# Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform

**Guido Marco Cicchini[a], Giovanni Anobile[b], and David C. Burr[a,b,1]**

[a]Neuroscience Institute, National Research Council, 56124 Pisa, Italy; and [b]Department of Neuroscience, Psychology, Pharmacology and Child Health, University of Florence, 50135 Florence, Italy

The mapping of number onto space is fundamental to measurement and mathematics. However, the mapping of young children, unschooled adults, and adults under attentional load shows strong compressive nonlinearities, thought to reflect intrinsic logarithmic encoding mechanisms, which are later "linearized" by education. Here we advance and test an alternative explanation: that the nonlinearity results from adaptive mechanisms incorporating the statistics of recent stimuli. This theory predicts that the response to the current trial should depend on the magnitude of the previous trial, whereas a static logarithmic nonlinearity predicts trialwise independence. We found a strong and highly significant relationship between numberline mapping of the current trial and the magnitude of the previous trial, in both adults and school children, with the current response influenced by up to 15% of the previous trial value. The dependency is sufficient to account for the shape of the numberline, without requiring logarithmic transform. We show that this dynamic strategy results in a reduction of reproduction error, and hence improvement in accuracy.

numerical cognition | predictive coding | approximate number system | Weber–Fechner law | serial dependency

Humans have a strong intuition of the spatial nature of numbers, usually (but not always) a horizontal "mental numberline," with numbers increasing from left to right (1–4). However, the nature of number mapping is not identical for all, but changes during development, starting from a nonlinear representation, well characterized as logarithmic (placing, for example, the number 10 near the midpoint of a 1–100 scale), then becoming more linear over the first years of schooling (3, 5, 6). Similarly, logarithmic-like numberlines have been demonstrated in indigenous Amazonian populations without formal mathematical schooling (4).

Several recent studies have shown that under certain circumstances even the math-educated tend to reproduce numbers logarithmically. For example, we showed that depriving attentional resources leads to logarithmic-like numberline responses (7), consistent with the possibility that the native logarithmic encoding emerges when attention is deprived. Other studies have shown that the use of unfamiliar numerical format (such as exponential) can induce a switch from a linear to a logarithmic-like response, even in math-educated adults (7, 8). Most recently, Dotan and Dehaene (9) have devised a clever technique to record the whole trajectory of the pointing response (across the face of a touchscreen), rather than just the endpoint: The response begins quite logarithmically, then corrects toward linear mapping by the time contact is made. All these studies have led many to interpret the logarithmic map as the direct reflection of the internal native number representation (4, 10–12) that becomes corrected over time by education but can emerge under special circumstances.

Whereas the nonlinear numberline is consistent with intrinsic logarithmic processes, other explanations have been suggested. One promising possibility is that the nonlinearity results from

a "central tendency of judgment" or "regression toward the mean" (7, 13, 14), which has been successfully applied to many perceptual tasks (15) and recently well described within the Bayesian framework (7, 16): Under conditions of uncertainty—such as under attentional load or unfamiliar numerical format—responses tend to be biased toward the mean of the stimulus distribution. Importantly, strategies of this sort can lead to reductions in error, particularly under conditions of high uncertainty or noise (7, 16–19). In the numberline task, regression toward the mean predicts a logarithmic pattern of results (7, 13), with a goodness of fit similar to a log-linear model (7).

How can we distinguish between these two plausible classes of explanations of the numberline nonlinearities? One major difference is that whereas the logarithmic transform is a static nonlinearity, regression to the mean is a dynamic process that requires continuous online updating. A strong prediction, therefore, is that there should exist serial dependencies between the response to the current stimulus and the strength of previous stimuli. In this study we measure intertrial dependencies and demonstrate strong and significant correlations between the current response and previous stimuli, clearly favoring the central tendency explanation of the nonlinear numberline. We go on to develop a simple Bayesian integration model to show how the intertrial dependencies can explain completely the nonlinear numberline, without resort to logarithmic encoding or other static nonlinearities.

## Results

We measured numberline mapping under high and low attentional demand. Five subjects were asked to indicate the numerosity

> **Significance**
>
> The ability to map numbers onto space is fundamental to measurement and mathematics. The mental "numberline" is an important predictor of math ability, thought to reflect an internal, native logarithmic representation of number, later becoming linearized by education. Here we demonstrate that the nonlinearity results not from a static logarithmic transformation but from dynamic processes that incorporate past history into numerosity judgments. We show strong and significant correlations between the response to the current trial and the magnitude of the previous stimuli and that subjects respond with a weighted average of current and recent stimuli, explaining completely the logarithmic-like nonlinearity. We suggest that this behavior reflects a general strategy akin to predictive coding to cope adaptively with environmental statistics.

of clouds of dots on a line demarcated by a single dot on the left and a 100-dot cloud on the right. In the single-task condition they simply responded to the numerosity, in the dual-task condition they first responded to a difficult visual conjunction task, then to the numerosity (*Methods* and ref. 7). The symbols of Fig. 1 show the average results for single-task (Fig. 1*A*) and dual-task (Fig. 1*B*) conditions. Agreeing with previous studies (7, 14), the data show a less linear pattern under conditions of high attentional load. The thin black curves show the best linear and log fits and the red a mixed model combining linear and logarithmic components:

$$R = a\left((1-\lambda)N + \lambda\frac{N_{max}}{\ln(N_{max})}\ln(N)\right),\qquad [1]$$

where $R$ is the response to numerosity $N$, $a$ is a scaling factor, $N_{max}$ the end of the numberline (equal to 100), and $\lambda$ a factor describing the logarithmic nonlinearity (0 for pure linear, 1 for pure logarithmic). This two-parameter fit was quite good (fitting parameters are given in the legend to Fig. 1). The logarithmic component $\lambda$ was 0.11 for the single-task condition, and 0.38 for the double task, reflecting the nonlinear number mapping under attentional load. The blue curves show the predictions of a Bayesian integration model, described later.
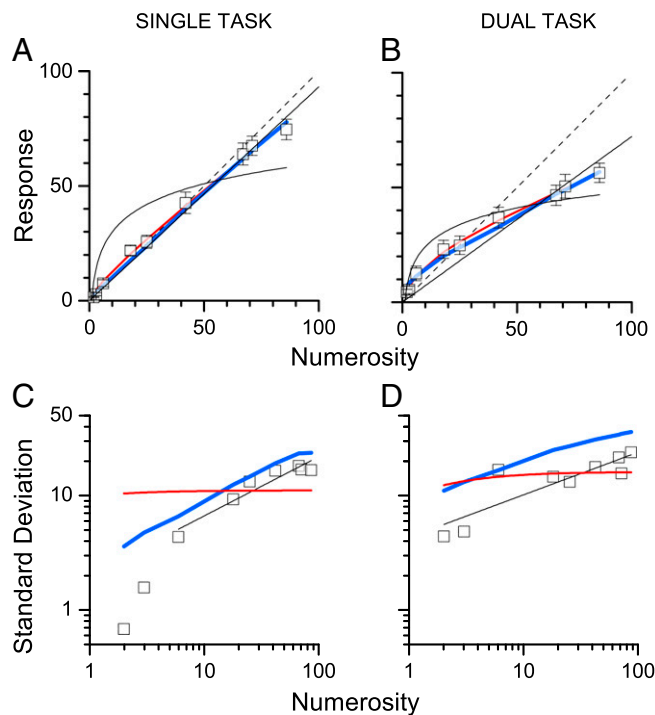


**Fig. 1.** Mapping number to space. The square black symbols show the average response location of subjects in the single-task (*A*) and dual-task (*B*) conditions. The thin black lines show the best-fitting linear and logarithmic models. The blue and red lines show, respectively, the best-fitting predictions for the Bayesian (Eq. **2**) and log-linear models (Eq. **1**). (*C* and *D*) SDs of responses as a function of number presented (on log axes). The thin black lines show the best-fitting linear regressions (outside the subitizing range), defining the index of the power law ($\alpha$ of Eq. **5**), used for the modeling. The blue and red curves show, respectively, the predicted SDs from the Bayesian and log-linear models. Goodness of fit was assessed by the coefficient of determination ($R^2$) and the Akaike information criterion (AIC), which takes into account degrees of freedom (45). The values of $R^2$ and AIC (respectively) for the log-linear model were 0.79 and 16.4 (single) and 0.94 and 27.0 (double). For the Bayesian model they were 0.72 and 17.0 (single) and 0.97 and 16.7 (double). Smaller AIC values denote better fits, with differences of 2 considered negligible.
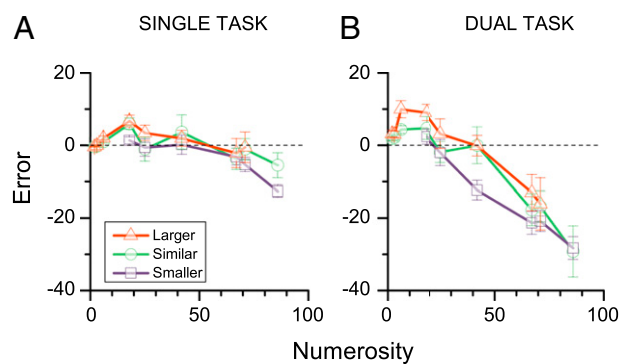


**Fig. 2.** The effect of previous stimulus magnitude. (*A* and *B*) Average response error (difference between response and stimulus magnitude) for trials parsed into three categories based on the magnitude of the previous stimulus: purple, previous stimulus at least seven numbers less; green, previous stimulus within seven of the current one; red, previous stimulus at least seven higher. The magnitude of the previous stimulus had a clear effect, particularly in the dual-task condition, and was statistically significant in both conditions.

Fig. 1 *C* and *D* show the SDs for the numberline judgments under the two conditions (on log–log coordinates). These can be taken as estimates of the thresholds for localization on the numberline. The thick lines show the output of two models (discussed later), and the thin straight lines best-fitting regressions (for use in modeling). In fitting the data, we excluded the subitizing range, because there is very good evidence that different mechanisms operate over that range (2). Furthermore, subitizing mechanisms are attention-dependent, operating only when there are sufficient attentional resources (20, 21). For that reason, it seemed safest to exclude those measurements from the fits of single-task data.

**Trialwise Dependencies.** The simplest direct prediction of the central-tendency explanation is that responses to trials preceded by a less numerous stimulus should on average be lower than responses to trials preceded by a more numerous stimulus. Fig. 2 shows that this prediction is borne out. The red points show response errors for when the previous trial was greater (by at least 7), blue points where it was at least 7 less, and the green points where it was similar ($|\Delta| < 7$). For the dual-task condition the errors for trials preceded by larger numbers were consistently higher than those for trials preceded by smaller numbers, with the "similar" category falling in between. The difference between the greater-than and less-than conditions is clearly significant [repeated measures ANOVA for the numerosities covered by both, $F_{(1,4)} = 29.9$, $P < 0.005$]. For the single-task condition (which showed little logarithmic tendency), the curves again separate (although less obviously), and again the difference is significant [$F_{(1,4)} = 15.9$, $P = 0.016$].

We next looked for more quantitative dependencies between the magnitude of the current responses and previous stimuli. Fig. 3*A* shows responses at four sample numerosities as a function of the numerosity of the previous trial for the single- (*Left*) and dual-task (*Right*) conditions. At low subitizing numerosities, the previous trials had very little effect, but at higher numerosities there was a clear dependency on the previous trial, with responses varying almost monotonically with the magnitude of previous stimulus. The dashed, color-coded lines show the robust linear regressions of the data, all of which have positive slope. The thick lines (in this and the other figures) are model predictions, discussed later.

We take the slope of these regressions as an index of the response dependency on previous stimuli and plot them in Fig. 3*B* as a function of the numerosity of the current trial (black squares). The dependencies are highest at midhigh numerosities, falling off at the lowest and the highest numerosities. We averaged the weights of Fig. 3*B* for the range of numerosities greater than 5
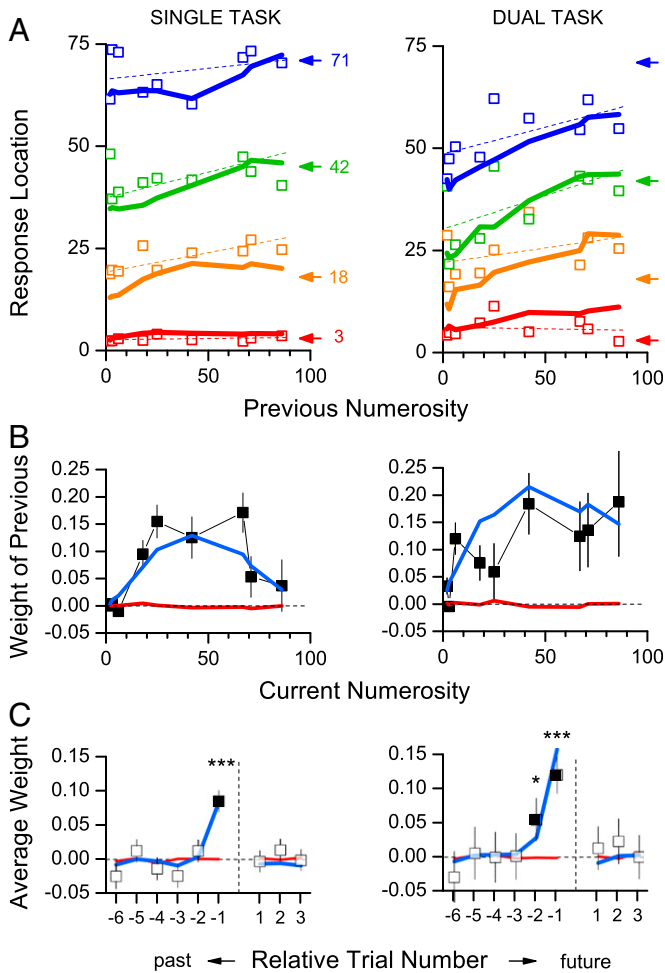
**Fig. 3.** Predicted and measured serial dependencies. (*A*) Response as a function of numerosity of the previous trial for single-task (*Left*) and dual-task (*Right*) conditions for the experiment data (□), together with the predictions from the Bayesian integration model of Eq. **2** (heavy color-coded lines), for the four representative numerosities. Values of $k$ are 2.75 for the single-task condition and 12.6 for the dual-task. The thin straight lines show best-fitting linear regressions to the data. The arrows show the veridical responses. (*B*) Average dependency (weight) on previous stimuli, as a function of current numerosity, given by the slope of regression lines such as those above. The thick lines show the predicted weights for the log-linear (Eq. **1**, red) and Bayesian (blue) models. The black symbols show the correlations in the data for the three datasets. The error bars mark ±1 SEM. (*C*) Average weights as a function past and future trial number, for the data (open squares) and the two model predictions (Bayesian integration blue, log-linear red). Asterisks indicate significance from zero: ***$P < 10^{-5}$, *$P < 0.05$, one-tailed bootstrap sign test.

(outside the subitizing range) and plot them as a function of past and future trials in Fig. 3*C*. The average weights of trial $i-1$ were strong (w = 0.08 for single-task, w = 0.12 for dual-task), and highly significant ($P < 10^{-5}$, bootstrap sign test). For the dual-task condition, there was also a significant dependency for trial $i-2$ (w = 0.04, $P = 0.04$, bootstrap sign test). That for the single-task condition was also positive (w = 0.01) but not significantly greater than zero. Importantly, there were no significant dependencies on future trials, a strong control against statistical artifacts.

**Bayesian Integration Model.** To show how intertrial dependencies can predict the nonlinear numberline, we implement a simple Bayesian integration model whose output is shown by the blue lines in Figs. 1 and 3. This model should be considered more of an existence proof to show that correlations can lead to

nonlinearities, rather than attempting to describe actual physiological mechanisms. The model resembles a Kalman filter, in that we assume that the expected response to any given trial $i$ is given by the weighted average of the numerosity of the current stimulus and the estimate of that of the past:

$$R_i = (1 - W_{i-1})N_i + W_{i-1}R_{i-1}, \qquad [2]$$

where $W_{i-1}$ is the weight assigned to the estimate of the numerosity of previous trial. In Bayesian terms this can be considered as a prior from previous stimulus history. Because subjects do not have direct access to the physical magnitude of the previous trial, we assume that the best estimate of it is the response $R_{i-1}$. Because $R_{i-1}$ in turn depends on $R_{i-2}$, the formula is clearly recursive, so it will accumulate evidence from the past without accessing the full stimulus history, or a rolling average of it.

How do we define the weight $W_{i-1}$? For two redundant cues, the weight of each cue should be proportional to its reliability $\rho$, the inverse of variance:

$$W_{i-1} = \frac{\rho_{i-1}}{\rho_i + \rho_{i-1}} = \frac{1/\sigma_{i-1}^2}{1/\sigma_i^2 + 1/\sigma_{i-1}^2} = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_{i-1}^2}. \qquad [3]$$

This formula expresses ideal weighting when trials $i$ and $i-1$ sample the same physical stimulus. In this case they do not, and the probability that they are different can be shown to vary with the square of their separation (22, 23). The formula now becomes

$$W_{i-1} = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_{i-1}^2 + (N_i - R_{i-1})^2}. \qquad [4]$$

We assume that the SD (thresholds) will, like most sensory discriminations, follow a power law:

$$\sigma = kN^\alpha, \qquad [5]$$

where $\alpha$ is the index of the power law and $k$ a free constant (the only degree of freedom in the model), giving the overall level of noise. Substituting Eq. **5** in Eq. **4** the weight given to the previous trial is

$$W_{i-1} = \frac{k^2 N_i^{2\alpha}}{k^2 N_i^{2\alpha} + k^2 N_{i-1}^{2\alpha} + (N_i - R_{i-1})^2}. \qquad [6]$$

Fig. 4 shows the general behavior of the model for various values of $\alpha$ and $k$. Note that $\alpha = 1$ (threshold proportional to $N$) describes Weber's law, $\alpha = 0.5$ describes a square-root relationship, commonly referred to as "shot-noise" or "Poisson noise," and $\alpha = 0$ describes constant noise, invariant with numerosity. The curves for different $\alpha$ (Fig. 4*A*) all show strong logarithmic-like compression. The major effect of changing the index is that for high $\alpha$ the deviation from veridicality is mainly at high numbers, whereas for lower values there is also a deviation toward the mean at low numbers (clearer on the expanded insert). It is clear that the general pattern of results does not depend on a specific noise regime. Even the implausible assumption of constant noise ($\alpha = 0$) causes a strong regression to the mean, logarithmic-like over much of the range. Simply adding the constraint of veridicality in the subitizing range would make the function quite like a logarithmic transform.

In the simulations of Fig. 4*A* the free parameter $k$ was adjusted to best fit the data in the dual-task condition (which showed the most nonlinearity). Fig. 4*B* shows the effect of varying noise level $k$ (fixing $\alpha$ at 0.36, the measured value for the dual-task condition). With increasing $k$, the functions become more curved, deviating more from veridicality in a logarithmic-like manner.

**Fig. 4.** Predictions of the Bayesian model. (*A*) Predicted numberlines for various power-law noise functions (Eq. **5**), for $\alpha$ ranging from 0 (constant noise) to 1. All values predict a negatively accelerating numberline, with the main effect of varying $\alpha$ visible at low numerosities (*Inset*): Low $\alpha$ causes greater deviation from veridicality at these low numerosities. The value of $k$ was chosen to best fit the data of the dual-task condition ($k = 12.6$). (*B*) The effect of varying the noise level $k$ (for $\alpha = 0.36$, the level best-fitting from the dual-task data). Increasing $k$ causes greater deviations for veridicality.

**Modeling the Data.** To model our numberline data, we calculated the value of the power-law index ($\alpha$ in Eqs. **5** and **6**) from the data (Fig. 1 *C* and *D*) by linear regression. This yielded a value of 0.52 (shot-noise) for the single-condition data and 0.36 for the dual-task condition. Note, however, that although using the data to determine $\alpha$ seemed the most assumption-free strategy, Fig. 4*A* shows that the choice of $\alpha$ is not fundamental for the general pattern of results. We then adjusted parameter $k$ (the overall level of noise) to give the best fits to the data. With only this one degree of freedom, the model captures the nature of the number-line data with a comparable or better fit than the two-parameter log-linear model (fit parameters in the legend to Fig. 1).

The blue lines of Fig. 3 show Monte Carlo simulations of the model, obtained by simulating 1,000 virtual subjects, assuming that their estimates of $N_i$ were corrupted by Gaussian noise of SD given by Eq. **5** ($\alpha$ and $k$ given by data and best fit, respectively). Fig. 3*A* shows simulated responses as a function of magnitude of previous response. In general, the pattern of results is similar to the data, with clear dependencies at all except the lowest numerosity. We then calculated linear regressions of the simulations and plotted these with those of the data in Fig. 3*B*. Again the model captures the trend in the data: low for low numerosities, higher for mid and high levels. It even captures the tendency of the single-task data to show weaker dependencies at the highest numbers. Fig. 3*A* show the predictions of overall dependencies, again capturing the trend in the data, predicting strong dependencies in both conditions for trial $i{-}1$ and also a measureable dependency at trial $i{-}2$, as was observed. The red curves of Fig. 3 *B* and *C* show the simulations of a memory-free system with partial logarithmic encoding (using the values of the fit of Eq. **1**, and the same values of $\alpha$ and $k$ as for the Bayesian model). As may be expected, the static linearity predicted zero correlations in all conditions. Although this prediction is obvious, it serves as a sanity check to ensure that the programs functioned correctly without introducing spurious, artifactual correlations.

Finally, we obtained fits to the data by simulating the behavior of 1,000 virtual subjects performing experimental sessions like the subjects, randomizing trial order for each simulation. For each trial the response was calculated as a weighted average of the current stimulus (corrupted by Gaussian noise with SD given by Eq. **5**) and the previous response:

$$R_i = (1 - W_{i-1})\left(N_i + \mathcal{N}\left(0, kN_i^\alpha\right)\right) + W_{i-1}R_{i-1}, \qquad [7]$$

where $\mathcal{N}$ represents the normal distribution and $W_{i-1}$ follows Eq. **6**. Responses that would have exceeded the numberline (because of a large draw of noise) are clipped at the boundary of the line. The current response is stored and retrieved on the following simulation without further noise corruption.

The blue curves of Fig. 1 show the simulated responses (*A* and *B*) and SDs (*C* and *D*). Clearly, the simulation describes the data well, particularly in the dual-task condition where it captures 97% of the variance (see the Fig. 1 legend for fit parameters). The thick blue lines of Fig. 1 *C* and *D* show the SDs of the simulated responses. These follow a trend similar to the measured SDs, approximating a power law of similar slope to the data, but slightly higher: by a factor of 1.2 for the single task and 1.7 for the dual task. This implies that the system assumes a slightly higher noise level than actually exists. The red lines show the predictions of the log-linear model, assuming that precision is dictated by the inverse of the slope of the log-linear Eq. **1**. In practice this leads to Weber's law when the model is fully logarithmic ($\lambda = 1$) and to constant noise when the mixture model is fully linear ($\lambda = 0$). However, we do not take the failure in predicted noise levels as strong evidence against the logarithmic model.

**Potential Advantages of Bayesian Integration.** Bayesian strategies are usually thought to be statistically advantageous, reducing the variability of sensory estimates (e.g., ref. 16). Fig. 5 shows how this may apply in this case. We simulated rms error as a function of the noise constant ($k$ of Eq. **5**) for a memoryless linear system and the Bayesian integration model. The Bayesian model predicts less error than the memoryless model, and the difference increases with noise level. The symbols studded into the lines show the noise level that best fit the dual-task adult condition.

Fig. 5*B* gives an intuition of how the weighted average can lead to an advantage. We take the example of adult responses under attentional load. Following Jazayeri and Shadlen (16), we partition the error into two orthogonal components, the bias (average accuracy) given by the distance of the average response from veridicality (the distance of the points of Fig. 1 from the equality line), and the precision, given by the SD of the scatter of responses around that mean. The Pythagorean sum of these two components gives the total error, on this graph represented by



**Fig. 5.** The effect of the Bayesian integration model on error. (*A*) Predictions of how error increases with noise for a memoryless system (red) and the Bayesian integration model (blue). The Bayesian model predicts less error, with the difference increasing as the noisiness of the system increases. The symbols indicate $k = 12.6$, the level that best fit the dual-task data. (*B*) The abscissa shows the bias (average deviation from veridicality) and the ordinate the precision (SD of the scatter of responses around their mean) for the dual-task adult data. Each data point refers to a different numerosity. The red curve shows the prediction of a memoryless linear model and the blue curve that of the Bayesian model (Eq. **2**). The black symbols show the data. The Bayesian model captures the pattern of the data: Both the model and the data show less total error (distance from origin) than the memoryless model.

the distance from the origin. While a memoryless linear system is essentially bias-free (near vertical red curve), both the Bayesian model and the data show increasing bias for larger numbers. The result of this is that the total error (distance from the origin) is about 33% less than that predicted by a memoryless system with the same amount of noise. The Bayesian model predicts the nature of the error, with positive bias for low numbers and negative for high. The total error predicted is about 1.7 times more than that actually observed (as we have seen in Fig. 1), suggesting that the system overestimates its internal noise. However, besides this small overestimation the model captures the pattern of the data.

**Nonlinear Number Mapping in Children.** Finally, we asked whether the current model may be useful in explaining previously reported numberline data for school-aged children. To this end, we reanalyzed the previously published (24) numberline data of 68 primary school children (ages 8–11 y). As with the adult data, we found a clear and statistically significant dependency of the current trial on trial $i-1$ ($P = 0.002$). The dependency on trial $i-2$ was not significant. A split-half analysis like that of Fig. 2 also revealed a significant difference between trials preceded by a lower numerosity and those preceded by a higher numerosity [$F_{(1,3)} = 35.9$, $P = 0.009$].

## Discussion

Two general conclusions can be drawn from this study. The first is that the characteristic nonlinearities in numberline mapping of young or unschooled subjects, or adults under attentional load, need not result from logarithmic encoding of number. We are able to account completely for the nonlinearities observed in three sets of data with a simple Bayesian model that performs a linear weighted sum between present and past stimuli. Obviously we cannot exclude that static nonlinearities also occur, but the nonlinear numberline cannot be taken as evidence for them. Second—and perhaps more importantly—we suggest that taking stimulus history into account may reflect a very general strategy of optimizing behavior to take into account environmental statistics. Because the physical world is largely stable and continuous over time, the recent past is a good predictor of the present (25). As Fischer and Whitney (26) have recently suggested for orientation, serial dependence of estimates of numerosity may reflect a basic mechanism to improve the efficiency of numerosity estimation.

**Logarithmic Transformation of Stimulus Magnitude.** Much evidence has been thought to reflect logarithmic encoding of number, including the approximation to Weber's law (thresholds proportional to numerosity) of range for both humans (27) and monkeys, and the logarithmic bandwidth of neurons selective to number (11, 28, 29). However, compressive nonlinearities do not necessarily implicate intrinsic logarithmic encoding (30–32). This is a very old controversy in psychophysics, dating back to Weber and Fechner, who interpreted the proportionality of sensory thresholds to stimulus intensity (now called the Weber–Fechner law) to reflect logarithmic processing (as the derivative of the logarithm is $1/x$) (33, 34). However, this was famously challenged by Stevens (35), who showed that the logarithmic-like behavior holds only at threshold. Weber–Fechner behavior for sensory attributes is now rarely interpreted to imply logarithmic encoding, usually adaptation or gain control, mechanisms that have been linked to optimal behavior (26, 36–38). So, the strongest evidence for logarithmic coding was the logarithmic numberline: Because that now has a more plausible explanation, there exists no evidence at all for logarithmic encoding of number in primate brains. Just as the dependency of increment thresholds on magnitude led to erroneous assumptions of logarithmic transformation of sensations, rather than dynamic gain control, so too has the logarithmic-like shape of the numberline led to assumptions of logarithmic encoding of number, rather than dynamic adaptive coding.

We also point out that although Weber's law is often assumed it is seldom actually observed. In a recent study (39) we showed that for discrimination of numerosity Weber's law applied over only a very limited interval, with the square-root law ($\alpha = 0.5$) operating over most of the range. In the current study, which required a pointing response (which presumably also introduced some noise), the power law exponents ranged from 0.36 to 0.5, well outside the Weber range. Although this is not particularly strong evidence against logarithmic encoding, it certainly does not support the concept. And recent evidence that monkeys can perform simple additions almost linearly (40) also speaks against logarithmic encoding.

**Modeling.** We model our results with a simple Bayesian-like model, which incorporates a weighted estimate of past stimuli in a recursive way, so responses tend to be drawn toward the mean. The degree to which they deviate depends on three factors: the estimated reliability (inverse variance) of the current stimulus, that of the previous stimulus, and the difference in magnitudes between the two (Eq. **4**).

The model (Eq. **2**) resembles closely a Kalman filter, with the weighting to previous stimuli acting like Kalman gain. However, we do not attach any particular significance to this similarity. Kalman filters are typically used to stabilize and maintain calibration in control systems by minimizing the difference between predicted and observed states. The filter does not completely recalibrate each time a difference between prediction and observation occurs, because that would render it unstable. In our case it is not clear why a Kalman filter should apply, as it is being stabilized or calibrated. Perhaps the apparent similarity of our model to a Kalman filter merely results from the fact that both incorporate a running average in the final estimate of magnitude. We believe that the particular form of modeling we have chosen is not unique, but acts as an existence proof to show how trialwise correlations can lead to logarithmic-like distortions.

The model could certainly be refined. For example, there is no active memory in estimating the mean. Stimuli older than 1-back affect the predictions only because of the innate recursiveness of the model: The best estimate of the magnitude of trial $i-1$ is considered to be the response to it, which includes a weighting from trial $i-2$, etc. Indeed, in the condition with the highest noise (dual task), we do find a significant dependency on trial $i-2$. However, it is conceivable that the system may actively construct an estimate of the mean as the prior, as has been suggested in previous work (7, 16, 18). It would be relatively simple to extend the model to incorporate a running average of the mean, which may lead to further improvements in the fits.

Perhaps the most important question is, Why should the current responses depend on the magnitude of previous stimuli? Within the Bayesian framework, the use of priors has typically been considered to be optimal. Indeed, also in this case, the prior does lead to an improvement of overall accuracy (Fig. 5). However, the improvement is not great, only about 33%. Is this a large enough advantage to be driving this effect? Possibly, but it is also possible that other factors are at work. For example, the dependencies may reflect an inherent hysteresis of the system. Indeed, priming (positive dependency on past events) is a ubiquitous phenomenon in psychology (41, 42), which may well reflect a general strategy to cope adaptively in the natural environment (37, 43).

Some readers may wonder why we did not look for correlations between responses, rather than stimuli. Because Eq. **2** predicts a dependency on the estimate of the previous numerosity, approximated by the response to it ($R_{i-1}$), it may seem sensible to search for dependencies on previous responses, rather than stimuli. However, this is psychophysically impractical, because response biases could lead to spurious correlations between successive responses. For example, a subject first responding consistently too low, then consistently too high, would produce a spurious correlation between neighboring trials. However, no such behavior can lead to spurious correlations between previous stimuli and current responses, because each stimulus was drawn independently.

As the main component of $R_{i-1}$ is given by $N_{i-1}$, showing a dependency of $R_i$ on $N_{i-1}$ is the strongest support possible for the model. That there is no dependency on $N_{i+1}$ is very strong proof against artifacts producing spurious correlations.

In our study we chose to use clouds of dot stimuli, because their sensory characteristics are more easily defined. However, many numberline tasks use symbolic stimuli, such as Arabic digits, and the logarithmic form of the numberline has been shown to be predictive of math performance in school-age children. We have yet to investigate whether the coding of symbolic numbers may be influenced by past history. Because the model does not depend on any particular power-law relationship, but works even with constant noise ($\alpha = 0$), there is no reason why the model should not predict performance in this situation too. If similar effects are observed, it may provide insights into why nonlinear numberline performance is associated with poor math performance.

## Conclusions

To summarize, we have shown that when mapping number to space subjects adjust responses to take into account recent history, fitting well with suggestions that the spatial representation of numbers is not an inert map but is a highly dynamic process (8). If this notion is correct, we are in a position to make predictions for a wide range of research. For example, our model predicts the serial dependency recently reported for perception of orientation (26). We also predict serial dependencies in other tasks, such as Dotan and Dehaene's (9) demonstration that numberline mapping responses are initially toward the "logarithmic" goal, before being corrected toward the linear response. We predict that the early responses should be highly correlated with prior stimuli, reflecting the intrinsic tendency to incorporate the prior, which is steadily reduced as evidence accumulates. We also expect to find strong trialwise correlations in many other experimental conditions, such as time reproduction (16) and even causality (44).

## Methods

Five adults naïve to the purpose of the study (mean age 26 y), all with normal or corrected-to-normal vision, participated in the study. Participants gave written informed consent. The experiments were approved by the local ethics committee (Azienda Ospedaliero-Universitaria Pisana n. 45060). Stimuli and procedures were similar to those in typical numberline experiments, described in detail elsewhere (e.g., ref. 7). Briefly, participants were presented with a cloud of dots and asked to indicate the quantity on a line demarcated by two sample numerosities. Each trial started with participants viewing a 22-cm "numberline" that remained visible throughout the trial with sample dot clouds representing the extremes: one dot on the left of the numberline and 100 on the right. Dot stimuli (half black, half white) were presented for 240 ms (in a circular region of 8° diameter) and were followed by a random-noise mask. The numerosities were 2, 3, 6, 18, 25, 42, 67, 71, 86; subjects responded by mouse click. As described in ref. 7, in the dual-task condition adults performed the task together with a color-orientation conjunction task on the central squares, before making the number-line judgment. In the single-task condition, everything was identical except they ignored the color-orientation stimuli and responded only to the number task. In each session, subjects were presented with two repetitions of each stimulus, a total of 18 trials, presented in random order. All subjects performed eight blocks of 18 trials each, randomly intermingling single- and dual-task conditions. No feedback was provided in any condition.

1. Galton F (1880) Visualized numerals. *Nature* 21:252–256, 494-255.
2. Dehaene S (1997) *The Number Sense: How the Mind Creates Mathematics* (Oxford Univ Press, Oxford).
3. Siegler RS, Booth JL (2004) Development of numerical estimation in young children. *Child Dev* 75(2):428–444.
4. Dehaene S, Izard V, Spelke E, Pica P (2008) Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science* 320(5880): 1217–1220.
5. Siegler RS, Opfer JE (2003) The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychol Sci* 14(3):237–243.
6. Booth JL, Siegler RS (2006) Developmental and individual differences in pure numerical estimation. *Dev Psychol* 42(1):189–201.
7. Anobile G, Cicchini GM, Burr DC (2012) Linear mapping of numbers onto space requires attention. *Cognition* 122(3):454–459.
8. Chesney DL, Matthews PG (2013) Knowledge on the line: Manipulating beliefs about the magnitudes of symbolic numbers affects the linearity of line estimation tasks. *Psychon Bull Rev* 20(6):1146–1153.
9. Dotan D, Dehaene S (2013) How do we convert a number into a finger trajectory? *Cognition* 129(3):512–529.
10. Dehaene S (2002) Neuroscience. Single-neuron arithmetic. *Science* 297(5587):1652–1653.
11. Dehaene S (2003) The neural basis of the Weber-Fechner law: A logarithmic mental number line. *Trends Cogn Sci* 7(4):145–147.
12. Dehaene S (2007) Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. *Sensorimotor Foundations of Higher Cognition*, eds Haggard P, Rossetti Y, Kawato M (Harvard Univ Press, Cambridge, MA), pp 527–574.
13. Karolis V, Iuculano T, Butterworth B (2011) Mapping numerical magnitudes along the right lines: Differentiating between scale and bias. *J Exp Psychol Gen* 140(4):693–706.
14. Anobile G, Turi M, Cicchini GM, Burr DC (2012) The effects of cross-sensory attentional demand on subitizing and on mapping number onto space. *Vision Res* 74:102–109.
15. Hollingworth HL (1910) The central tendency of judgements. *J Philos Psychol Sci Meth* 7(17):461–469.
16. Jazayeri M, Shadlen MN (2010) Temporal context calibrates interval timing. *Nat Neurosci* 13(8):1020–1026.
17. Körding KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. *Nature* 427(6971):244–247.
18. Petzschner FH, Glasauer S (2011) Iterative Bayesian estimation as an explanation for range and regression effects: A study on human path integration. *J Neurosci* 31(47): 17220–17229.
19. Cicchini GM, Arrighi R, Cecchetti L, Giusti M, Burr DC (2012) Optimal encoding of interval timing in expert percussionists. *J Neurosci* 32(3):1056–1060.
20. Vetter P, Butterworth B, Bahrami B (2008) Modulating attentional load affects numerosity estimation: Evidence against a pre-attentive subitizing mechanism. *PLoS ONE* 3(9):e3269.
21. Burr DC, Turi M, Anobile G (2010) Subitizing but not estimation of numerosity requires attentional resources. *J Vis* 10(6):20.
22. Roach NW, Heron J, McGraw PV (2006) Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. *Proc Biol Sci* 273(1598): 2159–2168.
23. Ernst MO (2007) Learning to integrate arbitrary signals from vision and touch. *J Vis* 7(5):1–14.
24. Anobile G, Stievano P, Burr DC (2013) Visual sustained attention and numerosity sensitivity correlate with math achievement in children. *J Exp Child Psychol* 116(2):380–391.
25. Dong DW, Atick JJ (1995) Statistics of natural time-varying images. *Network* 6(3):345–358.
26. Fischer J, Whitney D (2014) Serial dependence in visual perception. *Nat Neurosci*, 10.1038/nn.3689.
27. Jevons WS (1871) The power of numerical discrimination. *Nature* 3:363–372.
28. Nieder A (2005) Counting on neurons: The neurobiology of numerical competence. *Nat Rev Neurosci* 6(3):177–190.
29. Nieder A, Merten K (2007) A labeled-line code for small and large numerosities in the monkey prefrontal cortex. *J Neurosci* 27(22):5986–5993.
30. Gallistel CR, Gelman R (1992) Preverbal and verbal counting and computation. *Cognition* 44(1–2):43–74.
31. Brannon EM, Wusthoff CJ, Gallistel CR, Gibbon J (2001) Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychol Sci* 12(3):238–243.
32. Gallistel CR, Gelman R (2005) Mathematical Cognition. *Cambridge Handbook of Thinking and Reasoning*, eds Kolyoak KJ, Morrison R (Cambridge Univ Press, Cambridge, UK), pp 559–588.
33. Weber EH (1834) *De pulsu, resorptione, auditu et tactu. Annotationes anatomicae et physiologicae* (Koehler, Leipzig, Germany).
34. Fechner GT (1860) *Elemente der Psychophysik* (Breitkopf und Härtel, Leipzig, Germany).
35. Stevens SS (1957) On the psychophysical law. *Psychol Rev* 64(3):153–181.
36. Barlow H (1990) The mechanical mind. *Annu Rev Neurosci* 13:15–24.
37. Chopin A, Mamassian P (2012) Predictive properties of visual adaptation. *Curr Biol* 22(7):622–626.
38. Gepshtein S, Lesmes LA, Albright TD (2013) Sensory adaptation as optimal resource allocation. *Proc Natl Acad Sci USA* 110(11):4368–4373.
39. Anobile G, Cicchini GM, Burr DC (2014) Separate mechanisms for perception of numerosity and density. *Psychol Sci* 25(1):265–270.
40. Livingstone MS, et al. (2014) Symbol addition by monkeys provides evidence for normalized quantity coding. *Proc Natl Acad Sci USA* 111(18):6822–6827.
41. Maljkovic V, Nakayama K (1994) Priming of pop-out: I. Role of features. *Mem Cognit* 22(6):657–672.
42. Maljkovic V, Nakayama K (1996) Priming of pop-out: II. The role of position. *Percept Psychophys* 58(7):977–991.
43. Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. *Philos Trans R Soc Lond B Biol Sci* 364(1521):1211–1221.
44. Rolfs M, Dambacher M, Cavanagh P (2013) Visual adaptation of the perception of causality. *Curr Biol* 23(3):250–254.
45. Burnham KP, Anderson DR (2004) Multimodel inference understanding AIC and BIC in model selection. *Sociol Methods Res* 33(2):261–304.