Title

Ongoing Dynamic Calibration Produces Unstable Number Estimates

Journal

Journal of Experimental Psychology General, 151(9)

ISSN

0096-3445

Authors

Brockbank, Erik
Barner, David
Vul, Edward

Publication Date

2022-09-01

DOI

10.1037/xge0001178

Ongoing dynamic calibration produces unstable number estimates

Erik Brockbank[1], David Barner[1,2], and Edward Vul[1]

[1] University of California, San Diego, Department of Psychology, 9500 Gilman Drive, La Jolla, CA 92093

[2] University of California, San Diego, Department of Linguistics, 9500 Gilman Drive, La Jolla, CA 92093

Author Note

Experimental data reported in this manuscript were initially presented at the 2013 annual meeting of the Cognitive Science Society.

Correspondence concerning this article should be addressed to Erik Brockbank, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093. E-mail: ebrockbank@ucsd.edu. Data and materials available at: `https://github.com/erik-brockbank/estimation_drift`.

Abstract

Whether estimating the size of a crowd or rating a restaurant on a five-star scale, humans frequently navigate between subjective sensory experiences and shared formal systems. Here we ask how people manage this in the case of estimating number. We present participants with arrays of dots and ask them to report how many dots there are. Our results produce two novel findings. First, people's estimates are best fit by a *bilinear* function in log space, rather than the traditional power law described in previous literature. Second, we find that people's estimates do not have a stable *coefficient of variation* at higher magnitudes, and that the likely cause of this is a "drift" in people's estimate calibration over many trials which has not previously been identified. Building on these results, we present a model of the mapping function from subjective numerosity to symbolic number which relies primarily on a constrained set of previous estimates and familiar numerosities, rather than the robust internal scale used in existing models. Our model is able to generate an accurate mapping with limited data and reproduce notable aspects of estimation seen in our experimental results. This suggests that human number estimation, and perhaps other domains in which we must navigate between subjective representations and formal systems, is governed by a relatively simple decision process that primarily seeks to maintain consistency with previous estimates.

Keywords: numerosity; number; estimation; sampling; Bayesian modeling

Ongoing dynamic calibration produces unstable number estimates

Human reasoning and planning frequently involves mapping between internal states and formal systems: We can compare the weights of two rocks using just our subjective sense of weight, but to provide an estimate of one rock's weight in kilograms requires translating that subjective sense onto a formal metric scale. This task of expressing perceptual states in objective, standard systems is commonplace, from making time estimates to evaluating prices, yet it requires the unique ability to describe our internal representations of the world using abstract systems like number and value. What information do we use to accomplish this, and are such mappings stable? More broadly, *how do humans map from subjective internal states to formal systems?* In this paper, we approach this question using people's ability to estimate number.

Based on a quick glance at a display of many objects, humans can estimate the number of objects present using basic visual cues even when there is insufficient time to explicitly count them. Imagine, for example, the task of guessing how many people are in a large room. As you look around, you can get a rough sense of the number of people present based on the density of the crowd and the size of the room, and can do so faster than you could count each person individually. Past work suggests that it takes around 300ms per item to count individual items while estimation happens much faster (Simon & Vaishnavi, 1996). To estimate the size of a crowd requires that we convert the visual signals we receive from the world to an internal representation of magnitude, which can then be translated to a rough numerical estimate. How do humans accomplish this mapping from visual information to an estimated quantity based on limited signals from the environment?

A large body of research has examined the representations that support our internal sense of number and that form the basis of numerical reasoning tasks like estimation (for a recent review see Leibovich et al. (2017)). The predominant view in this literature is that people have an internal "Approximate" or "Analog" Number System (ANS) which allows for rough discrimination of numerical quantities across sensory

modalities (Brannon, 2006; Brannon & Terrace, 1998; Dehaene, 1997; Gallistel, 1990; Starkey et al., 1990). This system has been documented in a variety of animal species (Feigenson et al., 2004) and emerges in humans early in infancy (Xu & Spelke, 2000), though the role that it ultimately plays in the development of numerical reasoning remains controversial (Carey & Barner, 2019; Nieder, 2020). A distinct but related view is that, rather than being inferred from perceptual stimuli, number is available as a primary feature of perception. Compatible with this, numerical estimates are subject to visual adaptation effects, much like other visual properties such as color and motion (Burr & Ross, 2008). However, competing accounts emphasize that perceptual features of a quantity, such as size, area, and density, are highly correlated with number and that people struggle to infer number independent of these cues (Leibovich et al., 2017). This has led some to argue that our ability to estimate number stems from a more abstract "Generalized Magnitude System" without any internal number representation, or that insofar as we have an internal representation of number, it is assembled from our underlying sense of continuous magnitudes (Gebuis & Reynvoet, 2012; Leibovich et al., 2017; Lourenco & Longo, 2011; Walsh, 2003). Such a system, and the related question of whether humans and other animals selectively represent number via a system like the ANS, remains an area of active research (Clarke & Beck, 2021; Opstal & Verguts, 2013; Yates et al., 2012).

The present work is agnostic regarding the format of our internal representations of number. Whether humans have an internal sense that is *number specific* (e.g., the ANS) or assemble their sense of number from continuous magnitudes that simply *correlate* with number, the task of estimation requires mapping such inputs to formal representations such as number words and written numerals. In this way, it belongs to a broader class of problem, namely that of navigating between subjective, internal representations and quantitative external systems (Stevens, 1956). Influential work in psychophysics has shown that the mapping process from representations to external systems can be formally separated from the earlier mapping from stimulus to representation (N. H. Anderson, 1974;

Attneave, 1962; Birnbaum, 1974; Shepard, 1981; Treisman, 1964); for a review, see Gescheider (1988). In this vein, we consider it separately from questions about the underlying representational character of number information. Further, results using a range of psychophysical tasks such as estimating tone loudness and line length suggest that the mapping functions from various internal representations to external scales or categories are often agnostic across sensory domains (Collins & Gescheider, 1989; Zwislocki, 1983); we therefore expect results in this literature to inform questions about number estimation.

Critically, prior psychophysical investigations into how people map from internal representations to external systems have largely emphasized the limitations of this process (Miller, 1956; Shiffrin & Nosofsky, 1994). For example, in *absolute judgment* tasks, people are shown stimuli that vary along a single psychological dimension (e.g., tone loudness) and asked to provide the correct ordinal labels for the stimuli (e.g., 1-11) over many repeated presentations. In these tasks, people are typically only accurate for a handful of categories (exhibiting constrained *information transmission* from stimulus to response: Baird et al. (1970) and McGill (1954)), even when they can easily differentiate members of the category. Further, their responses show characteristic dependencies on the previous stimulus (Garner, 1953; Holland & Lockhead, 1968) so that in some cases, a participant's response in a given trial is well predicted by the stimulus and response in the previous trial (Mori, 1989). Researchers have offered a number of accounts for these effects, but the dependency on previous trials is difficult to explain in models that specify a fixed internal scale (Stewart et al., 2005). In contrast, Laming (1984) demonstrated that people's behavior in absolute judgment tasks can largely be explained by a model that has a highly limited internal scale and instead relies exclusively on a coarse relationship between current and previous stimuli to calibrate responses; despite the constraints on this model, subsequent work has shown support for such "comparison-based" or "relative" accounts of the mapping from internal scales to external judgments (Stewart & Brown, 2004; Stewart et al., 2002, 2005).

In contrast to categories of auditory tones or line lengths, people's estimates of

number do not show the same limitations in the mapping from internal representation to external values. Extensive prior experience with number categories allows people to map internal representations to a large (theoretically infinite) range of correct responses. In line with this difference, existing models of estimation have largely assumed a robust internal scale which forms the basis of the mapping from subjective representations to external values (Izard & Dehaene, 2008). This assumption builds on earlier psychophysical results as well; for example, Laming (1984) notes that the "relative" model of absolute judgment tasks above may not extend to domains such as color where people amass rich prior knowledge. Further, in *magnitude estimation* tasks, in which participants identify line lengths or tone loudness as described above, but are allowed to assign arbitrary numbers to each stimulus rather than identify its ordinal category (e.g., "tell me a number that seems as big as the line seems long" (Collins & Gescheider, 1989)), responses often reflect an accurate ordinal ranking of the stimuli (Collins & Gescheider, 1989; Zwislocki, 1983). These findings have been interpreted as suggesting that number is mapped to a stable underlying scale that is recruited for other psychometric judgments (Baird et al., 1970; Collins & Gescheider, 1989).

In line with the idea that number estimation relies on a relatively stable internal scale, research on human numerical estimation has demonstrated several robust features of people's mapping from subjective magnitude representations to symbolic number. First, when estimating quantities outside the *subitizing* range, people tend to underestimate (Kaufman et al., 1949). This relationship follows a power law (Indow & Ida, 1977), where the magnitude of people's errors is roughly proportional to the quantity being estimated. This error pattern is thought to produce a stable *coefficient of variation* (CoV) in estimates (Gallistel, 1990; Shepard et al., 1975) (though see also Testolin and McClelland (2021)). Second, the accuracy of people's estimates (i.e., the amount they underestimate, or in some cases overestimate) varies considerably across individuals (Krueger, 1982). Together, these findings suggest that people's mapping from internal quantity

representations to formal numbers is often systematically miscalibrated within individuals and exhibits reliable variability from person to person.

Perhaps the most comprehensive attempt to characterize the "interface between the system of verbal numerals and the non-verbal analog representations of numerosity" is presented by Izard and Dehaene (2008). They found that giving participants a reference array and telling them it had a magnitude which was either equal to, above, or below the true number calibrated all subsequent estimates, suggesting that the mapping from numerical representations to formal estimates is flexible and "globally" responsive to new information. Building on these results, they proposed a model of the mapping whereby people deploy a "response grid" overlaid on the mental number line, which is itself a Gaussian distribution of activation around the perceived magnitude. On this model, activations for a given stimulus produce corresponding activation of segments of the response grid, which is then used to provide a verbal estimate. Individual differences in estimation and calibration of participants via reference arrays amount to a "stretching or compression" of the response grid (Izard & Dehaene, 2008). This internal scale has an intuitive notion of numerical distance built in, allowing for rich numerical inference based on a given stimulus. Indeed, the model described in Izard and Dehaene (2008) predicts the robust empirical features of number estimation described previously and replicated in their own results: a) that participants have power law underestimation behavior over the range of estimates, b) that the degree of underestimation can be fit to individuals based on idiosyncratic stretching or compression of their response grids, and c) that estimates display *scalar variability* over increasing magnitudes, i.e., the degree of variability in estimates increases in proportion to the magnitude of the stimulus being estimated (we use scalar variability and constant CoV interchangeably in what follows).

Subsequent work aimed at uncovering the mechanisms of the "interface" described by Izard and Dehaene (2008) has largely focused on the role that *associative learning* and *structural analogy* each play in supporting the mapping from magnitude to number. Under

an "associative mapping" account, the process of mapping approximate magnitude representations to verbal number is one of learning to align a given number word to its corresponding magnitude representation (Lipton & Spelke, 2005; Nieder, 2020). However, proponents of a "structure mapping" view have noted that structural similarity between monotonic internal magnitude representations and the ordinal structure of the number line might allow for a mapping of magnitude to symbolic number which instead draws on notions of equivalent ordering and distance across the two systems without needing to map every number to a corresponding magnitude (Carey, 2009). Evidence from studies of estimation in young children supports the use of an associative mapping for small numbers (LeCorre & Carey, 2007), while the re-calibration results in Izard and Dehaene (2008) and similar work by others (Lyons et al., 2012) call into question the notion of a strong association between number and quantity for larger magnitudes. Indeed Sullivan and Barner (2013) suggest that humans use both structure and associative mappings to support estimation, though critically, they find little evidence that children or adults have associative mappings beyond magnitudes of about 12; further, they find that developmental improvements in estimation are better explained by improvements in structural analogy than improvements in accuracy or scope of associative mappings.

The results from Sullivan and Barner (2013) suggest that humans associate numbers with quantities in the world through some combination of associative and structure mappings (for a recent review, see Carey and Barner (2019)), and therefore that models like Izard and Dehaene (2008), which posit only a single mapping function, may not be adequate. Nevertheless, the "response grid" model in Izard and Dehaene (2008) provides one of the only formal accounts of how large number estimates are affected by calibration. It therefore offers the best existing model of how structure mapping—and thus most estimates—might work. The model also makes several concrete predictions about the form and stability of people's estimate functions that have not been robustly tested in previous studies, and which are important to understanding the mechanisms that support

estimation. In particular, it predicts that estimates should obey a power law for large magnitudes, and that estimates should exhibit scalar variability or a constant CoV. As we describe below, data regarding these predictions suggest an alternative model of estimation that not only explains calibration effects, compatible with structure mapping, but can also accommodate effects attributable to associative mapping.

In the present study, we tested the power law estimate function and stable CoV predicted by Izard and Dehaene (2008) to better characterize the relationship between associative and structure mappings over a large magnitude range. Our results contain two key findings. First, we show that individual estimates are better fit by a log-bilinear function rather than a power law, which may in turn capture the relationship between precise associative mappings of lower magnitudes and more flexible structure mappings for larger numbers. Second, we find that people display *dynamic* variability in their mapping from magnitude representations to verbal estimates over many trials. This produces an increasing coefficient of variation at higher magnitudes. We hypothesize that this latter *mapping variability* stems from an ongoing attempt to maintain calibration consistency with previous estimates. We argue that these findings are not easily incorporated into the "response grid" model of Izard and Dehaene (2008) and present an alternative model that produces numerical estimates based not on a stable internal calibration but on samples from prior trials and familiar magnitudes. This model is consistent with a large body of psychophysics work indicating that people's judgments of magnitude and categorizations of continuous stimuli are heavily influenced by the context of previous trials and show little evidence of having a stable internal scale (Laming, 1984; Stewart et al., 2006; Stewart et al., 2002, 2005). Despite the challenge of not having a reliable underlying scale, the model is able to reproduce key characteristics of human number estimation consistent with our experimental results. Altogether, our findings suggest that human number estimation, and other domains in which we must navigate between subjective representations and formal systems, is governed by a relatively simple decision process that seeks to maintain

consistency with previous judgments and prior experience with the relevant system. In this way, our results suggest that the process of mapping from internal representations to external systems may rely on computationally limited, domain-general processes even in settings where we have a great deal of calibrated experience.

Experiment

In this experiment, participants estimated number in dot arrays over many repeated trials that captured a large magnitude range; we investigate the form and stability of people's estimate calibrations across trials.

Participants

Participants were 24 undergraduate psychology students at the University of California, San Diego who received course credit for their participation. Informed consent was obtained from all participants in accordance with the Institutional Review Board's approved protocol.

Procedure

In each trial, participants were shown a series of dot arrays on a white background like those in Figure 1. The array of dots was presented for 250ms, and then subjects were prompted to type in their guess as to how many dots were in the array. Subjects were then asked to type in a second guess about the number of dots in the array. Our experimental results use the first of the two guesses. Participants performed 300 estimation trials over the course of the experiment and did not receive feedback on their estimates at any point.[1]

[1] The code for this experiment, as well as all data, analyses, and modeling code, are available at: `https://github.com/erik-brockbank/estimation_drift`.

Stimuli

The number of dots shown on each trial was sampled from a geometric distribution with a mean of 50, truncated at the low end so that displays had at least two dots. All the dots in an array were the same size (radius of 10 pixels), presented in red on a white background. The configuration of dots was randomly generated by drawing locations from a uniform distribution over the full display area (1024x768 pixels) with the constraint that the dots did not overlap. The range of stimuli did not control for changes in perceptual features that correlate with number, such as stimulus density, display luminance, or convex hull, since the impact of these non-numerical dimensions on underlying number representations is somewhat orthogonal to the question of how people generate number estimates on the basis of their internal representations. Figure 1 shows an example trial along with one representative subject's data from all 300 trials.

Results

The "response grid" estimation model in Izard and Dehaene (2008) makes four predictions about the overall character of people's number estimates: (1) estimates should follow a power law over increasing numbers; (2) estimate calibration across participants should reveal large individual differences due to idiosyncratic "stretching" of the response grid; (3) individual estimate calibration should be fairly stable over time (so long as participants don't receive feedback on their estimates); and (4) estimates should have a static coefficient of variation. Here we examine each of these predictions in turn.

Bilinear estimate function

Previous work on number estimation has proposed that people's estimates can be described by a power law, where a numerical estimate y based on the presented number x can be approximated by $y = \alpha x^{\beta}$. Individual fits for α and β reflect a participant's overall accuracy: Their tendency to underestimate can be described by a stable $\beta < 1$ (see for

example Izard and Dehaene (2008)). This power law produces a relationship that is linear in log space, i.e., $log(y) = log(\alpha) + \beta log(x)$. Figure 2 shows individual estimate data for each of the 24 experimental participants plotted on log coordinates. It's clear that participant estimates do not appear perfectly linear in log-log coordinates. We propose that the mapping function is not described by a linear relationship between log magnitude and log estimates, but bends such that small numbers are mapped more or less veridically onto number words, while large numbers show a systematic deviation from the identity line. Consider a bilinear function that is accurate up to some critical number C, and then deviates from the identity line with some log slope of S. This produces an estimate function of the following form:

$$\log(y) = \begin{cases} \log(x) & \text{for } x \leq C \\ S(\log(x) - \log(C)) + \log(C) & \text{for } x > C \end{cases} \tag{1}$$

The bilinear estimate function defined above produces parameter estimates that match the coarse patterns observed for individual estimates in Figure 2. Cutoff values (fit in log space across participants) averaged 1.175 ($sd = 0.34$), or around 15 in linear coordinates. Notably, this cutoff is well above the subitizing range explored in prior literature and typically described as about 5 (Kaufman et al., 1949; Mandler & Shebo, 1982), suggesting that the bilinear model characterizes estimation patterns beyond simply differentiating subitizing from power-law like estimates. Fitted linear slope estimates (above the threshold) averaged -0.25 ($sd = 0.11$), or around 0.56 in linear coordinates, reflecting the general pattern of under-estimation shown robustly in previous literature (Izard & Dehaene, 2008; Kaufman et al., 1949; Krueger, 1982).

Critically, the bilinear estimate function can account for data of individual subjects better than a simple line in log space with an intercept α and a slope β. Figure 2 shows best-fitting linear and bilinear curves for each participant on log-log coordinates. Averaged across participant fits, the bilinear model has a substantially lower Bayesian Information

Criterion (BIC) than the fitted power law function (bilinear: mean BIC = -402.03; power law: mean BIC = -346.24). In addition, across individuals, the BIC for the bilinear function was lower than the BIC for the power law function for 19 out of the 24 participants (in each case using $k = 2$ parameters and $n = 300$ estimates per person).

To ensure that the bilinear fit is not a result of factors specific to this experiment, we ran a similar model comparison using the data from Experiment 1 of Izard and Dehaene (2008). In that experiment, five participants completed five sessions of 600 trials each for a total of 3,000 estimates with stimuli in the range $[1, 100]$. Though this represents a lower range of number estimates than in our experiment, the distribution of individual responses is similar to the bilinear pattern exhibited in our data. Indeed, a model comparison with the estimate data reported in Izard and Dehaene (2008) shows that the bilinear function has a lower BIC than the power law function for all five of the participants (bilinear: mean BIC = -5217.1; power law: mean BIC = -4694.9; individual differences in BIC: subject 'ML' = 267.5; 'PQ' = 798.4; 'AL' = 83.7; 'DC' = 1102.5; 'BF' = 359.1).

Stable individual differences

Previous work which has assumed that estimate functions follow a power law has measured individual variability in accuracy using the fitted exponents of the power law mapping described previously (Krueger, 1982). However, given the finding that individual calibration can be described by a log bilinear function, this predicts that individual variability in estimate accuracy can instead be measured in the *bilinear* slopes fit to each individual's estimates at higher magnitudes. Large individual differences in bilinear estimate functions implies that there is a great deal of variability *between* subjects in their bilinear slopes and little variability *within* subjects over time. We evaluate this by separately estimating slopes for each participant from two distinct sets of trials, then calculating the pairwise correlation in these slope estimates across the two sets. In other words, for two sets of slopes A and B, where A_i is participant i's slope estimate from one

set of trials and B_i is participant i's slope estimate from the other set of trials, we calculate the pairwise correlation of A_1 and B_1, A_2 and B_2, etc. If subjects show a great deal of variability in their individual slopes and little change over time (i.e., large individual differences), then this correlation will be high, since A and B will each have high variance and high pairwise similarity. In contrast, if subjects are inconsistent in their own calibrations (leading to low pairwise similarity) or there is little variability between subjects (i.e., little variance in the slopes in A and B), this correlation will be low.

In this vein, we assess the individual variability in shape of the bilinear mapping via a *modular* split-half analysis. We divide the 300 trials into odd trials (1, 3, 5, ..., 299), and even trials (2, 4, 6, ..., 300) and determine the pairwise correlation between each participant's slope S (see Equation 1 above) estimated for each of the two halves. For any given participant, the slopes for these two split halves should be highly similar; whatever is true of their estimate calibration across even trials should be equally true in odd trials. Therefore, the correlation between these halves depends mostly on the variability in each set of slopes. If participants have a large amount of individual variability in their estimates, then the split-half correlation between participant slopes from the two sets of trials should be high; on the other hand, a low split-half correlation would reflect little stable variability between participants. Modular split-half correlations for the bilinear slopes were very high ($r = 0.96$; $t(22) = 15.6$, $p < 0.001$), revealing large, stable individual differences in estimate calibrations. Thus, individual differences persist under the bilinear model, not just the traditional power law estimation model. While prior work has mostly argued for the source of this individual variability in the acuity of people's magnitude *representations* (Gallistel & Gelman, 1992; Whalen et al., 1999), we show in the subsequent section that this may also stem from variability in people's *mapping* from magnitude representations to verbal estimates over time.

Within-subject calibration "drift"

The "response grid" model of estimation proposed by Izard and Dehaene (2008) predicts that human estimate calibration should be stable over time in the absence of feedback. They find that individual estimates are calibrated by feedback (both accurate and misleading) and that this calibration persists across many subsequent trials. To test the stability of the mapping function over time, we revise our earlier modular split-half analysis in favor of a *blocked* split-half analysis. In a blocked split-half analysis, we divide each participant's estimates into their first and second half of the experiment rather than even and odd trials. As with our modular split-half analysis, the blocked split-half correlation of subject estimate slopes was highly significant ($r = 0.79$; $t(22) = 6.13$, $p < 0.001$), indicating that people are very consistent in their idiosyncratic magnitude to number mappings. However, the blocked split-half correlation above is notably lower than the *modular* split-half correlation discussed previously—this difference is highly significant using the Fisher r-to-z transform ($z = -2.58$, $p = 0.0098$). The difference between modular and blocked split-half correlations provides a coarse indication that the slope of the magnitude-number mapping function is not stable within individuals over the experimental session; if the bilinear slopes we estimate for each participant's mapping function were stable over time, the best-fitting slopes in the first and second half of the experiment should not depart meaningfully from each other. Their correlations should therefore be similar to those estimated over the full range of the experiment on alternating trials.

To more precisely measure the change in estimate calibration over time indicated by our split-half results, we generalize the blocked split-half analysis to blocked split-$n$ths for $n = 30$. For a split-30th analysis, we divide our 300 trials into 30 subsets (rather than the two used for split-half), each one comprising 10 trials. For example, the fifth blocked split-30th subset will contain trials 41-50. This gives us a more fine-grained view of the change in estimate slopes over time. As in our blocked split-half analysis, we estimate the bilinear slopes for each of these trial subsets and compare them across participants. Figure

3a shows the correlations between calibration slopes across each of the 30 trial blocks. Trial blocks which are close to each other have a higher correlation than blocks which are farther, i.e., block 1 is more like block 2 than block 10. This suggests that the blocked split-half results described above reveal a broader pattern of decreasing reliability of estimate calibration over time: People's estimates *drift* in their calibration.

Though the overall pattern of drift can be seen in Figure 3a, we wish to quantify how much people's calibrations vary over the experiment. The blocked split-30th correlation between, e.g., block 1 and block 2 (the bottom-most red square in Figure 3a) measures the correlation of slopes estimated from two adjacent periods of time in the session which are on average separated by 10 trials. The same is true for the correlation between blocks 2 and 3, 3 and 4, etc. In general, if we calculate the correlation between subset i and subset $i + k$ from a blocked split-30th analysis, those subsets are separated by a *trial distance* of $300 * k/30$ trials. If slopes are drifting over the course of the experiment, we would expect the correlation of slope estimates to decrease with k—the separation between blocked subsets. This predicted decrease in slope correlations over increasing *trial distance* summarizes the pattern seen in Figure 3a: Trials in blocks 1 and 2, 2 and 3, and 3 and 4 ($k = 1$, average trial distance = 10) have a higher slope correlation than trials in blocks 1 and 10, 2 and 12, 3 and 13, etc. ($k = 10$, average trial distance = 100).

Figure 3b shows the correlation in bilinear slope estimates across *trial distances* between the blocks of trials shown in 3a. This correlation over trial distances is plotted in red. As a point of comparison, we calculate this same correlation after shuffling each participant's trial index, shown in green in Figure 3b; this represents an expectation about the stability of individual calibration slopes when we do not consider the time course of the experiment. While the shuffled correlation of estimate slopes remains stable, the blocked split-30th correlations decrease steadily over greater trial distances. The fact that the blocked split correlations remain fairly high even at a trial distance of 300 (hovering around 0.6 in Figure 3b) is likely a result of stable individual variability in estimate calibrations in

combination with people maintaining reasonably calibrated estimates throughout the task. However, despite the overall stability of estimate calibration across individuals, a linear regression on the correlations in the blocked split-30ths as a function of trial distance is significantly negative (95% confidence interval on the slope: (-0.0015, -0.0013); $t(433) =$ -24.95, $p < 0.001$). Meanwhile, the slope of the shuffled trial order correlations shown in green in Figure 3b is not significantly different from zero. This provides a robust confirmation that participant estimate calibrations *drift* over the course of the experiment. It is worth noting here that we describe participants' estimate calibration as a "drift" not in the directional sense of, e.g., drift-diffusion models (Ratcliff & McKoon, 2008), but instead as a random walk constrained by each participant's overall calibration tendency and their most recent estimates. These results are not predicted by the response grid model of Izard and Dehaene (2008); rather than a stable mapping from magnitude to number, we find evidence of a dynamic variability in people's estimate calibrations over time.

Increasing coefficient of variation at higher magnitudes

Prior work in number estimation has proposed that people have an idiosyncratic but stable *Weber fraction* which represents variability in their internal number representations (Gallistel & Gelman, 1992; Whalen et al., 1999). However, the subsequent mapping from these internal representations to verbal estimates is assumed to be noise-free, leading to a constant coefficient of variation in their estimates. In other words, the variability of their estimates scales with the magnitude of the estimates as a result of Weber noise in the underlying number representations (though see recent findings in Testolin and McClelland (2021)). However, in the previous section we describe evidence that the bilinear slope of individual estimates may wobble across many trials, causing participants' estimate calibrations to "drift" over time. This drift will naturally introduce variability in estimates above and beyond that produced by Weber noise, since it causes variability in the mapping from number representations to estimates (rather than just variability in the

number representations themselves). Further, this drift in the logarithmic slope of an individual's bilinear mapping will affect estimates at larger magnitudes more than smaller magnitudes, because the wobbling slope of the bilinear mapping introduces greater variability farther along the number line. As a consequence, the calibration drift described previously predicts that the coefficient of variation for participants' estimates will *increase* over increasing magnitudes. Some evidence of this can be seen in individual estimate data in Figure 2. When viewing estimates in log coordinates, a constant CoV amounts to a consistent variability at all (log) magnitudes (i.e., variance that increases in proportion to magnitude should be constant for multiplicative increases in magnitude). However, it's clear in Figure 2 that subjects appear to have *increasing* variance in their estimates at larger magnitudes, even when viewed in log coordinates. This suggests that CoV may be increasing for these participants, as predicted by the slow drift in estimate calibrations.

To test whether participants have an increasing coefficient of variation, we fit bilinear curves to their estimates as before; now, in addition, we fit a linear parameter to the variance of their log estimates to determine how much this variance is increasing as a function of log magnitude. Concretely, our previous bilinear estimate functions were fit using a normal distribution around the true log magnitude to determine the likelihood of each participant's log estimates. With a constant CoV, the standard deviation of this log normal distribution should be a fixed value which reflects the coefficient of variation. However, if CoV is increasing, then a linear function fit to the standard deviation of the log normal as a function of (log) magnitude should have a positive slope (if CoV is not increasing, this function will have a slope of zero). We fit slope and intercept parameters to the standard deviation of log estimate distributions for each participant.[2] Figure 4 shows the distribution of fitted slopes, which represent the increase in standard deviation as a function of log magnitude. The distribution of slope parameters is significantly greater than zero ($t(23) = 5.54$, $p < 0.001$; 95% confidence interval for the slope mean is [0.17, 0.38]); as log magnitude increases, the best-fitting standard deviation for the distribution

of log estimates increases as well, suggesting that participants have an *increasing* CoV. This pattern is predicted by an estimate process which involves ongoing updating of individual estimate calibrations—which would have a larger impact on the variability of larger estimates—but is not consistent with participants having a stable coefficient of variation (Izard & Dehaene, 2008).

Discussion

In this experiment, we sought to characterize the form and stability of the mapping function between representations of number and formal estimates. In particular, we tested the claim of Izard and Dehaene (2008) that this mapping function should respect a power law that varies across individuals, with a constant coefficient of variation that reflects a stable mapping from noisy internal magnitude representations. We presented participants with a large range of magnitudes and analyzed their estimates over the course of many trials. Results from this experiment provide two novel ways of thinking about how people map from perceptions of number to verbal number estimates.

First, the overall shape of people's estimates is best described by a *bilinear* mapping in logarithmic space from presented number to estimate. In this formulation, most people are highly accurate at estimates up to a threshold, after which their estimates depart from the identity line, most often underestimating. Critically, this shift in behavior for larger numbers does not simply reflect random responding, or a complete lack of systematicity. Instead, our results show that although participants are not always *accurate* at high magnitudes, they are nevertheless uniquely *calibrated*. Even over a distance of 300 trials, estimates exhibited far greater variability between participants than within. This log bilinear fit differs from the power law described in previous literature (Izard & Dehaene, 2008), but is notably consistent with results suggesting that people combine *associative*

[2] This new fit did not substantially change the bilinear parameters previously fit to participant estimates, though the addition of the slope parameter increased overall log likelihood of the fits.

mappings at lower magnitudes with more flexible *structure mappings* at higher magnitudes (Sullivan & Barner, 2013). The best fitting cutoff values for each participant, which averaged 14.9 in linear coordinates, may in part reflect the point at which they no longer rely on associative mappings and therefore generate less accurate estimates.

One alternative account of these results is that the fitted cutoff values do not reflect a transition from associative to structural mapping, but subitizing instead (Kaufman et al., 1949; Mandler & Shebo, 1982). On this view, the bilinear model simply reflects the role of subitizing at lower numbers. However, the average cutoff value of approximately 15 in our data is substantially higher than the typical subitizing limit of around 5. The accuracy of the bilinear function below the cutoff is therefore unlikely to *merely* reflect the increased accuracy of subitizing.[3] A second possible account of the data is that the cutoff values, rather than representing something like the transition from associative to structure mappings, are primarily a function of exposure time to the stimuli. Prior work has shown that estimation error can in part be explained by the amount of time participants have to foveate a number array (Cheyette & Piantadosi, 2019). However, the finding in these results is that participants who are given longer exposure (up to 3s) on some stimuli show less underestimation and lower Weber fractions in their responses. Such an effect ought to improve estimate calibration and reduce variance across the full range of numbers, rather than being restricted to some particular cutoff point, as we find.

The second contribution of the current results is to show that numerical estimation—in particular, the slope of the bilinear fit for larger magnitude estimates—is

---

[3] A further distinction between the regime below our cutoff of around 15 and subitizing is that subitizing is typically considered to be a matter of *precision* rather than calibration for low number estimates. Our data show subitizing—or zero-variance estimation—below about 6, but in the range of 6 to 15, while there is variance in estimates (in contrast to subitizing), there is no systematic miscalibration. Thus we believe that the calibrated regime below 15 is a different phenomenon than subitizing itself. Indeed, Kaufman et al. (1949) showed calibration up to about 15, but only identified the subitizing range as below 6, because that is the range to which zero-variance, high-confidence estimates were restricted.

subject to a slow "drift" over many trials (despite the relative stability noted previously). While the mapping of magnitudes to numbers may be consistent across a range of magnitudes at any point in time, these data suggest that it changes over time. This change in calibration slope over many trials, which is exaggerated at higher magnitudes, explains the variability in estimates above and beyond what would be expected by Weber noise in internal number representations. This in turn explains the observed increase of the coefficient of variation for larger magnitudes, an effect not accounted for by existing models. Why might previous results have failed to detect this increase in coefficient of variation? Given that an individual's slow drift in estimate calibration is best detected across many trials and a large range of estimates, previous studies may have lacked a sufficient number of trials at large magnitudes to detect such effects (see e.g., Frank et al. (2008), Frank et al. (2012), and Gallistel (1990), Gordon (2004)). More recent work by Testolin and McClelland (2021) has also called into question the notion of a stable CoV. After re-analyzing data from several well-known experiments on number perception and estimation, the authors find evidence of a *decreasing* CoV for estimates between 10 and 80 first reported in Revkin et al. (2008). This pattern contrasts with our results showing an increasing CoV for larger estimates, but this may also be due to the lower range of numbers estimated in Revkin et al. (2008). Future research should incorporate larger estimate ranges and total estimates to better quantify this effect.

Broadly, our results suggests that people not only have uncertainty in their subjective representations of number, but also a dynamic uncertainty in their mapping from these representations to formal number. One possible explanation of this dynamic uncertainty is that when estimating, people seek to maintain consistency with previous estimates, which could produce the sort of wandering calibration slopes seen in our data. Such an account might in theory apply to a range of settings where we regularly map from internal, subjective representations to formal systems. Indeed, this finding integrates number estimation with other domains of psychophysics where such effects have been

observed (Garner, 1953; Holland & Lockhead, 1968; Stewart et al., 2002). In the next section, we test this theory with a model of the mapping process from magnitude to formal number in which numerical estimates are generated with a goal of maintaining consistency with prior estimates. This approach builds on previous models in psychophysics which aimed to provide accounts of similar dependencies exhibited across multiple judgments (Laming, 1984; Stewart et al., 2005). We show that this model, with a limited set of cognitively plausible assumptions about people's numerical reasoning process, achieves an accurate mapping and produces characteristic patterns of bilinear estimation, individual variability, and calibration drift.

Modeling Number Estimation

The experimental results described previously suggest that central features of human number estimation are unaccounted for by existing models, namely the log-bilinear shape of the estimate function and the "slow drift" in estimate calibration. The latter reflects dynamic variability in the mapping from internal magnitude representations to verbal estimates which produces an *increasing* coefficient of variation at higher magnitudes. The "response grid" model of Izard and Dehaene (2008), which proposes a direct mapping between internal magnitude activation and verbal number estimates, does not readily incorporate these findings. While revisions to the response grid might allow for auto-correlated stretching and compression of the grid over successive trials to produce a drift similar to what participants exhibit, there is no principled way to produce the log-bilinear estimate function seen in our empirical results; nothing about the response grid formulation suggests that participants should be highly accurate up to a threshold of around 15 and then show power-law like estimates for greater magnitudes.

The response grid model also requires that people's internal magnitude representations offer a *distance* metric which can be mapped onto numerical distance for estimation, i.e., different magnitude representations have a psychophysical "distance"

which has a rough correspondence to differences in number. While this is likely defensible in the case of number estimation (Sullivan & Barner, 2014a), we cannot assume that any scale for which we have an internal representation will have the property of intuitive distance, nor that such distances will map cleanly onto the formal system (Laming, 1984): Consider, e.g., willingness-to-pay or how much you enjoyed a restaurant mapped onto a five-star review scale. Therefore, in the interest of generality, we seek a solution to number estimation which might plausibly inform the broader problem of navigating between psychophysical and formal scales.

To better account for our experimental data and to provide a more generalized solution to the problem of mapping internal states onto formal systems, we propose a model of number estimation that does not rely on a stable internal scale that corresponds to the external one. Instead, our model uses paired magnitude-number associations drawn from past experience to determine the most likely estimate on a given trial. At a high level, the model generates an estimate for a given magnitude by sampling previous trials, as well as more familiar magnitude-number mappings drawn from prior experience. A series of ordinal comparisons between the trial magnitude and the sampled magnitudes gives the model a set of parallel ordinal constraints on the corresponding number value for the current stimulus (i.e., if its magnitude is larger than the previous trial's magnitude, then its number estimate should also be larger). These ordinal comparisons, combined with a prior that reflects more experience with low numbers than high numbers, forms the basis for the model's estimate function.

Critically, estimates generated in this fashion don't rely on a stable or long-term internal scale, but instead are the result of ongoing calibration using items sampled from memory. This approach is broadly consistent with earlier work which has shown that rich commonsense inferences can be made using only simple operations performed over limited samples (Bonawitz et al., 2014; Stewart et al., 2006; Vul et al., 2014). Further, the model's estimate process does not require a notion of psychophysical distance which somehow maps

onto numerical distance, since its estimates are based on simple ordinal comparisons with previous experience; its accuracy then depends on the availability of relevant samples which allow it to calculate an estimate. The use of immediate context to support estimates rather than an *absolute* internal scale connects this model to prior work in psychophysics (Laming, 1984; Stewart et al., 2002, 2005), reflecting its generalizability beyond number estimation. In particular, it is similar to models of categorization which rely on the difference between the current and previous stimulus to make a decision (Stewart & Brown, 2004; Stewart et al., 2002, 2005). However, the current model differs from this prior work in a number of ways, including the combination of previous trials with familiar examples and the mapping onto a large integer range rather than a small number of ordinal categories.

Despite the considerable challenges in producing estimates with such a simple process, we show that this model is able to achieve a reasonably accurate mapping based on a limited set of data. Critically, we further show that the model's structure allows for a simple characterization of bilinear estimate patterns, individual variability, and calibration drift. Our model therefore suggests that many signature aspects of human number estimation may be explained by such an ongoing estimate calibration process. More broadly, we argue that human number estimation, as well as other settings in which we must map from subjective internal scales to formal external ones, are best described by this sort of simple reasoning process over limited samples.

Model Description

Representing magnitudes

Figure 5 provides an overview of the model's estimation process. First, presented with a number η (e.g., the sample stimulus shown in Figure 1), the model generates an internal magnitude representation m from a distribution $p(m) \propto \mathcal{N}(log(\eta), \sigma)$ for that trial stimulus.[4] The absolute value of this magnitude representation is assumed to have no bearing on the numerical estimate that the model will produce but allows us to formalize

the ordinal comparison between magnitudes. While the exact nature of this representation is a subject of active research (Carey & Barner, 2019; Cheyette & Piantadosi, 2020; Leibovich et al., 2017) we remain agnostic about the details of how magnitudes are represented, and our model is neutral with respect to differences between previous accounts.

A natural question is how fine grained this magnitude representation is. If the representation were such that it could distinguish any two numbers with 100% accuracy, the model would have a much easier task than if magnitudes of 50 were indistinguishable from 500. Prior work has suggested that the Weber fraction for people's number representations (the ability to distinguish between two distinct numbers relative to their magnitudes) remains constant as magnitude increases, though it can vary substantially across individuals (Whalen et al., 1999). This leads to a stable coefficient of variation in magnitude *representations* (distinct from the coefficient of variation in *estimates* described previously). Consistent with Whalen et al. (1999), and other similar findings (e.g., members of the Pirahã tribe have a fairly stable coefficient of variation in their magnitude representations even without words for larger numbers (Frank et al., 2008; Frank et al., 2012; Gordon, 2004)), we set the magnitude representations in our model to have a cognitively plausible, stable CoV of 0.24.[5] This means that our model uses a noisy magnitude representation consistent with previous research on human number reasoning.

Calculating the estimate

The model's task is then to select a reasonable number estimate y for this magnitude m (here, we use the range $[1, 1000]$ to match our experimental results). To calculate this, it relies on a set of sampled magnitude and number tuples $\{\mu_i, \gamma_i\} \in \mu, \gamma$.

[4] The assumption throughout this paper that magnitude distributions have variance which is constant in log space comes from prior work suggesting that people's internal magnitude representations are likely to be on a log scale (Izard & Dehaene, 2008). Increasing or decreasing this variance corresponds to more or less "noise" in peoples' approximate number sense (Piantadosi, 2016).

[5] This corresponds to a standard deviation of the $log10$ Gaussian magnitude representation equal to 0.1.

The vectors of magnitude μ and number γ are comprised of previous trial estimates as well as well-known mappings, and thus combine two different sources of information that might support number estimation. Though the *sources* of these mappings vary, the model has a unified process for generating an estimate given this information. The model uses Bayesian inference to select an estimate y sampled from the posterior distribution $p(y \mid m, \mu, \gamma)$. Following Bayes rule, $p(y \mid m, \mu, \gamma) \propto p(m \mid y, \mu, \gamma)p(y)$. Below we outline the process for calculating each of these terms.

The prior $p(y)$ is described by a power law distribution $p(y) \propto y^{-\alpha}$. We set $\alpha = 1$, thus favoring lower numbers overall. This reflects the fact that people have a great deal of experience with small numbers and relatively little experience with large numbers and closely describes the *need probability* function (J. R. Anderson & Schooler, 1991) for sets of increasing magnitude based on their frequency of occurrence in the natural world (Cheyette & Piantadosi, 2020; Dehaene & Mehler, 1992; Piantadosi, 2016). People may therefore be more likely to sample lower numbers as candidate estimates, all else equal.

For the likelihood $p(m \mid y, \mu, \gamma)$, the model assembles a stepwise likelihood distribution defined over candidate y values based on the likelihoods of sample estimates. To illustrate, $p(m \mid y, \mu, \gamma) = \prod_{j=1}^{n} p(m \mid y, \mu_j, \gamma_j)$ for each sample estimate with magnitude μ_j and number value γ_j. The likelihood for each sample $p(m \mid y, \mu_j, \gamma_j)$ has the form:

$$p(m \mid y, \mu_j, \gamma_j) = \begin{cases} p(m < \mu_j) & \text{for all } y < \gamma_j \\ p(m \geq \mu_j) & \text{for all } y \geq \gamma_j \end{cases} \tag{2}$$

Since magnitudes m and μ_j are drawn from Gaussian distributions centered at y and γ_j respectively, we can derive the probability of sampling a value less than 0 from a normal distribution centered at $m - \mu_j$ as $\Phi(\frac{m-\mu_j}{\sqrt{2\sigma^2}})$. While this might seem to violate the earlier constraint that the model only knows the ordinal ranking of magnitudes and not anything about their distance, suitably small σ in log space will make this probability $p(m < \mu_j)$ close to 1 or 0 for almost any two estimate magnitudes, rendering this effectively

a binary ordinal judgment. By taking the product of sample likelihoods $p(m \mid y, \mu_j, \gamma_j)$ in this fashion, the model can assemble a reasonable stepwise approximation to the overall likelihood $p(m \mid y, \mu, \gamma)$ (see Figure 5, step 2). This stepwise distribution is then scaled towards lower numbers by the prior $p(y)$ to produce the posterior $p(y \mid m, \mu, \gamma)$ defined above (Figure 5, step 3). Finally, to generate a number estimate y from the posterior $p(y \mid m, \mu, \gamma)$, the model raises the posterior distribution to an exponent δ whereby sampling from the posterior approximates the maximum *a posteriori* (MAP) estimate (Sanborn & Beierholm, 2016).

Sampling previous estimates

The model calculates a likelihood over y above by sampling from previous estimates and familiar magnitude-number mappings to get sample magnitude and number tuples $\{\mu_j, \gamma_j\} \in \mu, \gamma$. The model has free parameters for the number of samples n that it takes in each trial and k for the probability that a given sample comes from a well-known magnitude to number pairing (with probability $1 - k$ the model instead samples from the set of previous trial estimates). A sample $\{\mu_j, \gamma_j\}$ can then be defined as follows:

$$\{\mu_j, \gamma_j\} = \begin{cases} \text{Memory}(\alpha) & \text{with probability } k \\ \text{Trials}(\alpha) & \text{with probability } 1 - k \end{cases} \tag{3}$$

In the function above, Memory(α) is a function that returns magnitude, number tuples $\{\mu_j, \gamma_j\}$ from *familiar mappings*, and Trials(α) is a function that returns magnitude, number tuples $\{\mu_j, \gamma_j\}$ from *previous trial estimates*.

The familiar mappings between magnitude and number are described by a power law distribution over numbers in the range [1, 1000] (this was chosen to match our experiment but is not central to the model). In other words, Memory(α) returns a sampled number estimate y and a corresponding magnitude m that is close to the true value of the sampled estimate y: $p(y) \propto y^{-\alpha}$ and $p(m) \propto \mathcal{N}(log(y), \sigma)$. These m and y values returned

by Memory(α) form the sample tuple $\{\mu_j, \gamma_j\}$. The power law distribution used for these samples has a slope α that strongly favors sampling lower numbers ($\alpha = 4$).

Intuitively, the use of familiar mappings as a source of information during estimation is consistent with the idea that people have a pretty good sense of what 10 or 20 items looks like. The large slope of the power law distribution from which familiar mappings are sampled reflects the fact that people are far more likely to have an associative mapping to smaller numbers (Sullivan & Barner, 2013), due perhaps to their greater frequency (Piantadosi, 2016), the use of subitizing for especially low numbers (Carey & Barner, 2019; Feigenson et al., 2004), or simply their greater information content (Cheyette & Piantadosi, 2020).

The previous trial estimates, like the well-known mappings, are sampled from a power law distribution, in this case over the n previous trial indices. Trials(α) samples a "lag" value L where $p(L) \propto L^{-\alpha}$, which dictates the previous estimate index from which to sample an estimate (therefore L will be defined over the range [1, n]). The sampled lag yields a tuple of an estimated number $y = y_{t-L}$ and its corresponding magnitude $m = m_{t-L}$ for the current trial index t. We use $\alpha = 1$ for the previous trial estimates, which favors the immediately preceding estimates but retains some dependency on earlier estimates; the model is most likely to sample the preceding trial, then the one before that, etc. As with the familiar mappings, previous trial estimates are sampled without replacement from the power law distribution over previous estimate indices.

The use of previous estimates to support calibration reflects the idea that people might rely in part on previous estimates in order to make an estimate in the current trial that feels "coherent," i.e., calibrated similarly to previous estimates. The power law function from which previous estimates are sampled is consistent with the fact that insofar as people may be calibrating their current estimate in part based on what they said previously, this would be most likely for the immediately preceding trials. Here, the model does not have any knowledge of whether the magnitude corresponding to the sampled

estimate reflects an accurate mapping. Instead, it simply has access to the magnitude representation m that corresponded to a particular estimate y. Where this estimate happened to be accurate, the model will benefit from such a well-calibrated sample, but if it was inaccurate, this may support the sort of miscalibrations described in Izard and Dehaene (2008). The combination of samples from previous trials and familiar number mappings accords with work in psychophysics arguing that for highly familiar domains like color, people may exhibit more 'absolute' categorization, while in less familiar domains people show a great deal of sensitivity to local context (Laming, 1984). Here, number estimation presents a sort of *hybrid*, with lower number estimation more likely to rely on some amount of absolute judgment (samples from memory), while estimation of higher numbers may rely more on local context (preceding trials).

Summary

The sampling process described previously yields a set of length n where each element is a {magnitude, number} tuple $\{\mu_j, \gamma_j\}$ which is either a noisy but accurate associative mapping from "familiar magnitudes" or the magnitude and corresponding estimate from a previous trial, with the relative proportions of familiar mappings and previous trials in a set of samples determined by k. The model's sampled estimates allow it to compute a stepwise likelihood function over possible estimates y for the stimulus magnitude m as a product of the likelihoods of each sampled estimate via the process described above. This likelihood is then scaled by the prior and normalized to generate a posterior distribution over number values which the model samples to produce an estimate.

On this account, if the sample parameter n (dictating the number of samples that will inform each estimate) is large, the model draws on a richer set of previous experiences for its estimation; when the probability of sampling from a known mapping k is also large, the model has a more reliable set of guideposts mapping from magnitude to number which it uses to make a novel estimation. If k is small, the model relies primarily on previous

estimates it has made to calibrate its mapping from magnitudes to estimates. Given this, the proposed model will be trivially successful for a suitably large n and k and will be hopelessly inaccurate with sufficiently low n and k. Thus, we begin by asking whether the model can achieve human-like performance with a cognitively plausible number of samples.

Model Results

Using the procedure described above to generate estimates for each trial, we tested our model with the 300 trials from each of the 24 participants in our experiment. We begin by identifying values of n and k that allow for reasonably accurate estimates. Next, we evaluate how well the model produces the characteristic features of human estimation described in our experimental results. Throughout the remainder of this section, comparison of the model to human performance is done using the participant data from the previous experiment. We use the average of participants' two estimates in each trial, which provides a less noisy set of responses and therefore a more conservative bar for model accuracy. In what follows, we evaluate our model with four central claims:

Claim 1: *The model produces an accurate mapping from magnitude to number even with relatively few samples.*

Claim 2: *The model produces human-like bilinear estimation patterns and underestimation.*

Claim 3: *The model produces variability that is similar to human individual differences in estimate calibration.*

Claim 4: *The model produces human-like drift in estimate calibration.*

Accurate estimates with few samples (Claim 1)

Although the model is greatly hindered by making only ordinal magnitude comparisons on each trial, it achieves reasonable performance with a limited number of samples. We hold the probability that a sample comes from a familiar mapping constant at $k = 1.0$ to see how the model performs under idealized conditions and evaluate model

performance for varying numbers of samples n. Figure 6 (top) shows model performance alongside the same estimates for three sample participants. With $n = 20$ samples drawn from familiar number mappings, the model produces a reasonably accurate function from internal magnitude to number. Model estimates cluster around the identity line at lower numbers and don't deviate substantially more than people do at higher magnitudes.

To quantitatively compare human and model estimates, we plot model estimate mean squared error (MSE) for increasing values of n samples alongside human MSE from the estimates in our experiment data. Figure 6 (bottom) shows MSE of human estimates compared to model estimates for increasing numbers of samples. The model reliably surpasses overall human accuracy at 15–20 samples. This finding is robust to alternative values of k: With $k = 0.5$, the model drops below human MSE at a similar n to $k = 1.0$. Thus, reasonable estimate performance by the model doesn't hinge on idealized learning conditions. This result is compatible with prior research showing that adults may have on the order of 15 strong *associative mappings*, i.e., numbers for which they have direct and accurate mappings from magnitude to number (Sullivan & Barner, 2013). We therefore find support for *Claim 1*, that under reasonable parameter values, the model is able to attain an overall accuracy that is comparable to humans and resembles in broad strokes the character of human estimation.

Bilinear estimation with human-like underestimation (Claim 2)

Underestimation at larger magnitudes is perhaps the most salient and well-documented feature of human number estimates. In our experimental data, this was characterized as a log-bilinear estimation function and was shown to have a better fit than a simple power law mapping. One possible account of the underestimation pattern is that if people are more likely to encounter lower numbers in everyday experience, they will likely be more calibrated in estimating lower numbers; therefore, when encountering a higher number than they are used to seeing, participants might fall towards more familiar (but still

plausible) numbers in their estimates, thus producing a general pattern of underestimation. Such a tendency might even be considered rational, given the power law governing "need probability" of increasing integers (Cheyette & Piantadosi, 2020; Piantadosi, 2016). Consistent with this, the model's prior on lower numbers and the power law sampling of familiar mappings at lower values together produce a bias towards lower estimate values.

To compare model bilinear estimation fits to the full set of participant data, we plot model cutoff and bilinear slope parameters alongside the fitted parameters for our 24 experimental subjects (model settings remain at $n = 20$ samples, $k = 0.5$ probability of estimates from familiar mappings). Figure 7a shows the aggregate set of human estimates with a single cutoff and slope parameter alongside cutoff and slope values for a matched set of model estimates. These fits are nearly indistinguishable, reflecting the overall trend of the model to underestimate similarly to humans. Figure 7b shows the distributions of cutoff and slope parameter values fit to individual subjects, with the average model cutoff and slope values when simulating individual participants shown in red. The average model fit is well within the range of human estimates, particularly for the fitted slope. Finally, Figure 7c presents the same data in finer detail: Human cutoff and slope estimates are plotted together with average model cutoff and slope values. This comparison shows that the model occupies a position comparable to human estimates in fitted "cutoff-slope" space. Broadly, Figure 7 reflects *Claim 2* outlined above: The model is able to capture the human patterns of bilinear fit and underestimation at higher magnitudes.

Human-like individual differences (Claim 3)

In our experimental data, we found large individual differences in estimate calibration at higher magnitudes, which we described as varying bilinear slopes fit to the estimates. In line with this, we consider here how a model of the mapping from subjective magnitudes to verbal estimates might capture variability across participants. The model as we've described it so far has free parameters for the number of samples n that participants

use to assemble an estimate and the probability k that each sample comes from familiar number mappings or previous estimate trials. These mappings are sampled anew at every trial, though the distribution over familiar mappings and previous trials heavily constrains this re-sampling. We modify this *baseline* model in favor of an *individual differences* model that uses a fixed and limited set of familiar mappings across all trials. We vary this fixed set across the model's participant simulations to capture individual variability in the range of numbers for which participants have a strong recognizable mapping. To generate each subject's estimates, the model samples exclusively from this subject's fixed set of mappings rather than sampling each time over the full range of *possible* mappings.[6]

Formally, recall that an individual's associative mappings are expressed as vectors μ, γ, where individual magnitude to number mappings sampled for a given estimate are expressed as {magnitude, number} tuples $\{\mu_j, \gamma_j\} \in \mu, \gamma$. For each set of 300 subject estimates generated by the model, we now populate the vectors μ, γ once with I unique magnitude to number mappings (these are sampled with the same Memory(α) function that generated familiar mappings *for each trial* in the baseline model, using the same α). This set is then fixed for all the subject's trials; each individual estimate draws samples $\{\mu_j, \gamma_j\}$ from this constrained set of memories. To illustrate, if $I = 10$, a given subject's set of mappings is most likely to include numbers in the range 1–10 and highly unlikely to contain, e.g., 230. However, the set of larger numbers that *do* get sampled for each subject will likely vary across subjects. For each estimate trial, samples from among the set of mappings I are drawn from the same power law distribution that initially generated I, with the probabilities initially assigned to each number y normalized across the set of mappings in I. The baseline model can be seen as a special case of this general model in which I is equal to 1,000, or the full range of numbers that our model considers when

[6] Though the associative mapping concept is useful within the domain of number estimation, these stable mappings might simply be thought of as "memories" or reliable associations in other problems that involve mapping from subjective internal representations to formal systems.

producing estimates.[7]

Within this framework, the free parameter I gives us a knob with which to tune individual variability.[8] High values of I will be closer to the baseline model conditions, where participants have very similar distributions from which they sample associative mappings for each estimate. However, low values of I will create more idiosyncratic distributions of associative mappings across subjects, thereby changing the mappings that each subject is likely to draw on for a given estimate. We define a *low variability* model with I = 1,000 associative mappings (i.e., one for every possible number estimate) and a *high variability* model with $I = 10$ mappings (we set $n = 20$ and $k = 0.5$ as above to maintain continuity). We are interested first in whether the low value of I in the high variability model has the desired effect of increasing the variability of bilinear slopes estimated across individual "subject" simulations by the model. Second, we want to know whether this high variability model produces individual differences in actual estimates comparable to what is seen in participant data from our experiment.

We find that the high variability model produces large variability in fitted slopes and that the corresponding estimates are similar to human individual differences. In Figure 8a, we plot the distribution of slopes for the high and low variability models, with the average human slope indicated by a dashed line. The high variability model has a notably larger distribution of fitted slopes than the low variability model, though both models have large mass around the slopes best fit to participants. How then does the variability of estimates for the high variability model compare to human participants? In Figure 8b, we plot the (modular) split-half correlation of fitted slopes for human estimates from our

[7] In fact there is a slight difference between the baseline model and the individual variability model with I = 1,000 because the baseline model will re-sample the magnitude value associated with a given number for each estimate, whereas the individual variability model keeps everything about these mappings fixed from the outset.

[8] There are a number of ways we might have implemented individual differences in this model; the current approach simply suggests that people have different prior experiences and thus different stable associations.

experiment alongside the high and low variability model, as well as the baseline model fits. Recall that in our experimental data, we used this same measure to assess the degree of individual variability in participant estimates. The low variability and baseline models are comparable in their split-half correlations, as expected, while the high variability model attains a split-half correlation similar to humans. This suggests that by giving the model a sampled set of familiar associative mappings which is stable across estimates but varies between participant simulations (maintaining all other parameters as before), we're able to produce a variability of estimate calibrations that is close to the individual differences between human subjects in our experiment. In line with *Claim 3*, our model offers a simple account by which we might explain the large individual variability in human estimates.

Human "drift" across many trials (Claim 4)

In our experimental data, we show that estimate calibration exhibits a *slow drift* as bilinear calibration slopes wobble over the course of many trials. We hypothesized that this drift is a result of continual updating of the mapping from magnitude to formal number as more data is encountered. Our model formalizes this prediction through its ongoing dependence on previous estimates. Here we show that, as with bilinear underestimation and individual variability, the model's estimate calibration drifts similarly to humans, providing evidence that human drift is explained by an effort to maintain consistency with one's previous estimates.

In order to estimate a number for a perceived magnitude, our model relies on a combination of familiar magnitude to number mappings and previous trial estimates. While the familiar mappings are fairly stable (particularly in the individual variability model considered above), the mappings from previous estimates are inherently dynamic. The model's k parameter determines the proportion of samples on a given trial that come from these more stable mappings. When this number is low, the model's estimates will be more dynamic and its auto-correlation higher as a result of relying more heavily on

previous estimates. We therefore expect that so long as the model's k value allows for sufficient dependence on previous estimates, it should "drift" as human calibration does.

Our measure of model estimate drift was calculated using the same process as in our experimental results. Model estimates were divided into bins of 10 consecutive trials, creating 30 such bins for each model "participant" over the 300 total trials. We then calculate a best-fitting log bilinear cutoff and slope parameter for the estimates in each bin using maximum likelihood estimation. The pairwise correlations between each participant's slopes in different trial blocks are aggregated by the average "trial distance" between blocks and the drift in estimate calibration described in our experiment results is revealed by the decrease in this slope correlation at greater and greater trial distances.

A notable feature of our empirical results was that despite the slow drift in estimate calibration, pairwise slope correlations were very high at low trial distances (0.8 – 0.9) and remained high even at trial distances approaching 300 (> 0.5). We hypothesized that this was due to the high individual variability in slope estimates, as well as people remaining fairly well-calibrated across more distant trials. We are therefore interested in the degree to which our model can capture these additional empirical features of the human estimate data as well. To explore this, we calculate the estimate drift for both the *baseline* model and the *high variability* model described previously. The degree to which the baseline model estimates drift provides an indication of how much the reliance on previous estimates by itself produces a slow drift in estimate calibration; then, the inclusion of the individual variability model provides an indication of how much individual variability of the sort we explored in the previous section contributes to the high overall slope correlations at both short and long trial distances. For continuity with previous results, both models use $n = 20$ samples for each estimate and a probability $k = 0.5$ that each sample in a given estimate draws on familiar mappings. As above, the individual variability model, which was able to simulate human individual differences by reducing the number of unique familiar mappings for each subject, samples from $I = 10$ stable memories

for each participant run of the model.

Figure 9 shows the drift in slope correlation over trial distance for participant and model data. The human data reflects the pattern first illustrated in our experimental results: At low trial distances, human estimates have a high correlation of fitted slopes, reflecting the stability of individual estimate calibration at close blocks of trials as well as the individual variability of fitted slopes across subjects. However, as trial distance increases, the correlation of fitted slopes decreases, reflecting the fact that human estimate calibration seems to be subject to an ongoing updating process throughout the task which makes more distant trial blocks less similarly calibrated. The baseline model data in Figure 9 shows a qualitatively similar pattern, with the correlation between fitted slopes in more adjacent blocks of trials decreasing gradually as trial distance increases (i.e., correlation is higher for more adjacent compared to less adjacent blocks of trials). This illustrates that the model, even in its most basic architecture, produces a drift in estimate calibration over time. At the outset, the model has a much lower correlation of slope estimates than humans due to lower "individual differences" for the model across simulated participants compared to human estimates. Nothing about the baseline model changes between "participant" estimate simulations, thus reducing the individual variability that can contribute to a correlation coefficient compared to 24 different human participants.

In Figure 9, the individual variability model has a gradual decrease in correlation of fitted slopes at greater trial distances (i.e., a drift in calibration). However, the individual variability model also has a much higher auto-correlation of slopes at the outset and maintains a higher correlation over increasing distances. This is consistent with the idea that the drift in human estimate calibration is a function of both ongoing updating of the estimate function over time, in combination with large individual differences in overall calibration. The individual variability model provides a reasonable approximation of this, though it does not reach a slope correlation as high as human estimates.

Given the success of the high variability model in capturing individual differences

in estimate calibration in the previous section, we might have expected it to exhibit slope correlations more similar to humans in the current analysis. However, while the previous section examined only *split-half* slope correlations, the current analysis is based on *split-30th* correlations and is therefore a more sensitive measure of individual differences. Second, the high variability model simulates individual differences by allowing for an idiosyncratic set of familiar mappings that each participant uses. In the previous section, we showed that this simple modification can produce variability in overall estimate calibration that is similar to humans. However, it is unlikely that this is the *only* source of individual variability in human estimate calibration. Other sources of stable individual variability not captured by our model might further increase the correlation of human estimate slopes. Therefore, it is perhaps not surprising that the subtle measure of estimate calibration over time shown in Figure 9 does not have a slope correlation as large as human estimates for a given trial distance, even with our high variability model.

Importantly, our model provides a plausible account of how calibration drift might arise in humans. Our model seeks to maintain a dynamic "coherence" in its estimates by continually updating the mapping from magnitude to number estimates based on the magnitudes of previous trials and the corresponding estimates produced. This process, in combination with variability across model runs, produces drift in estimate calibration that is similar to the pattern seen in humans. If people are also updating their mapping from magnitude to number based on new data they receive and trying to maintain some ongoing coherence with their most recent estimates, then our model offers a proof of concept that this process could explain the drift seen in human estimation.

General Discussion

We investigated the process by which people translate between perceptions of magnitude and formal representations of number when making numerical estimates. In particular, we asked what sort of mapping best explains people's ability to calibrate their

estimates and how that mapping might work. Our experimental results produce two novel findings. First, we show that participants' individual estimate functions are best modeled as a *bilinear* function in log space rather than as a simple log linear function, contrary to previous proposals (Kaufman et al., 1949; Krueger, 1982). Under this formulation, people are highly accurate up to a unique threshold, after which their estimates exhibit a sublinear relationship with numerical magnitude. Second, we show that the slope of this bilinear function varies not only across individuals (as shown in previous findings) but *within* individuals over many trials, suggesting that people's estimate function is subject to an ongoing updating process that may incorporate information from previous estimates.

Recent research addressing how people learn to map magnitude representations to symbolic number estimates has made a distinction between *associative mappings*, in which magnitude representations correspond to unique number values, and *structure mappings*, in which number values are assembled through more relative notions of distance and ordering of magnitude representations (Sullivan & Barner, 2013). Indeed, evidence suggests that people use a combination of both associative and structure mappings, with associative mappings mostly detected for smaller integers and developmental changes in estimation accuracy corresponding to improved structure mapping (Sullivan & Barner, 2013, 2014a, 2014b). Our experimental results bear on this existing work in two ways. First, the finding that people's estimates are best described with a log-bilinear function has an obvious isomorphism to the use of associative and structure mappings in estimation. Future work should explore the relationship between associative mappings and the "cutoff" found in our bilinear model, and further between structure mapping of higher magnitudes and the idiosyncratic bilinear slopes fit to individual estimates. Second, the finding that people show a dynamic uncertainty in the mapping from magnitude representations to number estimates, which causes a "drift" in their estimate calibrations, offers a refinement of our understanding of structure mappings. The proposed explanation for this drift, that people are continually updating their mapping function to be consistent with prior estimates, is

consistent with the structure mapping account, but suggests that such structure mapping is not a static process but is instead a dynamic one.

Our empirical results are not easily accommodated by existing models of the mapping from internal representations to formal estimates. We therefore offer a computational model of the process by which people might accomplish this mapping from internal magnitude representations to symbolic number. In the tradition of "decision by sampling" models of Stewart et al. (2006) and earlier psychophysical models of absolute judgment for novel stimuli (Laming, 1984; Stewart et al., 2005), our model assumes only that people have the ability to sample the magnitudes and corresponding number values from a limited set of prior estimates and "familiar mappings." Despite these constraints, our model is able to generate estimates with human-like levels of calibration using only ordinal comparisons between an observed magnitude and the sampled estimates.

We evaluate the model by its ability to reproduce the characteristic patterns of human mappings from internal magnitude to number described in our experimental results. First, we show that the model achieves human-level performance with a limited number of samples (15–20), which by itself was not a given since the model has a highly limited set of operations and knowledge to compute an estimate. Next, we show that model estimates, under reasonable conditions of the free parameters, are characterized by a log-log bilinear fit which strongly resembles the bilinear character of human estimates discovered in our experiment. We then show that a simple extension of the model produces individual variability which is comparable to the individual differences present in human estimates. Finally, we show that the calibration of the model's estimates is subject to a drift over the course of many trials which is similar to the pattern of human estimate calibration. With this latter result, we offer a candidate explanation for the source of human calibration drift, namely a reliance on sampled prior estimates to coordinate the current estimate, which produces a high correlation in estimate calibration between nearby estimates that decreases over large trial distances.

While the model offers several novel results, there are a number of ways in which future work might further validate it. First, the results presented here are based on human estimates for numbers drawn from a geometric distribution that extends the range of numbers used in prior work but that nonetheless favors smaller numbers. Since the model relies on previous estimates to inform its current decision, the model's behavior may be dependent to some extent on the underlying stimulus distribution, along with the exponential distribution from which familiar mappings are sampled. In the current work, these distributions were chosen to reflect the probability of encountering and needing particular number representations (Piantadosi, 2016). However, prior work in psychophysics has shown that people's mapping functions may indeed be malleable given different distributions of stimuli (Haubensak, 1992). In this vein, future work might explore the sensitivity of the model to much larger magnitudes or to stimulus distributions that differ from the one used in the current experiment.

Additionally, as noted above, support for the model comes primarily from its ability to capture a wide range of empirical phenomena in number estimation—including the novel "drift" in calibration observed over many trials—along with its generalizability to broader psychophysical domains. There are a number of robust behaviors exhibited in prior psychophysical work which the model might account for (for a review, see Stewart et al. (2005)). In particular, sequential effects of *assimilation* and *contrast*, whereby people's responses are pulled closer to the preceding trial magnitude and away from more distant trials, have been exhibited across a range of psychophysical judgments (Garner, 1953; Holland & Lockhead, 1968; Ward & Lockhead, 1970). These effects are difficult to account for in models that rely on stable internal scales, but can emerge somewhat naturally from models in which responses are calibrated based on previous trials (Stewart et al., 2002, 2005). Whether this and other classic psychophysical effects can be produced by our model is difficult to say because of its reliance on both previous estimates and familiar mappings, but the model's tendency to sample immediately preceding trials for comparison (similar to

the weighting of differences in Stewart et al. (2005)) might in principle yield assimilation and other effects. Recent work has explored the distinct role of stimulus and response in assimilation of facial expression perception (Hsu & Wu, 2020); given the parameterization of stimuli (familiar mappings) and prior responses (earlier trials) in our model, similar investigations offer a potential avenue of future work.

Finally, though the model captures a range of behavioral phenomena in estimation, its parameters were not fit to our empirical data, limiting the ability to do precise model comparison. The decision not to fit model parameters was both practically and theoretically motivated. In contrast to existing computational models which describe estimation at the level of aggregate behavior (Izard & Dehaene, 2008), our model offers a process account of individual trial-level responses based on samples from memory of familiar mapping and previous trials. Fitting the model to individual trial/subject data would therefore require estimating the specific bundle of historic, and transient, exemplars available to an individual subject on a given trial, which is not currently possible. Instead, we show that this process produces effects which are broadly consistent with observed behavior in estimation (similar to related sample-based models, e.g., Stewart et al. (2006)). At a theoretical level, we are not aware of other similar process accounts for comparison, thus limiting the value of precise model parameter estimation even if it were possible.

The current results have significant implications for the study of numerical reasoning, as well as for broader questions about the nature of how people reason about systems and scales in the world using internally calibrated representations. First, we show that people display measurable uncertainty not just in their magnitude representations but in the way they express these representations as number estimates over time; our experimental results quantify the variability in people's mapping from representation to number. Further, we provide novel evidence that this mapping is best described by a function that is *bilinear* in log space, rather than a simple power law. It is tempting to conclude based on these results that estimation is governed by two distinct processes, one

which allows for accurate estimates up to a threshold and then a second which produces error-prone estimates above the threshold. However, our modeling results indicate that this need not be the case. We show that a unified process of selecting estimates via ordinal comparison to a set of sampled magnitude-number pairs is able to account for robust features of human estimation, including the bilinear estimate function and "slow drift" described in our experimental results. Critically, while previous attempts to model the process by which people estimate number have emphasized a (somewhat) stable internal mapping from magnitude to verbal number (Izard & Dehaene, 2008), our modeling results suggest that this may not be necessary. Instead, we show that the ability to calibrate the present magnitude via ordinal comparison to samples drawn from memory is sufficient to generate accurate and distinctly "human" estimate patterns. This opens the door to future work aimed at understanding the degree to which children's estimation patterns, or other forms of numerical reasoning altogether, might be described by this model.

In addition to offering a unified account of the process by which people generate estimates from subjective representations of magnitude, our model raises a number of questions about the development of this mapping and the individual differences seen in estimate calibration. Our modeling results suggest that overall estimate accuracy, and individual differences in calibration, can be approximated through differences in the number and range of "familiar" magnitude to number mappings that people have, particularly for larger numbers. This suggests the intriguing possibility that estimation ability among children and the putative relationship between estimation and more general numerical reasoning (Halberda & Feigenson, 2008) might be improved, or individual differences among adults lessened, through mere learning of a broader range of associative mappings of the sort our model relies on.

More broadly, the current results tie number estimation into the general challenges people face when mapping between subjective, internal scales and the systems we use to communicate about them. The task of navigating between internal representations of our

everyday experience and formal systems is a part of intuitive reasoning across a range of domains. We regularly make estimates based on fairly concrete representations, e.g., whether we will be able to carry a heavy suitcase to the car or how long it will take to go grocery shopping, as well as more abstract estimates, such as whether the price of concert tickets exceeds how much we expect to value the experience. By recruiting domain-general processes such as sampling relevant "memories" and basic comparison between the current stimulus and those memories, the model outlined here attempts to solve this more general problem with an approach that is not restricted to number estimation. In doing so, we offer a bridge between work in psychophysics which has emphasized the extent to which mapping from internal representations to external systems can be done without a robust internal scale (Laming, 1984; Stewart et al., 2002, 2005) and prior work in number estimation, which has largely assumed a stable scale with an intuitive notion of psychological distance as the basis for numerical reasoning (Izard & Dehaene, 2008). Laming (1984) observed that a model which assumes a limited internal scale could successfully capture a range of behavioral phenomena when categorizing stimuli like auditory tones or line lengths, but might be unable to account for behavior in domains like color (and presumably number), where people have a great deal of prior experience. By incorporating both previous trials and more familiar mappings into our model's estimate process, we offer an account of how prior knowledge and consistency with earlier responses might come together in a domain like number to produce calibrated responses that nonetheless "drift" over time. In this way, we hope our model provides a more generalized view of people's ability to navigate the range of external scales we use every day based on differing amounts of prior knowledge and experience.

Finally, a number of previous results have suggested that complex human judgments of various kinds can be performed via simple cognitive operations over sampled data from memory or the world around us (Bonawitz et al., 2014; Stewart et al., 2006; Vul et al., 2014). By showing that number estimation—and the more general problem of

mapping psychophysical representations to analog formal scales—can be solved using a similar approach, we provide further evidence that the ability to sample and compare (i.e., "decision by sampling" (Stewart et al., 2006)) constitutes a core component of our *algorithmic toolbox* and a critical feature of domain-general human intelligence.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396–408.

Anderson, N. H. (1974). Algebraic models in perception. *Handbook of perception*, *2*, 215–298.

Attneave, F. (1962). Perception and related areas. In S. Koch (Ed.), *Psychology: A study of a science*. McGraw-Hill.

Baird, J. C., Romer, D., & Stein, T. (1970). Test of a cognitive theory of psychophysics: Size discrimination. *Perceptual and Motor Skills*, *30*(2), 495–501.

Birnbaum, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics*, *15*(1), 89–96.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, *74*, 35–65.

Brannon, E. M. (2006). The representation of numerical magnitude. *Current opinion in neurobiology*, *16*(2), 222–229.

Brannon, E. M., & Terrace, H. S. (1998). Ordering of the numerosities 1 to 9 by monkeys. *Science*, *282*(5389), 746–749.

Burr, D., & Ross, J. (2008). A visual sense of number. *Current biology*, *18*, 425–428.

Carey, S. (2009). Where our number concepts come from. *The Journal of philosophy*, *106*.

Carey, S., & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in cognitive sciences*.

Cheyette, S. J., & Piantadosi, S. (2020). A unified account of numerosity perception. *Nature Human Behaviour*, 1–8.

Cheyette, S., & Piantadosi, S. (2019). A primarily serial, foveal accumulator underlies approximate numerical estimation. *Proceedings of the National Academy of Sciences*, *116*(36), 17729–17734.

Clarke, S., & Beck, J. (2021). The number sense represents (rational) numbers. *Behavioral and Brain Sciences*, 1–57.

Collins, A. A., & Gescheider, G. A. (1989). The measurement of loudness in individual children and adults by absolute magnitude estimation and cross-modality matching. *The Journal of the Acoustical Society of America*, *85*(5), 2012–2021.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*, 307–314.

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*, *108*(3), 819–824.

Frank, M. C., Fedorenko, E., Lai, P., Saxe, R., & Gibson, E. (2012). Verbal interference suppresses exact numerical representation. *Cognitive psychology*, *64*(1), 74–92.

Gallistel, C. R. (1990). *The organization of learning*. The MIT Press.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1–2), 43–74.

Garner, W. (1953). An informational analysis of absolute judgments of loudness. *Journal of experimental psychology*, *46*(5), 373.

Gebuis, T., & Reynvoet, B. (2012). The role of visual information in numerosity estimation. *PLoS One*, *7*.

Gescheider, G. A. (1988). Psychophysical scaling. *Annual review of psychology*, *39*(1), 169–200.

Gordon, P. (2004). Numerical cognition without words: Evidence from amazonia. *Science*, *306*(5695), 496–499.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental psychology*, *44*(5), 1457.

Haubensak, G. (1992). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 303.

Holland, M. K., & Lockhead, G. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics*, *3*(6), 409–414.

Hsu, S.-M., & Wu, Z.-R. (2020). The roles of preceding stimuli and preceding responses on assimilative and contrastive sequential effects during facial expression perception. *Cognition and Emotion*, *34*(5), 890–905.

Indow, T., & Ida, M. (1977). Scaling of dot numerosity. *Perception & Psychophysics*, *22*(3), 265–276.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*, 1221–1247.

Kaufman, E., Lord, M., Reese, T., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*, 498–525.

Krueger, L. (1982). Single judgments of numerosity. *Perception & Psychophysics*, *31*(2), 175–182.

Laming, D. (1984). The relativity of 'absolute'judgements. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 152–183.

LeCorre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395–438.

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*.

Lipton, J., & Spelke, E. (2005). Preschool children's mapping of number words to nonsymbolic numerosities. *Child development*, *76*(5), 978–988.

Lourenco, S. F., & Longo, M. R. (2011). Origins and development of generalized magnitude representation. *Space, time and number in the brain*. Academic press.

Lyons, I. M., Ansari, D., & Beilock, S. L. (2012). Symbolic estrangement: Evidence against a strong association between numerical symbols and the quantities they represent. *Journal of Experimental Psychology: General*, *141*(4), 635.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of experimental psychology: general*, *111*(1), 1.

McGill, W. (1954). Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 93–111.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Mori, S. (1989). A limited-capacity response process in absolute identification. *Perception & Psychophysics*, *46*(2), 167–173.

Nieder, A. (2020). Neural constraints on human number concepts. *Current Opinion in Neurobiology*, *60*, 28–36.

Opstal, F. V., & Verguts, T. (2013). Is there a generalized magnitude system in the brain? behavioral, neuroimaging, and computational evidence. *Frontiers in psychology*, *4*, 435.

Piantadosi, S. (2016). A rational analysis of the approximate number system. *Psychonomic bulletin & review*, *23*(3), 877–886.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological science*, *19*(6), 607–614.

Sanborn, A., & Beierholm, U. (2016). Fast and accurate learning when making discrete numerical estimates. *PLoS computational biology*, *12*(4).

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive psychology*, *7*(1), 82–138.

Shepard, R. N. (1981). Psychological relations and psychophysical scales: On the status of "direct" psychophysical measurement. *Journal of Mathematical Psychology*, *24*(1), 21–57.

Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations.

Simon, T., & Vaishnavi, S. (1996). Subitizing and counting depend on different attentional mechanisms: Evidence from visual enumeration in afterimages. *Perception & Psychophysics*, *58*(6), 915–926.

Starkey, P., Spelke, E. S., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition*, *36*(2), 97–127.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *The American journal of psychology*, *69*(1), 1–25.

Stewart, N., Chater, N., & Brown, G. (2006). Decision by sampling. *Cognitive psychology*, *53*(1), 1–26.

Stewart, N., & Brown, G. D. (2004). Sequence effects in the categorization of tones varying in frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 416.

Stewart, N., Brown, G. D., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 3.

Stewart, N., Brown, G. D., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological review*, *112*(4), 881.

Sullivan, J., & Barner, D. (2013). How are number words mapped to approximate magnitudes? *The Quarterly Journal of Experimental Psychology*, *66*(2), 389–402.

Sullivan, J., & Barner, D. (2014a). The development of structural analogy in number-line estimation. *Journal of Experimental Child Psychology*, *128*, 171–189.

Sullivan, J., & Barner, D. (2014b). Inference and association in children's early numerical estimation. *Child development*, *85*(4), 1740–1755.

Testolin, A., & McClelland, J. L. (2021). Do estimates of numerosity really adhere to weber's law? a reexamination of two case studies. *Psychonomic Bulletin & Review*, *28*(1), 158–168.

Treisman, M. (1964). Sensory scaling and the psychophysical law. *Quarterly Journal of Experimental Psychology*, *16*(1), 11–22.

Vul, E., Goodman, N., Griffiths, T., & Tenenbaum, J. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in cognitive sciences*, *7*(11), 483–488.

Ward, L. M., & Lockhead, G. (1970). Sequential effects and memory in category judgments. *Journal of Experimental Psychology*, *84*(1), 27.

Whalen, J., Gallistel, C., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*(2), 130–137.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11.

Yates, M. J., Loetscher, T., & Nicholls, M. E. (2012). A generalized magnitude system for space, time, and quantity? a cautionary note. *Journal of Vision*, *12*(7), 9–9.

Zwislocki, J. (1983). Group and individual relations between sensation magnitudes and their numerical estimates. *Perception & psychophysics*, *33*(5), 460–468.

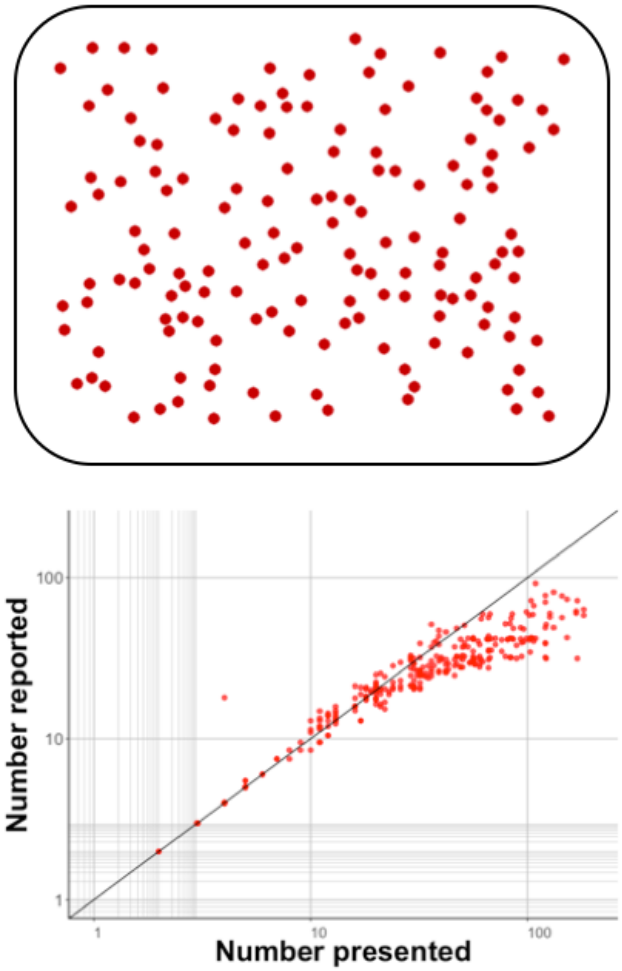


Figure 1

(Top) Participants saw 300 trials in which an array of n dots was briefly presented and participants made a guess as to the number of dots shown. (Bottom) A representative subject's data over all 300 trials with number presented on the x-axis and number reported on the y-axis (both log scale).

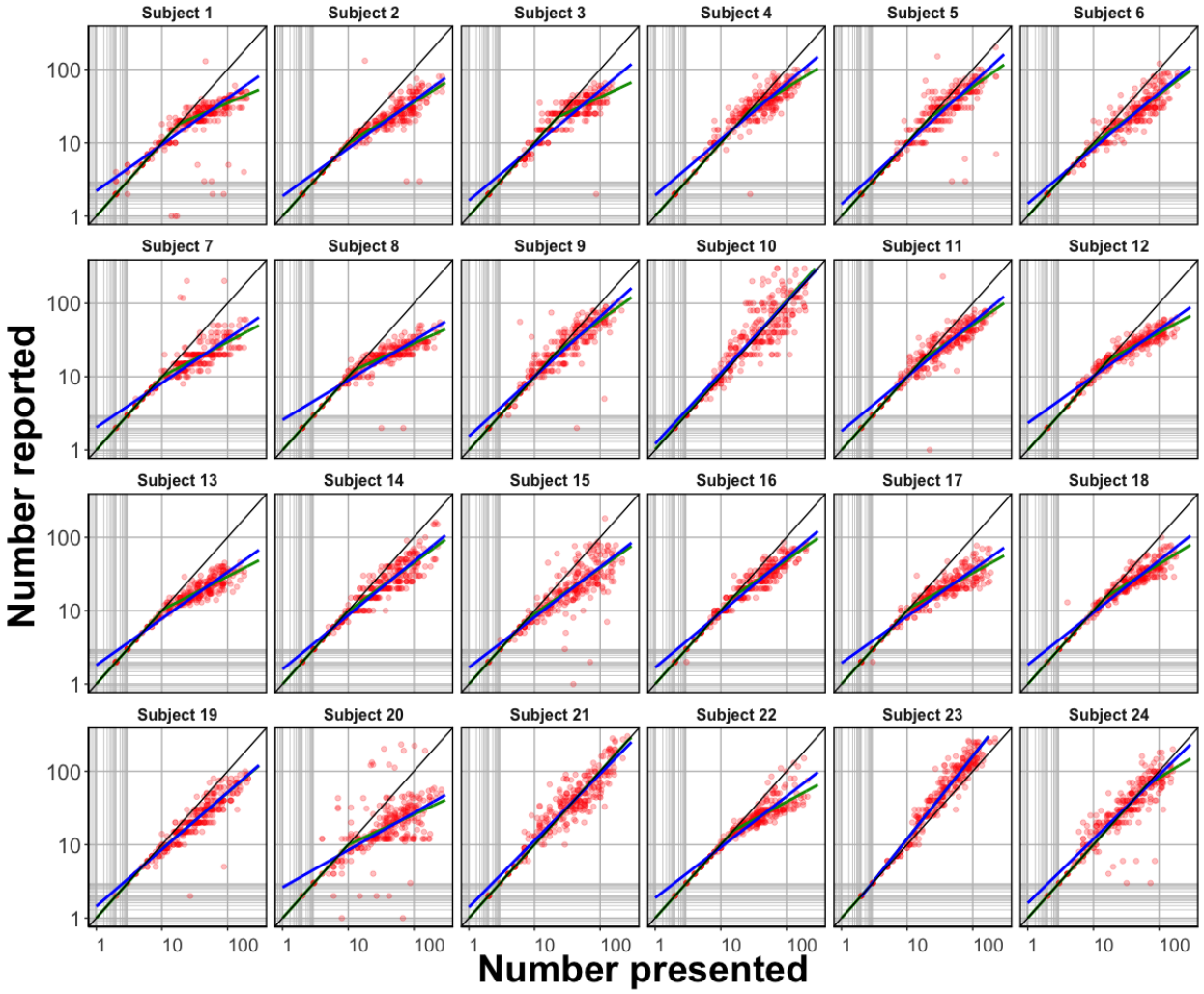
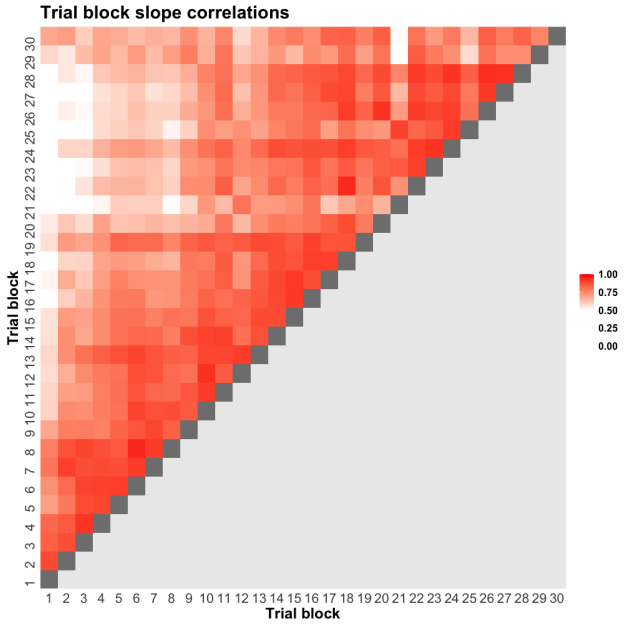
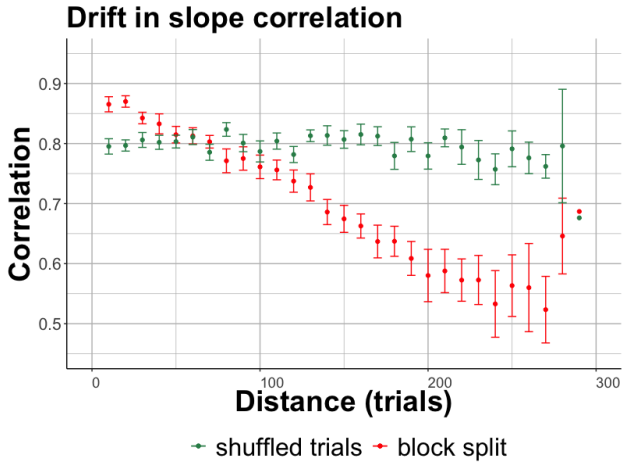


Figure 2

Individual subject estimation data (red points) along with best fitting linear (blue) and bilinear (green) mapping functions in log space. Some of our conclusions may be seen in the raw data alone: (1) systematic mis-estimation occurs for larger, but not smaller, numbers, (2) participants show individual differences in their estimation biases, and (3) estimate variability in log space increases with magnitude.



(a) *Correlation in trial slopes across blocks of 10 trials for all participants. Calibration "drift" is reflected in the high correlation in blocks close to each other (near the diagonal) and lower correlation between more distant trial blocks (further from the diagonal).*



(b) *Correlation in trial slopes by trial distance for distances between trial blocks shown above (red), compared to the same correlation when trials are shuffled (green).*

Figure 3

Decreasing correlation in trial slopes over more distant trial blocks (a) and over time compared to the high correlation across all distances when trial order is shuffled for each participant (b).

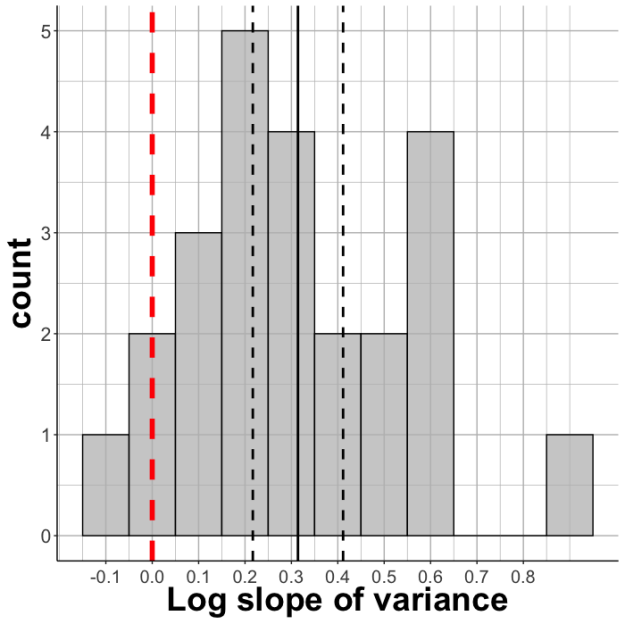


Figure 4

Distribution of slope parameters for the standard deviation of the (log) normal distribution used to determine each participant's estimate likelihoods. Slopes greater than zero represent an increase in variability of log estimates as a function of log magnitude, i.e., an increasing coefficient of variation. The red line indicates the expected average of 0 and the black line indicates the mean of the fitted slopes, with 95% CI indicated by the dashed black lines.

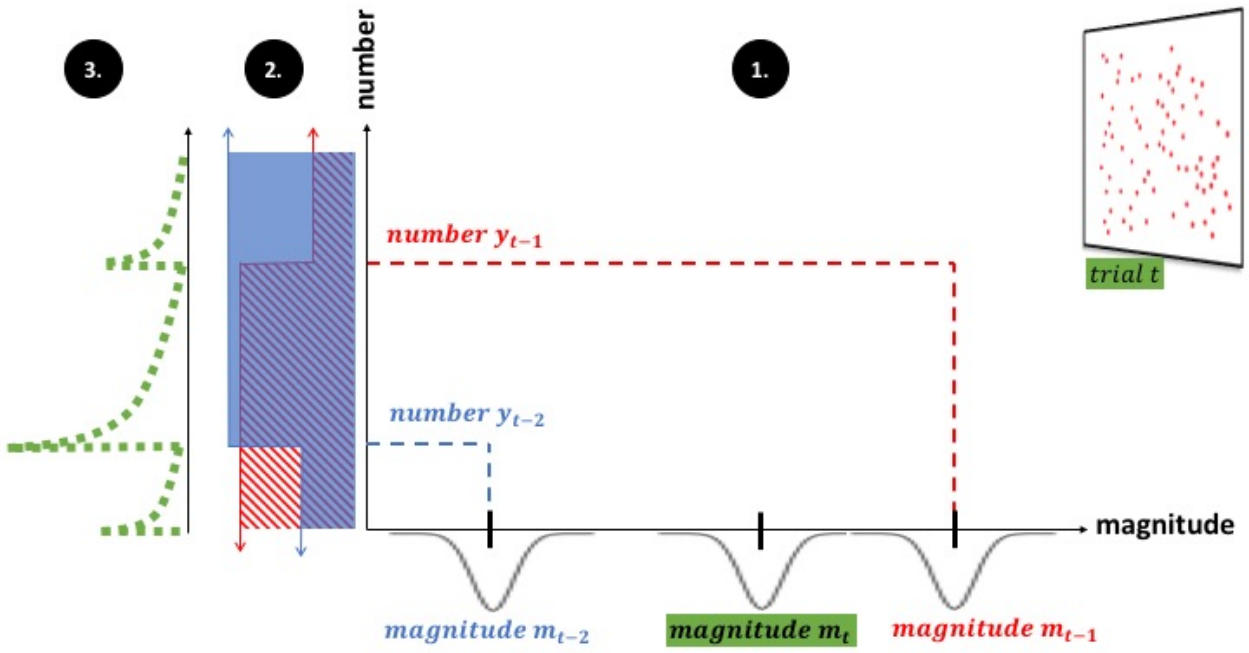


Figure 5

An overview of the estimation model. (1) a new trial t has an approximate magnitude m_t. (2) an ordinal comparison between m_t and magnitudes from sampled estimates (shown here for trial $t-1$ and $t-2$) produces a likelihood function over possible number estimates for m_t. The product of these individual sample likelihoods forms the general likelihood function. (3) combining the likelihood with a prior favoring lower numbers, the estimate for trial t is drawn from a posterior distribution over number estimates shown in green at far left.

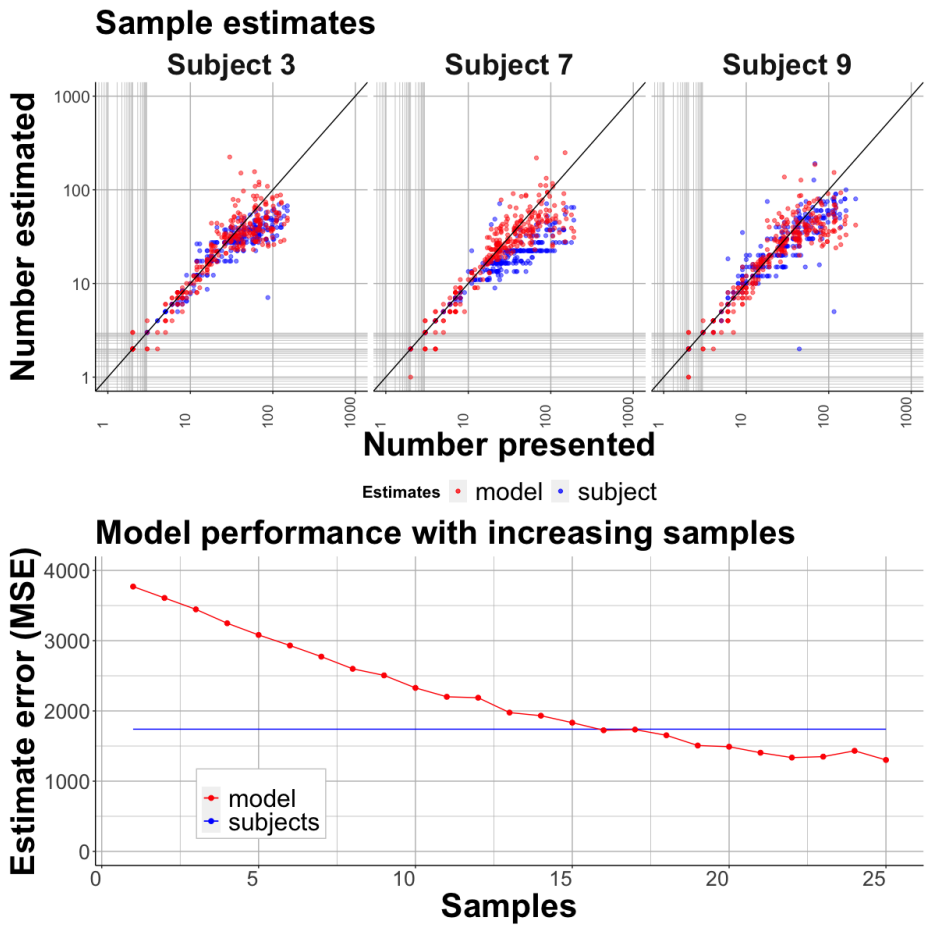
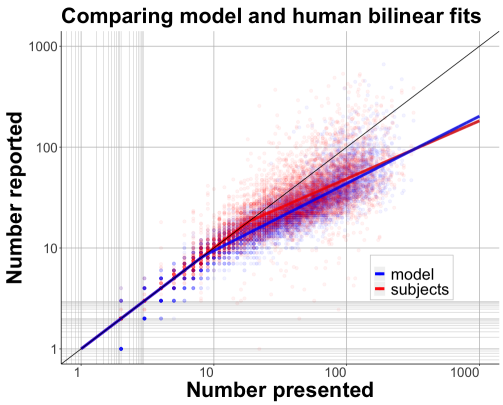
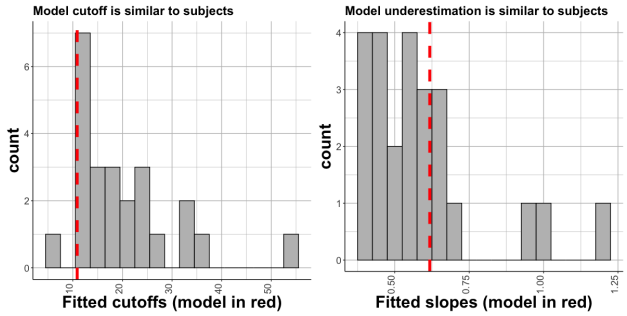


Figure 6

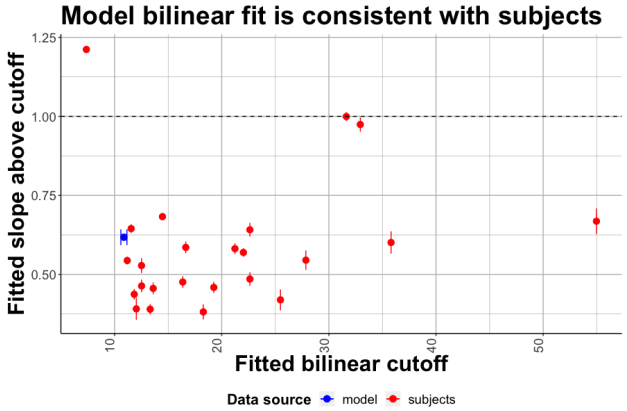
(Top) Model performance plotted alongside participant estimates for three sample participants. Using only $n = 20$ samples and a probability $k = 1.0$ that each sample is drawn from a familiar number mapping, the model achieves reasonable performance qualitatively. (Bottom) Mean Squared Error (MSE) of model estimates plotted alongside human estimates for comparison. With a probability $k = 1.0$ that a given sample comes from a reliable benchmark, the model is equivalent to human performance after only around 15 samples.



(a) *Model bilinear estimation alongside human results for all experiment runs reflect aggregate similarity.*



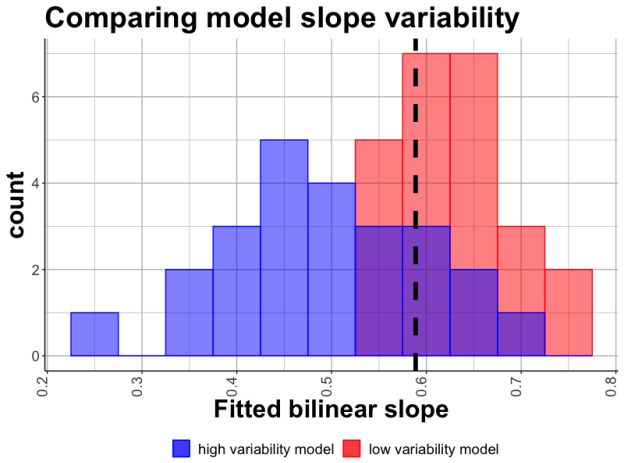
(b) *Distribution of bilinear parameter fits across participants. Average model bilinear parameter values are shown in red.*



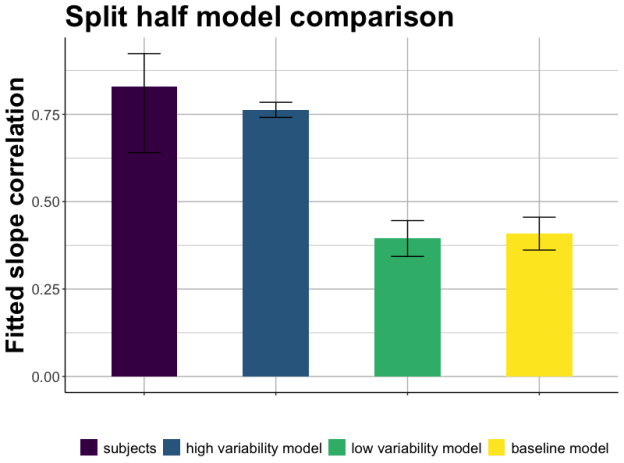
(c) *Bilinear slope and cutoff values across participants with fitted cutoff on x, and slope on y. Average model bilinear parameter values are shown in blue with error in either direction.*

Figure 7

Three views of how our model captures the bilinear shape of human estimates.



(a) *Fitted bilinear slope values across "participant" estimations for the high and low individual variability models. Average human slopes indicated by the dashed line.*



(b) *A split-half correlation indicates that the high variability model has individual variability in estimate slopes closer to human levels, compared with the low variability and baseline models (error bars represent 95% CI for subject data, SE of 10 model runs).*

Figure 8

Comparison of individual variability models to subject data. The high variability model, which is meant to capture human estimate patterns, samples from 10 familiar mappings while the low variability version represents a null comparison which samples familiar mappings during each trial from the full number range.

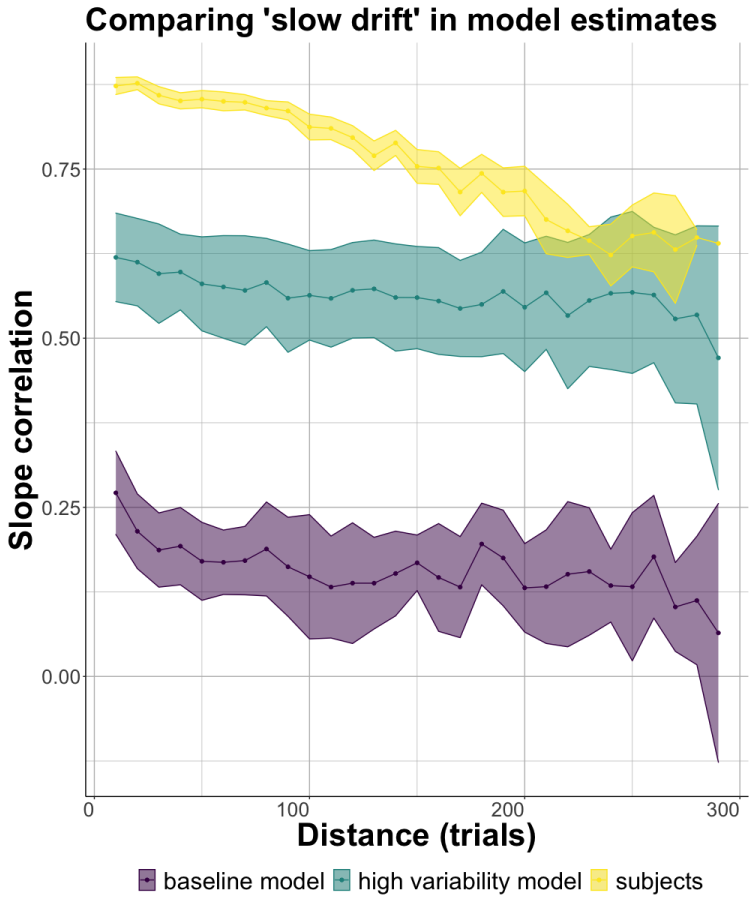


Figure 9

Comparison of estimate calibration "drift" for human subjects and model versions. Both model versions have the same qualitative pattern as human subjects (i.e., smoothly decreasing autocorrelation over greater trial distances). The high individual variability model is closer to human patterns of drift, though neither model's autocorrelation decreases as steeply as the empirical data. Ribbons reflect SEM of human subject data and SD of multiple model runs, respectively.