & Zemel, 2003), and report the results of the model comparison. The results support the population code model and have important implications regarding the nature of the analog number representations.

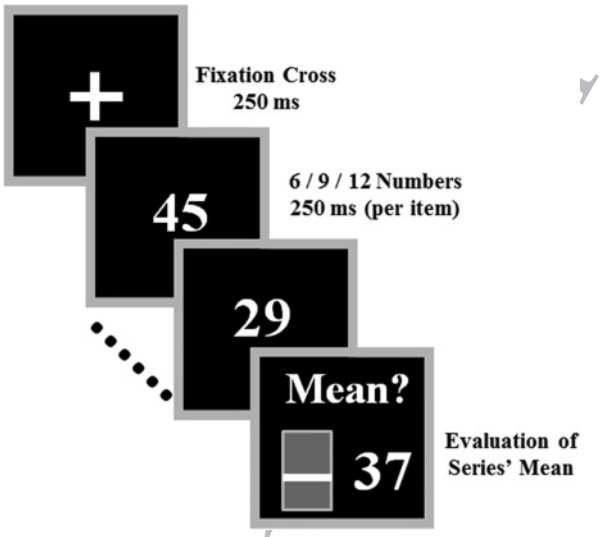

Figure 1: Depiction of a typical trial. Following the number sequence, participants were asked to convey the average by sliding a mouse-controlled bar. The corresponding number was displayed—here, for example, 37.

2 Method

2.1 Participants. The eighteen participants in the experiment were undergraduate students recruited through the Tel Aviv University Psychology Department's participant pool; they were naive to the purpose of the experiment and had normal, or corrected-to-normal, vision. Informed consent was obtained from all subjects. Participants were awarded either course credit for their participation or a small financial compensation (40 NIS, equivalent to about $10). They received a performance-dependent bonus of an additional 10 to 20 NIS. All procedures and experimental protocols were approved by the ethics committee of the Psychology Department of Tel Aviv University (application 743/15). All experiments were carried out in accordance with the approved guidelines.

2.2 Stimulus Materials and Procedure. The basic setup of a trial is depicted in Figure 1. Each trial began with a central fixation cross (250 ms) after which a sequence of two-digit numbers was presented (white arabic numerals on a black background; each number displayed for 250 ms; without blank interstimulus intervals. The sequence set size (i.e., the quantity of displayed numbers) consisted of 6, 9, or 12 items (randomly intermixed). The only instructions participants received were to convey as

accurately as possible the sequence's average by vertically sliding a mouse-controlled bar set on a number ruler between 0 and 100 (the number corresponding to the bar's location was concurrently displayed) and pressing the left mouse button when reaching the desired number (see also Brezis et al., 2015). After completing 20 practice trials, participants underwent 720 experimental trials divided into six blocks. Each block terminated with performance feedback (block-average correlation) and a short, self-paced break.

In order to the test the effect of variance, magnitude, and set size on the accuracy of the evaluations, we used a factorial design. The sequences of numbers were generated so as to create a $2 \times 2 \times 3$ design of variance (low/high), magnitude (low/high), and set size (6, 9, and 12) in the following manner. To achieve sequences of low magnitude and of low and high variance, we sampled, independently for each set size, the numbers from a gaussian distribution (mean = 30) with either low or high standard deviation (low = 8, high = 16). In case two identical numbers were sampled successively, the entire sequence was shuffled in order to prevent successive presentation. Numbers above 90 and below 10 were discarded. In addition, sequences were resampled in case the average magnitude of the low- and high-variance sequences was not identical. To generate high-magnitude sequences with identical standard deviations, we used the same procedure, yet with the mean of the gaussian distribution = 55. The order of the presented sequences was randomly determined. Overall 740 sequences were generated (20 practice and 180 sequences per condition). All stimuli were generated using Matlab and were presented on a gamma-corrected ViewSonic (Walnut, CA) 17 inch monitor viewed at a distance of 41 cm. The screen resolution was set to 1024×768 pixels, and the monitor had a refresh rate of 60 Hz. We obtained participants' evaluation of the sequence average and response time (RT; measured from sequence's offset until mouse button press) in each trial. No data were discarded.

2.3 Dependent Variables. We used two measures to quantify each participant's precision of evaluations: (1) Pearson correlation across trials between the sequence's average and estimations and (2) root-mean-square deviation (RMSD), given by $\sqrt{\sum_{i=1}^{n}(x_i - \mu_i)^2/n}$; where x_i corresponds to the estimated average of sequence i and μ_i corresponds to that sequence's actual average. Note that lower values of RMSD correspond to higher accuracy. Since these measures are strongly correlated, we report only the correlation for the overall task accuracy (it provides a simple measure of task performance, with 0 corresponding to guessing and 1 to perfect accuracy, subject to scaling), but we report the RMSD when we examine differences between conditions (since RMSD is less sensitive to differences between the actual trials in each condition). All effects reported in one measure remain significant when tested using the other measurement.

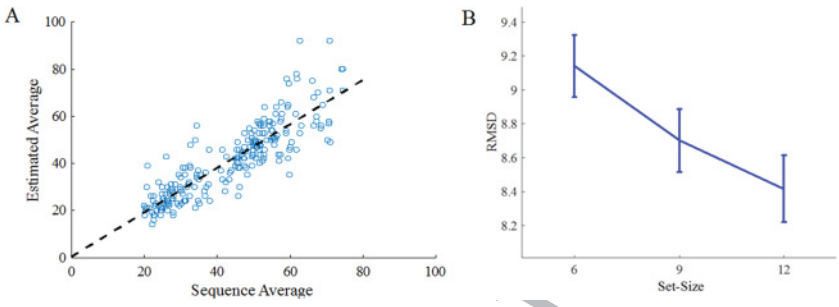

Figure 2: Performance in the averaging task (A) Estimations of a representative participant (median correlation score; Pearson correlation = 0.87). (B) Effect of sequence set size on estimation accuracy (RMSD). As set size increases, RMSD decreases (accuracy increases). Error bars denote 1 within-participant standard error of the mean.

3 Results

3.1 Behavioral Results. Overall, and despite the very rapid presentation (250 ms per numeral) and the large quantity of numbers (up to 12 items per sequence), the participants exhibited high accuracy in estimating the average of the numerical sequences. Participants' Pearson correlation across trials between the sequence's average and estimations is remarkably high and significantly larger than 0 ($r = 0.83$ (SD = 0.138); $p < 0.0001$ for all participants; see Figure 2A for a single representative participant). Importantly, this result is found separately for the low ($r = 0.74$ (SD = 0.17); $p < 0.00001$ for all participants except one with $p = 0.03$) and high mean conditions ($r = 0.62$ (SD = 0.15); $p < 0.0001$; for all participants), suggesting that the high sensitivity exhibited by the participants is for the actual presented numerical values rather than for the distributions' means from which the numbers were drawn. This is because reliance on the generative means would necessarily result in null correlation within each condition. Thus, there must have been some form of averaging of the specific numerical sequences within each mean condition.

In addition, each participant's square deviations between the estimation and the actual average are significantly lower than the square deviations obtained by randomly shuffling the participant's responses across trials and comparing the average square deviations between the actual mean and the shuffled estimations, across 500 independent shuffles (actual = 88 (SD = 72); shuffled = 455; $p < 0.001$, for all participants). We find no learning effect in the experiment, as the correlation did not increase across the six experimental blocks ($p > 0.7$).

Furthermore, consistent with our previous findings (Brezis et al., 2015), we find that participants' RMSD decreased (accuracy increases) as a

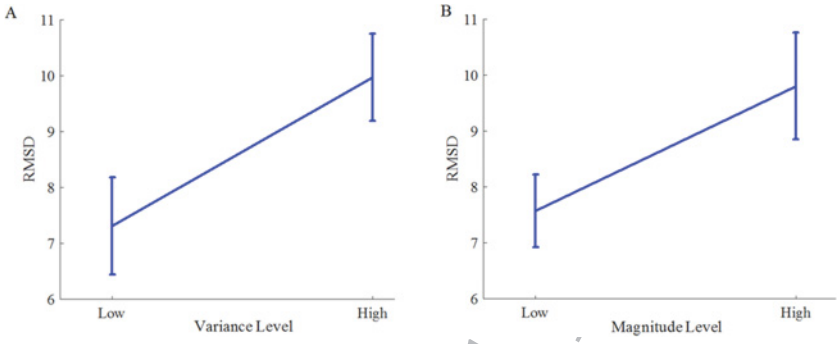

Figure 3: Behavioral results. The effect of variance (A) and magnitude (B) on the accuracy (RMSD) of estimations. Higher levels of variance or magnitude result in lowered accuracy (higher RMSD). Error bars denote 1 within-participant standard error of the mean.

function of set size ($6 = 9.14$; $9 = 8.7$; $12 = 8.42$. ANOVA of within-participant effect of set size on RMSD: $F(2, 34) = 4.64$; $p = 0.016$; Mauchly's W, for equal variance = 0.999; see Figure 2B). This indicates that noise at the encoding of the elements is averaged out during the averaging process, thus supporting the hypothesis that perceptual-like, non-rule-based processes underlie numerical averaging in the task. We also observe a recency bias in the 6- and 9-number conditions (but not in the 12-number condition), as indicated by regression weights assigned to each sample as a function of its location within the sequence (ANOVA of within-participant effect of temporal regression weight; $p < 0.01$; for the 6 and 9 set size, see the online supplement). This result replicates previous findings, according to which evaluations are more influenced from samples arriving later in the sequence (cf. Brezis et al., 2015).

To test the effect of sequence variance on the accuracy of the estimations, we compared the RMSD between the low- and high-variance conditions.[1] We find that the RMSD for low-variance sequences is lower (i.e., performance is better) than under high variance (low = 7.31; high = 9.97; $t(17) = -7.62$; $p < 0.0001$; see Figure 3A). To test the effect of the magnitude of the number sequence on the accuracy of the estimations, we compared the RMSD between the low- and high-magnitude conditions. We find that the RMSD for low-magnitude sequences is lower (i.e., performance is better) than under high magnitude (low = 7.57; high = 9.80; $t(17) = -4.46$; $p < 0.0005$; see Figure 3B). No differences were found in

[1]Note that we did not compare correlations for this analysis, as the numerical range between the conditions is dissimilar, a property that can result in a distorted estimation of the correlations.

reaction times (RT) between the set size conditions ($p > 0.5$) or variance conditions ($p > 0.3$). We find a significant effect of RT in the magnitude condition: RTs are longer in higher-magnitude trials ($t(17) = 3.2$; $p = 0.005$). This result indicates that the decrease in accuracy in higher-magnitude trials does not reflect the RT-accuracy trade-off; rather, it is a genuine decrease in performance.

Thus, we find strong evidence that both variance (∼30%) and magnitude dramatically (∼30%) deteriorate the accuracy of approximate numerical averaging. To account for these effects, we tested several competing computational models.

3.2 Modeling Approximate Numerical Averaging. In order to account for approximate numerical averaging, we considered three classes of alternative models: a population coding–based model (see Georgopoulos et al., 1986); a running average model, which is based on a leaky integrator; and a heuristic model, whereby the midrange of each sequence is estimated.

We first describe each model in detail and then present the fitting procedure and the results.

3.2.1 Population-Coding Model. It has been suggested that approximate numerical averaging relies on lower-level processes of summary statistics extraction (Dehaene, 2001; Verguts & Fias, 2004; Van Opstal et al., 2011; Dotan, Friedmann, & Dehaene, 2014; Brezis et al., 2015), whereby upon exposure to a multitude of continuous features (e.g., spatial orientation, circle diameter), observers exhibit high sensitivity to the average of the ensemble (Ariely, 2001; Chong & Treisman, 2003; Alvarez & Oliva, 2008). Thus, according to this hypothesis, sensitivity to summary statistics arises as a result of a process where the activation of dedicated feature-tuned neural populations is pooled together and the centroid of this activation profile, representing the features' average, is extracted (Georgopoulos et al., 1986; Pouget et al., 2003). Building on previous neurocomputational studies (Brezis et al., 2015), we suggest that such a population-coding model can account for the observed effects of variance and magnitude of the numerical sequences.

The model assumes that each number (10–90), activates a distinct gaussian distribution over a layer of broadly tuned numerosity detectors (layer 1; see Figure 4A), with a standard deviation, σ. Upon the presentation of a number, each unit or neuron responds probabilistically by triggering a number of spikes sampled from a Poisson distribution with a mean, λ, determined by the corresponding numerical tuning curve (see Figure 4B). Each successive number presented triggers an additional, accumulated probabilistic neural activation (see Figure 4C).[2] Note that the

[2]The population model can also include a recency component: the neural profile of previous items may be subject to decay and thus receive less weight in the population response. We do not include this factor here, but see section 4.

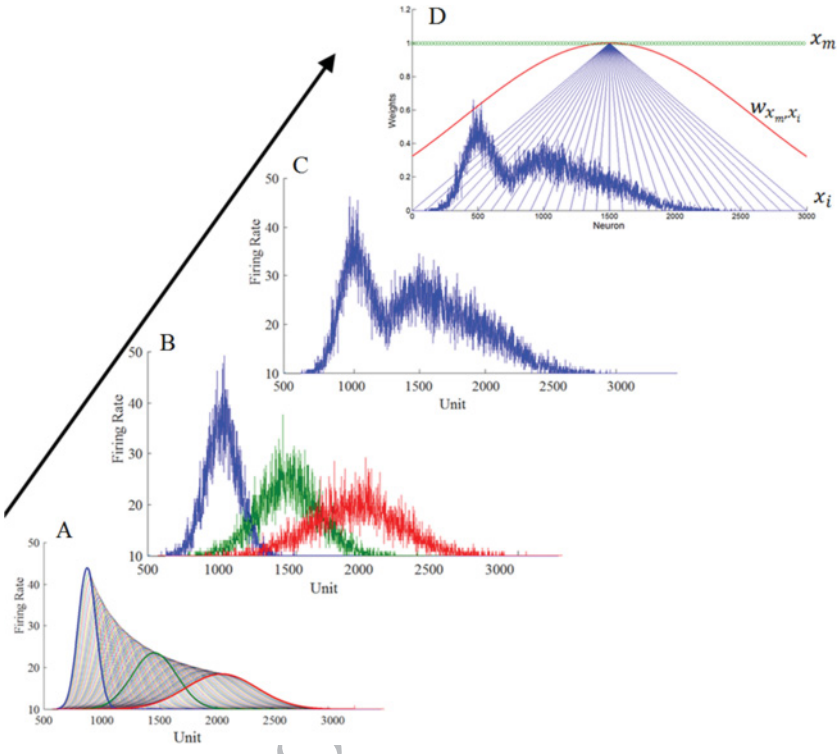

Figure 4: The suggested biologically plausible population-coding model for perceptual and numerical approximate averaging. Each panel represents a stage in the evaluation process. (A) The underlying tuning curves of the number-tuned units. (B) A noisy activation profile for a given three presented numbers (here, 20-50–80). (C) The superimposed overall activation profile of the network, for the given three numbers. (D) Extraction of the activation profile center of mass using a winner-take-all competition in a second layer. Each unit in the second layer is connected to all the units in the first layer with a parabolic weighting function (red curve).

Poisson random variability is the only source of item-dependent variability in the model. Without it, the model's evaluations would not improve with set size. The output of this population code is obtained by a weighted average (center of mass) and has an extra noise variable that corresponds to decision or motor variability. To obtain the center of mass in a biologically plausible mechanism (i.e., to decode the perceived average from the population response), we introduce an additional fully connected layer (layer 2) with each node in layer 2 corresponding to the same number representations as in layer 1 (i.e., 10–90), so that the effect of activation in layer 1 maps

topographically to layer 2 (see Figure 4D). Thus, each unit in layer 2 is connected to all the nodes in layer 1, yet with a connectivity weight matrix that is given by a diminishing parabolic function in the form of

$$w_{x_m,x_i} = -(x_i - x_m)^2, \tag{3.1}$$

where x_i is the number representation of unit i at layer 1 and x_m is the number representation of unit m at layer 2.

The function's maximal weight is always assigned to the topographically corresponding unit in layer 1 (i.e., to the unit in layer 1 that is tuned to the same number that is represented by the unit in layer 2). The unit with the maximal level of activity at layer 2 (identified by lateral winner-take-all competition) corresponds to layer 1's center of mass—that is, to the perceived sequence's average due to the following considerations:

- The activity in layer 2 is given by

$$L_2(x_m) = \sum_i w_{x_m,x_i} * L_1(x_i). \tag{3.2}$$

- Given the weight matrix described in equation 3.1, the derivative of layer 2 activity is

$$\frac{dL_2}{dx_m} = 2\sum_i (x_i - x_m) * L_1(x_i). \tag{3.3}$$

- The maximal activity in layer 2 is thus described by

$$\sum_i (x_i - x_m) * L_1(x_i) = 0. \tag{3.4}$$

As can be seen, x_m satisfies the center-of-mass definition (i.e., the sum of weighted distances equals zero), given the activity profile of layer 1.

Thus, as we have demonstrated, the suggested neuronal architecture is mathematically equivalent to the arithmetic center of mass, usually assumed in models of population coding.

We tested three variants of the population-coding model:

1. Two free parameters: σ (standard deviation) of the gaussian tuning curve, representing low-level perceptual noise at the individual item, and late noise added to the winning unit at layer 2—a stochastically independent internal noise sampled from a normal distribution (with mean 0 and standard deviation $\hat{\sigma}$), which reflects the general noise associated with the decoding process and motor noise (Solomon, Morgan, & Chubb, 2011).

2. Three free parameters: two parameters as in model 1 and, in addition, δ—a parameter that determines the linear increase of the standard deviation of the tuning curves as a function of the number magnitude.
3. Three free parameters: two parameters as in model 1 and ψ—a parameter that determines the logarithmic spacing of the tuning curves (an alternative manner to account for the decrease in differentiation as a function of magnitude; see the discussion in Dehaene, 2003).

3.2.2 Midrange Model. We tested a heuristic rule-based model, whereby evaluations are given by a noisy estimation of the sequence's midrange (Myczek & Simons, 2008):

$$\dddot{X}(\min(X) + \max(X))/2 + \ddot{\varepsilon}, \tag{3.5}$$

where X is the noisy vector of the number sequence (with σ determining the internal noise of each item) and $\ddot{\varepsilon}$ is a stochastically independent internal noise sampled from a normal distribution (with mean 0 and standard deviation $\hat{\sigma}$), which reflects the general noise associated with the decoding process and motor noise (Solomon et al., 2011).

We tested one variant of the heuristic midrange model:

1. Three free parameters: σ, determining the noise at the item level (ε); $\hat{\sigma}$—late noise; and δ—a parameter that determines the linear increase of σ as a function of the number magnitude

3.2.3 Running Average Leaky-Integrator Model. An alternative model that can account for approximate numerical averaging is the exponential moving average (EMA) model, whereby following each item, the current estimation of the average is being updated in a weighted manner. This can be achieved by assuming leaky integration in the form of

$$\dddot{X}_i = (1 - 1/\tau) * \dddot{X}_{i-1} + \tau * (x_i + \varepsilon), \tag{3.6}$$

where $\dddot{X}_i$ is the current estimated average at time $_i$, τ is the decay constant, x_i is the numeric item, and ε is a stochastically independent internal noise, sampled from a normal distribution (with mean 0 and standard deviation σ),

We tested several variants of the EMA leaky-integration model:

1. Three free parameters: σ, determining the noise at the item level (ε); τ, which determines the integrator's leak; and late noise added to the integrator's outcome—sampled from a normal distribution, with mean 0 and standard deviation $\hat{\sigma}$, and which reflects the general noise associated with the decoding process and motor noise (Solomon et al., 2011)

2. Four free parameters: three parameters as in model 1 and δ—a parameter that determines the linear increase of σ as a function of the number magnitude

In addition to these two variants, we also tested two variants in which the weight of the update is a function of the number's probability within the experimental probability distribution (thus incorporating prior beliefs into the estimation). Because these variants had higher Bayesian information criterion (BIC) values, we do not report them here (see the supplement for the models' description and fitting results).

3.3 Fitting Procedure. In order to obtain estimations of the best-fitting parameters and goodness of fit for each model, we fitted the models to the participants' evaluations on a trial-by-trial basis, separately for each participant. For any given set of parameters, 1000 simulations of each actual trial were generated using the actual number sequence that was presented. This allowed us to obtain a distribution of the model's average estimations. Assuming the model's estimations are normally distributed (Shapiro-Wilk test of composite normality; $SW = 0.998$; $p = 0.53$; Shapiro & Wilk, 1965; see also Razali & Wah, 2011), we obtained a probability density function reflecting the probability of the model to generate each possible average estimation. Given the probability function, we assigned a likelihood for the observed evaluation for each trial (i.e., the value of the probability function given the actual observed value). The likelihood for all of the data was calculated by multiplying the likelihood for the separate trials. Finally, parameters were estimated by maximizing the likelihood term using an exhaustive grid search (see Table 1 for the best-fitting parameters of each model).

3.4 Modeling Results. We find that the best model in terms of both maximal likelihood and Bayesian information criterion (BIC, which penalizes additional parameters) is variant 2 of the population coding model.

In addition, we show in Figure 5 the predictions of the behavioral results made by simulating the best-fitting model using the best-fitting parameters (generalizability criterion; Busemeyer & Wang, 2000; Ahn, Busemeyer, Wagenmakers, & Stout, 2008). The population code model accounts for all the behavioral patterns we reported in the data: set size, variance, and magnitude effect. Conversely, the running-average leaky-integrator model fails to predict the set size effect (note that variant 1 of this model, which does not assume Weber-like representations, predicts the set size effect yet not the magnitude effect).

4 Discussion

Approximate numerical averaging is fundamental to situations in which humans make decisions among alternatives that are characterized by

Table 1: Summary of Fitting Results Showing Best-Fitting Parameter Values for Each Model.

Model	Maximum Log Likelihood	BIC	Parameters		
Population 2	−838.61	1693.7	$\delta = 0.017$ [0:0.01:0.2]	$\sigma = 21.38$ [0:1:30]	$\hat{\sigma} = 6.00$ [0:1:15]
Population 1	−846.63	1704.2		$\sigma = 24.83$ [0:1:30]	$\hat{\sigma} = 7.60$ [0:1:15]
Running average 2	−842.97	1707.9	$\delta = 0.175$ [0:0.01:0.3]	$\sigma = 2.15$ [0:1:30]	$\hat{\sigma} = 5.88$ [0:1:15] $\tau = 4.16$ [0:1:10]
Population 3	−851.52	1719.5	$\Psi = 0.02$ [0:0.004:0.04]	$\sigma = 22.41$ [0:1:30]	$\hat{\sigma} = 7.66$ [0:1:15]
Running average 1	−853.00	1722.4		$\sigma = 16.77$ [0:1:30]	$\hat{\sigma} = 5.94$ [0:1:15] $\tau = 4.22$ [0:1:10]
Heuristic	−857.40	1731.2	$\delta = 0.105$ [0:0.01:0.3]	$\sigma = 2.44$ [0:1:30]	$\hat{\sigma} = 4.33$ [0:1:15]

Note: The range of parameters tested is in brackets.

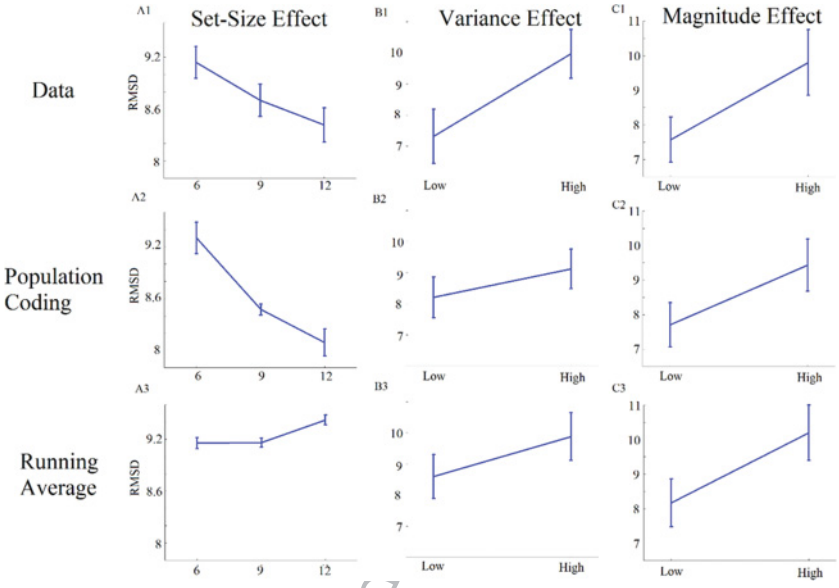

Figure 5: Data and model predictions of the main behavioral results. Top row: Behavioral data. Middle row: Population model. Lower row: Running average model. Models' predictions were generated using each model's best-fitting parameters. Error bars denote 1 within-participant standard error of the mean.

values (e.g., buying stocks, participating in lotteries) and is the basis for higher processes, such as preference formation (e.g., hiring an employee, choosing an apartment mate; Usher, Russo, Weyers, Brauner, & Zakay, 2011). In this study, we sought to characterize the mechanism that underlies the remarkable ability of humans to rapidly extract the average of a sequence of numerical values.

Toward this aim, we have tested how approximate numerical averaging is affected by the set size, variance, and magnitude of the numerical sequences. Using a specially designed experimental paradigm, we have shown that estimations improve as set size increases. Furthermore, we find that, similar to nonnumerical perceptual averaging processes, such as size and orientation averaging (e.g., Ariely, 2001; Chong & Treisman, 2003), numerical averaging markedly deteriorates as variance or magnitude increases. To the best of our knowledge, these two pivotal perceptual properties were never shown to apply directly to rapid numerical averaging. These results thus validate the assumption that approximate numerical averaging relies on perceptual-like processes rather than on purely symbolic (arithmetic) ones, and they provide important behavioral constraints for potential models aimed at accounting for approximate numerical averaging.

To account for approximate numerical averaging, we have developed a novel, biologically plausible population-coding model that solves the question of decoding the perceived average from the population response profile. We first demonstrated that this model is mathematically equivalent to a population vector or to the extraction of the center of mass. Next, we compared the population-coding model with competing mechanisms that are central to numerical cognition and perceptual processes (running average via leaky-integration and heuristic midrange calculation), using both qualitative and quantitative tests.

In our quantitative test, we fitted the models on a trial-by-trial basis and find that the population-coding model is superior to the alternative models in terms of both likelihood (which reflects the goodness of fit) and BIC (which penalizes additional degrees of freedom). Specifically, our data support a uniform (nonlogarithmic) representational space of the numerical tuning curves, with tuning curve variance that increases as a function of the numerical magnitude (Brannon, Wusthoff, Gallistel, & Gibbon, 2001; Dehaene, 2003).

In our qualitative test, we simulated the estimations of each model using the best-fitting parameters and tested the models' predictions of the reported behavioral effects. We find that only the population-coding model could account for all the behavioral patterns, thus lending additional strong support to the suggested population-coding model. The model's improvement with set size stems from the Poisson stochastic firing rate of the units responding to the presented numerical items. With each additional item, the noise averages out, resulting in a more reliable activity profile. The late (motor) noise, added to the evaluation, is sequence length independent and thus cannot account for improvement with set size. To account for the magnitude effect, the model assumes that larger numerical values are represented with broader tuning curves. One possible motivation behind this assumption is that larger numbers are relatively less frequent in everyday life (assuming that frequency determines the degree of precision with which the items are encoded; Yang & Maunsell, 2004).

A central assumption of our model is that dedicated feature-tuned neural populations are mapped onto a symbolic representation of numbers. This assumption relies on prior suggestions offered by Stanislas Dehaene and colleagues, who argue, based on theoretical and experimental considerations, that representations of quantities (such as numerosity) and magnitudes are being associated, through extensive learning, to representations of numerical (symbolic) values, yielding an approximate numerical system, which is akin to a perceptual system (Dehaene, 2011). Accordingly, the symbolic numerical values are being implicitly transformed into representations of quantity or magnitude, and these representations are those on which the perceptual-like system of population coding operates. Our behavioral and modeling results reported here offer indirect support for this assumption by demonstrating that approximate symbolic numerical

processing shows key properties of perceptual processing. Future studies should investigate this assumption by identifying the relations between perceptual representations of magnitude and quantities and those of symbolic numbers.

Although our model comparison results show strong preference for the population-coding model and are consistent with previous modeling results (Brezis et al., 2015, 2016), future studies should develop and test additional alternative models. Moreover, because previous research has indicated that temporal biases, such as recency or primacy, often occur in perceptual tasks (e.g., Kiani, Hanks, & Shadlen, 2008; Bronfman, Brezis, & Usher, 2016), it is likely that such biases will also be found in approximate numerical averaging. Indeed, a temporal regression analysis for our data shows that the evaluations are recency biased—a result that is consistent with prior reports (see the supplement and Brezis et al., 2015). While temporal biases were not directly addressed in our population-coding model, it is possible to extend the model to account for it. For example, one can introduce temporal decay in the activation buildup (currently implemented as a simple summation of the activation profiles). Such decay is likely to result in an increased influence of the later-arriving samples.[3] Future research should aim specifically at testing the question of temporal biases in numerical averaging and offer and test specific computational accounts for it. Of particular challenge is the fact that recency is diminished or even vanishes in the longer set size (of 12 items). Finally, while previous research has causally implicated the parietal cortex in approximate numerical averaging (Brezis et al., 2016), future studies should further test and characterize the brain regions that participate in this important human capacity.

References

Ahn, W. Y., Busemeyer, J. R., Wagenmakers, E. J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, *32*, 1376–1402.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*, 392–398.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

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

Ashby, N. J. S., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1153–1162.

[3] Note that since the running average model explicitly takes recency into account while the population model does not, the model comparison result in favor of the latter is a conservative one.

Beach, L. R., & Swenson, R. G. (1966). Intuitive estimation of means. *Psychonomic Science*, *5*, 161–162.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (2005). The Iowa Gambling Task and the somatic marker hypothesis: Some questions and answers. *Trends in Cognitive Sciences*, *9*, 159–162.

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*, 887–905.

Betsch, T., Plessner, H., Schwieren, C., & Gütig, R. (2001). I like it but I don't know why: A value-account approach to implicit attitude formation. *Personality and Social Psychology Bulletin*, *27*, 242–253.

Brannon, E. M., Wusthoff, C. J., Gallistel, C. R., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science*, *12*(3), 238–243.

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*, 1700–1713.

Brezis, N., Bronfman, Z. Z., & Usher, M. (2015). Adaptive spontaneous transitions between two mechanisms of numerical averaging. *Scientific Reports*, *5*, 10415.

Bronfman, Z. Z., Brezis, N., & Usher, M. (2016). Non-monotonic temporal-weighting indicates a dynamically modulated evidence-integration mechanism. *PLoS Comput. Biol.*, *12*, e1004667.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(2), 281–294.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*, 171–189.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*, 393–404.

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

Dehaene, S. (2001). Subtracting pigeons: Logarithmic or linear? *Psychol. Sci.*, *12*, 244–246 (discussion 247).

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, *7*, 145–147.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics.* New York: Oxford University Press.

Dotan, D., Friedmann, N., & Dehaene, S. (2014). Breaking down number syntax: Spared comprehension of multi-digit numbers in a patient with impaired digit-to-word conversion. *Cortex*, *59*, 62–73.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *233*, 1416–1419.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*, 517–523.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697.

Kiani, R., Hanks, T. D., & Shadlen, M. N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *Journal of Neuroscience*, *28*, 3017–3029.

Malmi, R. A., & Samson, D. J. (1983). Intuitive averaging of categorized numerical stimuli. *Journal of Verbal Learning and Verbal Behavior*, *22*, 547–559.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Attention, Perception, and Psychophysics*, *70*(5), 772–788.

Ossmy, O., Moran, R., Pfeffer, T., Tsetsos, K., Usher, M., & Donner, T. H. (2013). The timescale of perceptual evidence integration can be adapted to the environment. *Current Biology*, *23*(11), 981–986.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, *26*, 381–410.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Rusou, Z., Zakay, D., & Usher, M. (2017). Intuitive number evaluation is not affected by information processing load. In J. Kantola, T. Barath, S. Nazir, & T. Andre (Eds.), *Advances in human factors, business management, training and education* (pp. 135–148). New York: Springer.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology*, *7*, 82–138.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, *11*, 13.

Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *Frontiers in Neuroscience*, *6*, 70.

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences*, *109*, 9659–9664.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

Usher, M., Russo, Z., Weyers, M., Brauner, R., & Zakay, D. (2011). The impact of the mode of thought in complex decisions: Intuitive decisions are better. *Frontiers in Psychology*, *2*, 37.

Van Opstal, F., De Lange, F. P., & Dehaene, S. (2011). Rapid parallel semantic processing of numbers without awareness. *Cognition*, *120*, 136–147.

Van Overwalle, F., & Labiouse, C. (2004). A recurrent connectionist model of person impression formation. *Personality and Social Psychology Review*, *8*(1), 28–61.

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: a neural model. *Journal of Cognitive Neuroscience*, *16*, 1493–1504.

Wulff, D. U., & Pachur, T. (2016). Modeling valuations from experience: A comment on Ashby and Rakow (2014). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(1), 158–166.

Yang, T., & Maunsell, J. H. (2004). The effect of perceptual learning on neuronal responses in monkey visual area V4. *Journal of Neuroscience*, *24*(7), 1617–1626.

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin and Review*, *12*(3), 387–402.

Received April 30, 2017; accepted September 1, 2017.