Categories of Large Numbers in Line Estimation

David Landy

Indiana University & University of Richmond

Arthur Charlesworth

University of Richmond

Erin Ottmar

Worchester Polytechnic Institute

Keywords: Numerical Cognition, Mathematical Cognition, Concept Development, Experimental Psychology

Author Note

Thanks to Zach Davis, Brian Guay, and Megan DeLaunay for helpful comments on this research throughout its development, and to John Opfer, Nora Newcombe, Bob Siegler, and Mark Ashcraft for helpful comments on the manuscript. This research was partially supported by IES Grant R305A110060.

Abstract

How do people stretch their understanding of magnitude from the experiential range to the very large quantities and ranges important in science, geopolitics, and mathematics? This paper empirically evaluates how and whether people make use of numerical categories when estimating relative magnitudes of numbers across many orders of magnitude. We hypothesize that people use scale words—*thousand, million, billion*—to carve the large number line into categories, stretching linear responses across items within each category. If so, discontinuities in position and response time are expected near the boundaries between categories. In contrast to previous work (Landy, Silbert, and Goldin, 2013) which suggested only that a minority of college undergraduates employed categorical boundaries, we find that discontinuities near category boundaries occur in most or all participants, but that accurate and inaccurate participants respond in opposite ways to category boundaries. Accurate participants highlight contrasts within a category, while inaccurate participants adjust their responses toward category centers.

Keywords: mathematical cognition, concepts, numerical reasoning, number processing, category adjustment theory,

Use of categorical information in placing numbers on a line

Introduction

Reasoning in mathematics often involves taking structures well-defined in particular concrete domains and extending them to new, less accessible contexts. For example, exponents are often presented initially to learners as repeated multiplications: x^2=x*x. This idea is then extended to zero exponents, fractional and real exponents, and even complex-valued exponents. Similarly, the natural number system extends the perceivable, countable numbers far outside any possible experience. For instance, currently the largest known prime number (the 48th Mersenne prime) would contain 17,425,170 *digits* if written as an Arabic numeral.

This paper focuses on a very elementary instance of the extension of structures from the concrete to the abstract: the extension of the feasibly countable numbers into the range somewhat beyond them—about 10^6-10^{12} (one million to one trillion). These numbers fall into an important liminal group: so large that experiences of the relevant numerosities are vanishingly rare, but small enough to play important roles in many sciences including geology, physics, astronomy, and macroeconomics, and in political contexts such as debates about taxes and national deficits. In this paper, we'll call these *large numbers*. There is little reason to think that evolution would have specially prepared us to deal with quantities of this magnitude. How do we use these representations to make sensible quantitative judgments?

Large Numbers through Extending Numerical Resources

Understanding of very small number magnitudes has been attributed to properties of a logarithmically scaled (Dehaene et al, 2008) or linearly scaled (Leslie, Gallistel, and Gelman,

2008) internal system—or to the joint action of both systems (Kanayet, Opfer, & Cunningham, 2010). In particular, Harvey et al (2013) have found evidence for neural systems sensitive to presented numerosities up to around 20. Both adults and children are able to respond quite linearly to much larger numbers in at least some contexts (Siegler & Opfer, 2003) and monotonically in others (Izard & Dehaene, 2008; Sullivan & Barner, 2012). Although debate exists over the specific processes implementing this capacity (Ebersbach, Luwel, Frick, Onghena, & Verschaffel, 2008; Moeller, Pixner, Kaufmann, & Nuerk, 2009, Sarnecka & Cohen, 2014; Hurst, Monahan, Heller, & Cordes, 2014; Rouder & Geary, 2014; Slusser, Santiago, & Barth, 2013; Cohen & Blanc-Goldhammer, 2011; Barth & Paladino, 2011), the basic phenomena are not in question.

While the upper bound of these systems is currently unknown, it is clear that as numbers get larger, these direct mental 'representations' must end. At some point even a log-based neural scale must run out of space. People simply cannot have lognormal neural response systems that span the natural numbers, since brains are finite in size, nor can people have created the entirety of the natural numbers through a completed infinity of recursions. Typical environments do not require the individuation of sets as large as 10^6 (1 million) or 10^9 (1 billion), making these *large numbers* a plausible place to look for a transition in the construction of magnitude estimates. How do we deal with these quantities when we do encounter them? One possibility is that we induce new tokens as needed throughout life, extending a process essentially identical to that used to induce numbers small enough to encounter frequently. For instance, numbers in the range of 10^9 might be recursively generated through successorship (Leslie, Gelman, & Gallistel, 2008; Gelman, 2011). On this account, our cognitive systems do not represent a completed

infinity, but represent the natural numbers through unbounded extensibility. A plausible but (to our knowledge) novel variant is that a lognormal representation system could be extended to include large numbers, as needed, by the creation of new, lognormal units as needed, distributed on a logarithmic scale. While we cannot hope to represent the Mersenne primes this way, we might well deal with impressively gargantuan government deficits using the same log-normally distributed quantity representations that seem to be shared across many other species. To the degree that linear representation schemes and logarithmic internal resources are combined to induce roughly linear number behaviors for smaller numbers (Opfer & Siegler, 2003), a plausible hypothesis might extend this process to larger numbers, extending linear behavior out to any desired range by calibrating logarithmic representations to culturally provided forms.

Large Numbers through Categories of Smaller Ranges

Extending the boundaries of pre-existing numeric representations is not the only way to deal with larger numbers. It might instead be that when numbers exceed some endogenous or exogenous boundary, strategies for capturing magnitude shift become more complex than a more-or-less direct mapping from the relevant quantities into a spatial layout. For instance, one might subdivide the in-principle homogenous number line into segments or categories like ‘small’, or ‘large’, linearizing each sub-range of numbers (cf Laski & Siegler, 2007). Landy, Silbert, and Goldin (2013) found suggestive evidence for such mediated reasoning processes in the behavior of some college-students and adults recruited online. On a number-line placement task with boundaries of 1 thousand and 1 billion, about 40% of participants placed marks in a piecewise linear pattern of *overestimation*: the position of 1 million was very wrong (about 35% from the left edge of the page), but numbers between 1 million and 1 billion were placed

extremely linearly, as were numbers between 1 thousand and 1 million (see Figure 1). Note that on this task, also used in the current article, the correct linear location of 1 million is quite far to the left. Because there are 1 thousand millions in 1 billion, the location of 1 million lies about one thousandth of the way from one thousand to one billion. Our interpretation is that the overestimation pattern results from participants successfully applying their understanding of linearity over the two sub-ranges and simply adjoining the two line segments to yield a combined mapping[1], with the millions range slightly larger than the thousands[2]. More broadly, this strategy is reflective of 'reuse' style accounts of higher-level cognition, in which abstract capacities are achieved through coordinating disparate resources and strategic behavior rather than constructing special-purpose or domain-specific processing modules (Anderson, 2010). All this is not to

---

[1] Several minor points are worth noting: Nearly all participants in these populations can correctly model the relevant number words as numerals, and vice versa. In Landy et al., 2013, results were similar when *all* stimulus numbers were over 1 million, suggesting that these patterns are not a result of particular stimulus distributions. Furthermore, very similar results were obtained on a version of this task when the endpoints were 1 and 1 billion instead of 1 thousand and 1 billion (Guay, Davis, DeLaunay, Charlesworth, & Landy, under review)

[2] These responders were classed as *segmented linear* in Landy et al 2013; we will refer to the class as *overestimation* here, as we are interested in whether the "linear" response pattern also results from segmented or piecewise approaches. The term 'piecewise' seems more common than 'segmented' in the literature, and captures better the idea of separate linear pieces of the line.

suggest that responders necessarily had a firm notion of the relative magnitudes of the scale (see Rips, 2012 for some evidence that they do not), but rather that overestimation behavior results from making some decision about how to align separate above-1-million and below-1-million linear scales.

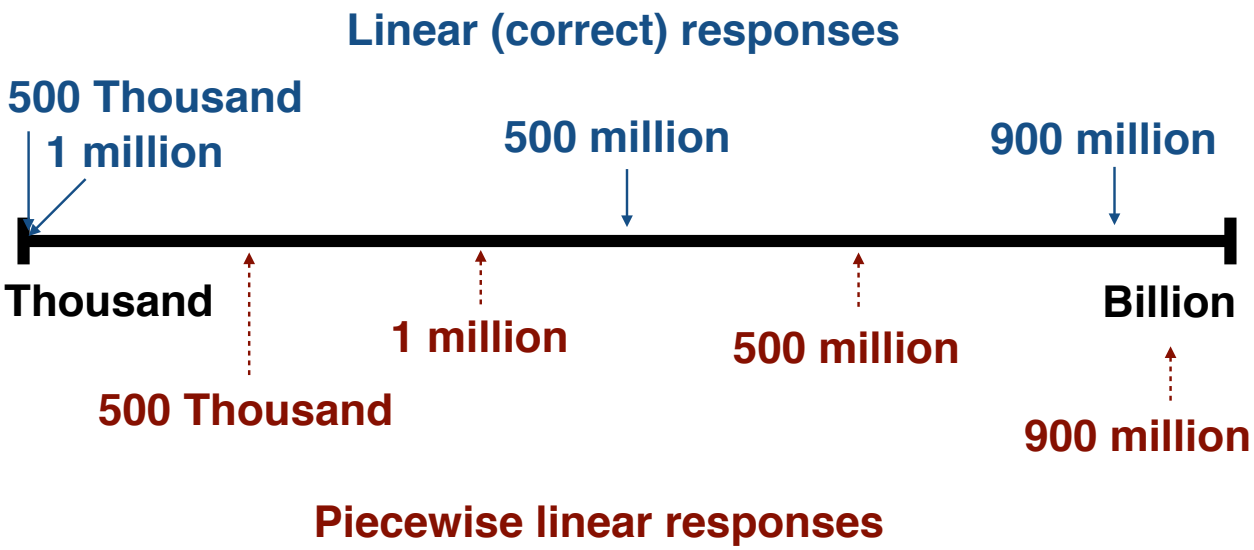


Figure 1: Illustration of typical response patterns from Landy, Silbert, & Goldin, 2013. On a line marked from 1 thousand to 1 billion, two groups of responses appeared. Approximately half of respondents gave roughly linear responses (shown above the line). In this strategy, 1 million is placed close to the extreme left-hand edge, while 500 million is placed at the midpoint. The other 40% of participants gave piecewise linear or "overestimation" responses, in which 500 thousand lies halfway between 1 thousand and 999 thousand, while 500 million lies halfway between 1 and 999 million. However, in this strategy, the location of 1 million is systematically biased toward the middle, resulting in substantial overestimation of all numbers, especially those near 1 million.

Scale Words as Categories that Bias Responses

A primary goal of the current paper is to understand how both linear and overestimation responses to large numbers may result from categorical responding. In this section, we cover two ways that categories have been shown to affect metric behavior in past research: assimilation to category prototypes, and highlighting of contrasts within categories.

Category adjustment theory is an approach often used in spatial memory that asserts that when categorical information is available and informative, it is often used to reduce error when reconstructing from memory (Huttenlocher, Hedges, & Duncan, 1991). In a typical task, a subject might observe the location of a dot, and after some delay be asked to place a dot in the same location as the original. To the degree that recall fails to precisely locate the dot, categorical information such as whether the dot was on the right or the left can be used to constrain its likely position by shifting memory estimates toward the distribution mass of the category. This reduces variance, at the cost of bias in the direction of the category. Resnick and colleagues (Resnick, Newcombe, & Shipley, under review; Resnick & Shipley, 2013; see also Laski & Siegler, 2007) have suggested that magnitudes contextualized as distances in space or time may be biased by available semantic categories. Combining this insight with the idea that scale terms such as thousands, millions and billions serve as organizing categories (Landy, Silbert, & Goldin, 2012), we here explore the possibility that the scale words themselves provide useful categories that bias number line placements in abstract tasks. It is important to emphasize that although prior work has *suggested* a role for category-based reasoning, no previous work has empirically demonstrated the role of category adjustment or categorical contrasts in large number estimation.

The pattern of results that appears in spatial memory tasks can be used to make specific predictions about the discontinuities likely to appear in a number line placement task under the reuse hypothesis, if we replace precision in memory with some facility at precisely transforming numbers into line positions: participants who have high-precision transformations are likely to exhibit low variability and low bias toward prototypes; those with lower-precision number transformations are likely to exhibit higher variability and more bias toward prototypes. Under the reasonable assumptions that (a) less linear-responding participants make lower-precision transformations and (b) the prototypical 'million' and 'thousand' items lie at the midpoint of their respective ranges, the category adjustment model then predicts that less linear responders will show a 'gap' or 'step' discontinuity at one million, and substantial inward bias away from the far left and right extremes (see Figure 2, third panel). When knowledge of the correct location is very good, however, prototype information will be used much less.

On the other hand, categories have also been shown to lead to the highlighting of contrastive information. If two elements fall into the same category, but are importantly different, that can lead to a highlighting of salient contrasts, and a biasing of perception toward extreme values (Hovland, Harvey, and Sherif, 1957; Coren & Enns, 1993). In particular, on perceptual tasks (Goldstone, 1995; Lochhead, 1988; King, 1988), or when memory delays are very short (Crawford, Landy, & Presson, 2014), spatial responses may be biased systematically away from category centers. This is a reasonable strategy if categories are well differentiated, and emphasizing within-category differences is deemed important. Consistent with this pattern, linear responders—who by hypothesis are likely to have relatively precise number estimations--are then less likely to show gaps, and more likely to show *overlap discontinuities* (Figure 2,

rightmost panel) and also to more closely approach the endpoints of the range. Finally, response times are often slow near a category boundary (Ashby & Maddox, 1989). Since both groups must make a categorical determination before placing a number into a range, reuse predicts slow response times in the vicinity of 1 million for both typical response patterns.

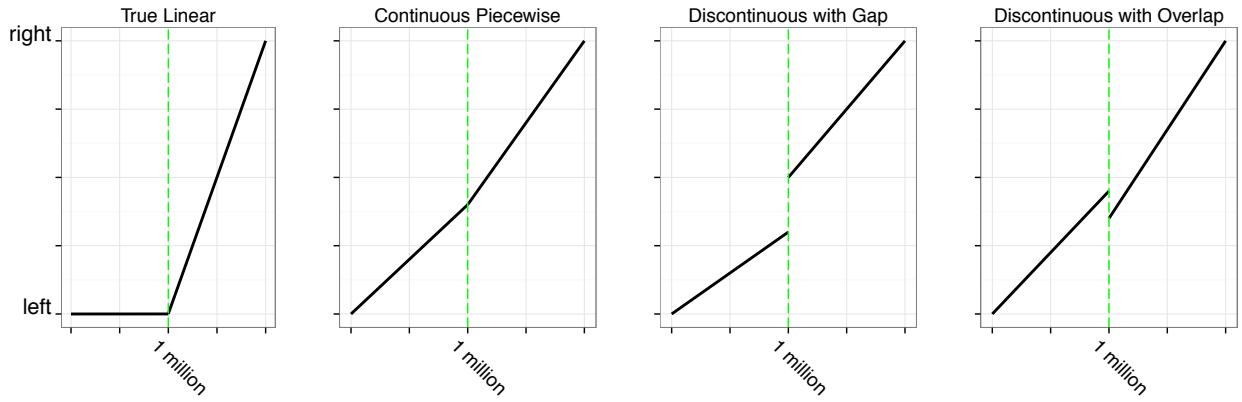


Figure 2: Possible response patterns on the number line placement task. Unlike the number line shown to participants, which showed 1 thousand and 1 billion without showing 1 million, and in which the normative location of 1 million was on the extreme left, the x-axis here is scaled to place 1 million at the center, with a linear scale on each side, which makes it easier for readers to see the visual detection of discontinuity at 1 million. The left panel indicates truly linear behavior. The next panel indicates continuous placement with 1 million shifted to the right of its normative position. The two right hand panels illustrate two possible discontinuities. The third panel illustrates a jump or gap discontinuity. The rightmost panel illustrates a non-monotonicity or overlap discontinuity, in which some numbers in the thousands would be placed to the right of some numbers over 1 million.

The idea that learners may gain by treating as categorical the recursive structure of the numbers has been explored as a method for teaching appropriate number line behavior. Thompson & Opfer (2010) demonstrated positive learning benefits when children were explicitly taught to treat words like "hundred" as categories, aligning them initially with words like "cherries". However, Young & Opfer (2011) analyzed the resulting data and, although students benefited from categorical training, there was no evidence for piecewise linear behavior by children on numbers under 100,000. To our knowledge numbers over 1 million have not been studied.

Response models

A mathematical formalization of the category bias model is presented in detail in the appendix, and summarized here. We assume that people first assign a mapping between regions of the line and appropriate categories. In the current case (in which the categories *thousands* and *millions* are relevant), this is accomplished by assigning a location to the boundary value, *one million*. When a new stimulus appears, a subject initially categorizes it, and then assigns it a position within its categorically assigned sub-range. The position is determined by combining information about the prototypical member of the category, and linear information about the magnitude in question, relative to the endpoints of the subrange. Prototype bias can manifest in either of two manners: to mitigate noisy linear information, resulting in adjustment *toward* the prototype (third panel of Figure 2), or to contrast with other salient members of the category, resulting in adjustment *away* from the prototype (fourth panel).

Response times have not received nearly as much attention as response positions in number line tasks, and expectations about response times are correspondingly poorly articulated

(though see Ashcraft & Moore, 2012). Here, we used a very simple model based on the same general assumptions as the model of response position. We assume response times come from three separate, sequential processes: non-decision based response time, determining which category the stimulus belongs in, and locating the stimulus value in the relevant line. We then fitted each person's response as following from an ex-Gaussian model (Heathcote, Popiel, & Mewhort, 1991), fitting non-decision response time as normally and decisional time as exponentially distributed. Non-decision based response time is assumed to be independent of the stimulus; locating the stimulus is treated as potentially linear in the stimulus magnitude, and the categorical decision difficulty is assumed to fall off with the psychological distance of the stimulus from the category boundary, which as usual is assumed to be 1 million (Ashby & Maddox, 1994).

Full details of both models can be seen in the appendix. The primary prediction of the category adjustment account is sharply increased response times near, but not at, the decision boundary of 1 million. These should be symmetric around the boundary point in psychological space, but since this psychological space is likely not to correspond to numerical space directly, this may be asymmetric around the value 1 million**.**

Words and numerals

A secondary goal of the current study was to explore in detail the integration of numerals and words. For small numbers (under about 1 million), there are two typical notational formats: numerals and number words. Larger numbers in the range of 10^6-10^{14} are often represented as numerals such as *500,000,000*, or hybrid notations such as *500 million.* The question of whether small numbers are represented in notation-invariant ways is a rich one (Campbell & Clark, 1988;

Campbell, 1994; Dehaene, Piazza, Pinel, & Cohen, 2003; McCloskey, 1992; Cohen-Kadosh, 2009), as is the connection between symbolic and non-symbolic forms (Lyons, Ansari, & Beilock, 2012); much less is known about the representation of larger numbers. Previous examinations have not had the power to detect possible differences between hybrid and numeral formats (Landy et al., 2013), or have not varied format (Resnick et al., 2013); one goal of the current study is to more closely compare different formats. To the degree that different notations are treated differently, it is quite plausible to predict that categorical effects would be stronger for hybrid numbers, since the difference between millions and thousands is arguably more salient for hybrid numbers, and since verbal labels may serve as especially strong inducers of categories (Lupyan, Rakison, & McClelland, 2007; Tanasescu, 2007).

Summary

Placements by adults on large number lines have previously been shown to follow two empirical patterns: overestimation, and roughly linear responding. The central theoretical question of this paper concerns the mechanisms that subserve these responses, especially for the proportion of participants who respond nearly linearly on number lines involving large numbers. How do they achieve this accuracy? There are two clear possibilities: one is that people extend a specialized linear number line constructed for smaller number ranges. The other possibility is that while people start this way on smaller ranges, they shift strategies as numbers become larger, and that linearity on these larger numbers is achieved, like overestimation behavior, achieved through the use of *multiple* lines: one for numbers (say) between 1 and 999 thousand, and one for numbers between 1 and 999 million (more precisely, the theory posits one line for each of the *scale words*, e.g., thousand, million, and billion, trillion, and so on, as well as one for

the number under 1,000). In this paper, we (1) explore which explanation best explains biases and boundary behavior, especially among linear responders, and also (2) provide more stringent tests of categorical responding among the overestimation group than have previous studies.

The two theoretical accounts—extension vs. categorization—can be discriminated by examining closely behavior near 1 million. If linear response patterns are achieved by extending small numbers lines performance should be very close to linear, and any strong deviations are likely to be symmetric and continuous, roughly fitting power-law or linear performance (Barth & Paladino, 2011; Opfer, Siegler, & Young, 2011). If linear responses, like overestimations, result from categorization, then although performance in the aggregate may be close to linear for some participants, both the linear and non-linear responders are likely to show evidence of 'joining' their lines: for instance, there might be discontinuities in placement behavior in the vicinity of 1 million. In order to apply the latter strategy discussed above, participants must first make a judgment about which line a particular item belongs on—whether the element is smaller or larger than 1 million. After this, they place the mark appropriately relative to the endpoints of the selected line segment. If so, linear responders might still have a discontinuity in the line at one million, even though they place 1 million at the correct location.

The categorization account makes several specific predictions, beyond the discontinuity at 1 million: (1) more linear participants will exhibit less response variability; (2) as a result, less linear responders will show a gap discontinuity at 1 million, and more linear responders may also show a discontinuity (but not necessarily a gap discontinuity); (3) because prototype adjustment affects responses at both extremes of the category, participants with larger discontinuities in the middle should also show a greater tendency to avoid outside edges. Finally, (4) the two step

process leads to a predicted response time peak at or perhaps slightly to the right of 1 million, since determining the correct category for numbers near the category boundary is predicted to be more difficult. All of these predictions follow directly from the application of category adjustment theory, using the response models developed in the appendix; none are obviously consequences of the extension hypothesis.

To explore these questions, an experiment was conducted in which college aged participants were instructed to place numbers linearly on a line marked from 1 thousand to 1 billion. We were particularly interested in the following questions: Will participants show evidence of scale reuse through prototype-consistent discontinuities at the subjective location of 1 million? Will these discontinuities appear even for otherwise highly linear individuals? If so, what is the shape of these discontinuities? And do these discontinuities lead to a large increase in response times near the value of 1 million?

Experiment

Method

Participants.

200 participants were recruited from Amazon's Mechanical Turk (*MTurk*). *MTurk* is an online marketplace in which participants volunteer to complete typically short online tasks in exchange for typically slight compensation. Participants recruited from *MTurk* have been found to behave similarly to other participants on a range of cognitive tasks when experiments are carefully conducted (Crump, McDonnell, & Gureckis, 2013), and have been extensively used as subject populations in prior work. In this case, data collected on *MTurk* were informally

compared to data collected with live participants in the lab; no differences in patterns were observed. This task took about 20 minutes to complete.

Procedure.

Instructions showed an image of a small line (labeled from 0-8), and indicated the placement of '6' as a sample. Participants were informed that the endpoints would be larger than in the sample, but to continue placing the numbers in the same way.

Participants were then shown a number line with "1 thousand" under the left end, and "1 billion" under the other ("1,000" and "1,000,000,000" in the numeral condition). Because the study took place online, the physical length of the stimulus line cannot be determined. Participants were sequentially presented numbers in a random order, and selected with the mouse their chosen location for each number. Participants made 182 number line placements. Stimulus numbers were selected to sample the ranges under 1 million and over 1 million roughly evenly. Twenty numbers under 1 million, and twenty numbers above it, were chosen to be integers with one or two significant non-zero digits, and to be close to uniformly spaced within the two subranges. Because the numbers in the vicinity of 1 million were of particular interest, the range just over 1 million was over-sampled: the exact numbers "1 million", "2 million", "3 million", and "4 million" were also included. Landy et al., (2013) found little shift in participant behavior in response to adjustments of the range of stimuli used; the same was expected here.

The experiment started with 10 warm-up trials, with distinct stimuli in the same range as the test stimuli. Each test stimulus was estimated 4 times by each participant; judgments were untimed and separated into blocks of 43 unique stimuli, each presented in a random order. The

stimuli presented in the test phase are presented in Table 1. Because the effect of number representation is of interest here, format was manipulated between participants: 100 participants received *numerals*, such as "54,000,000"; the other 100 received *hybrid notation* stimuli, as in "54 million". The two stimulus types essentially serve as independent samples to validate conclusions. However, there were no noticeable differences between formats, so they are collapsed here.

Table 1. Stimuli used in the Experiment

Number Range	Stimuli Used
Thousands	10, 60, 100, 150, 230, 250, 310, 380, 420, 480, 500, 580, 640, 680, 720, 780, 840, 890, 940, 950
Millions	1, 2, 3, 4, 60, 100, 150, 230, 250, 310, 380, 420, 480, 500, 580, 640, 680, 720, 780, 840, 890, 940, 950

Analysis and Results

Figure 3 presents individual participant responses. Responses were analyzed in two steps. First, large "order of magnitude" errors were culled. Second, a hierarchical Bayesian model was fitted to the data, assuming (as in Landy et al., 2013) two populations of participants, differing in where they placed 1 million. We use high-density intervals of the population posteriors to capture the patterns in the data.

Culling of Order of Magnitude Errors.

Several responses were highly compatible with the idea that the participant incorrectly encoded the stimulus by two or three orders of magnitude, e.g., reading “600 thousand” instead of “600 million.” A two-step process was used to prune these data: first a continuous piecewise linear model was fitted to the data for each subject. Each response was removed from analysis if the distance from the predicted position for an order of magnitude error was less than 25% of the distance of the predicted position for the actual stimulus (3% of responses fit this criterion; see Figure 3). Then, the pruned data were submitted to the MCMC model-fitting process. This cleaning process made the results more precise, but affected none of the conclusions reached here compared to analysis of the full data set.

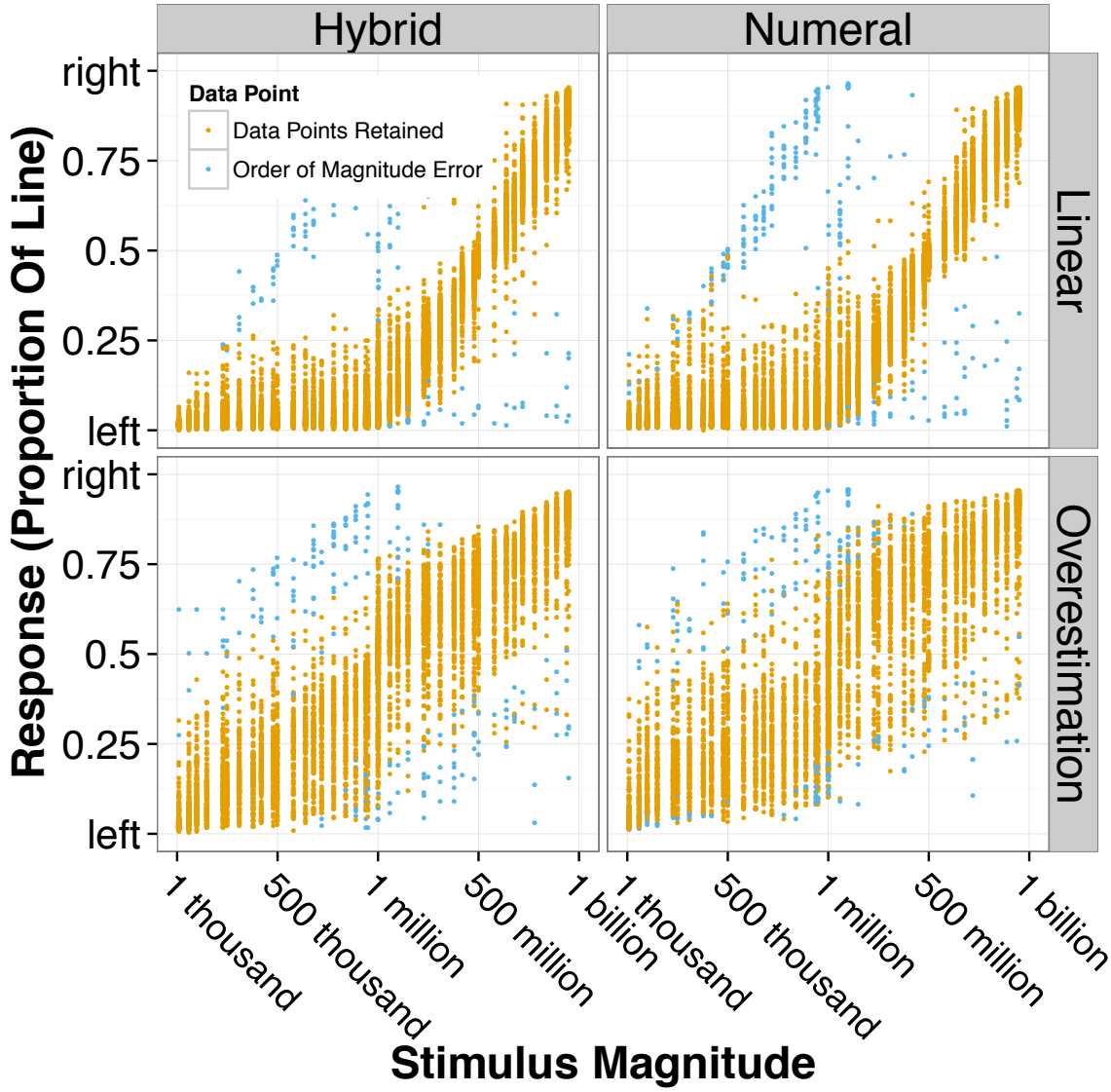


Figure 3: Raw participant responses, divided into the hybrid and numeral conditions (columns), and participants fitting the linear and overestimation response patterns (rows) Responses marked in blue were coded as order of magnitude errors. Responses marked in yellow were retained in the reported analysis. Note that in this figure, the x-axis is not normative: the numbers under 1 million have been 'stretched' to take up the same space as the numbers over 1 million. This makes the patterns clearer and easier to see. Remember that while in the figure, 1

million is placed in the center halfway between 1 thousand and 1 billion, no representation like this was shown to participants.

Model Fitting Procedure.

The data were fit using a hierarchical Bayesian procedure. Based on the results of Landy et al., 2013, two populations were expected: one group who placed responses very nearly linearly, and one group who showed a strong overestimation pattern. Thus, the model included two populations of respondents: one population was predicted to be 'close to linear'—that is, to place 1 million near to its normative location. The other population was predicted to strongly overestimate, with a million point near 0.4 or so. We also allowed each group to differ at the population level in degree of prototype-bias: two population-level prototype groups were constructed, each assigned with vague population-priors centered on 0. One set of priors was used across these populations for the remaining parameters. Table 2 presents the full set of variables; Table 3 presents the prior and highest posterior densities of the population-level parameters for the two populations.

Although Bayesian analyses allow mixing across strategies at the item level, for simplicity we fitted each participant as making all their judgments based on a single model, with a hyper-parameter governing the distribution of probabilities. As can be seen in Figure 4, the resulting probabilities were heavily bimodal, with approximately 30% of participants having more than an 0.8 estimated probability of being produced by the overestimation model, about 63% a less than 0.2 probability (and therefore probably produced by the linear model), and just 7% of participants' response patterns falling between these two extremes.

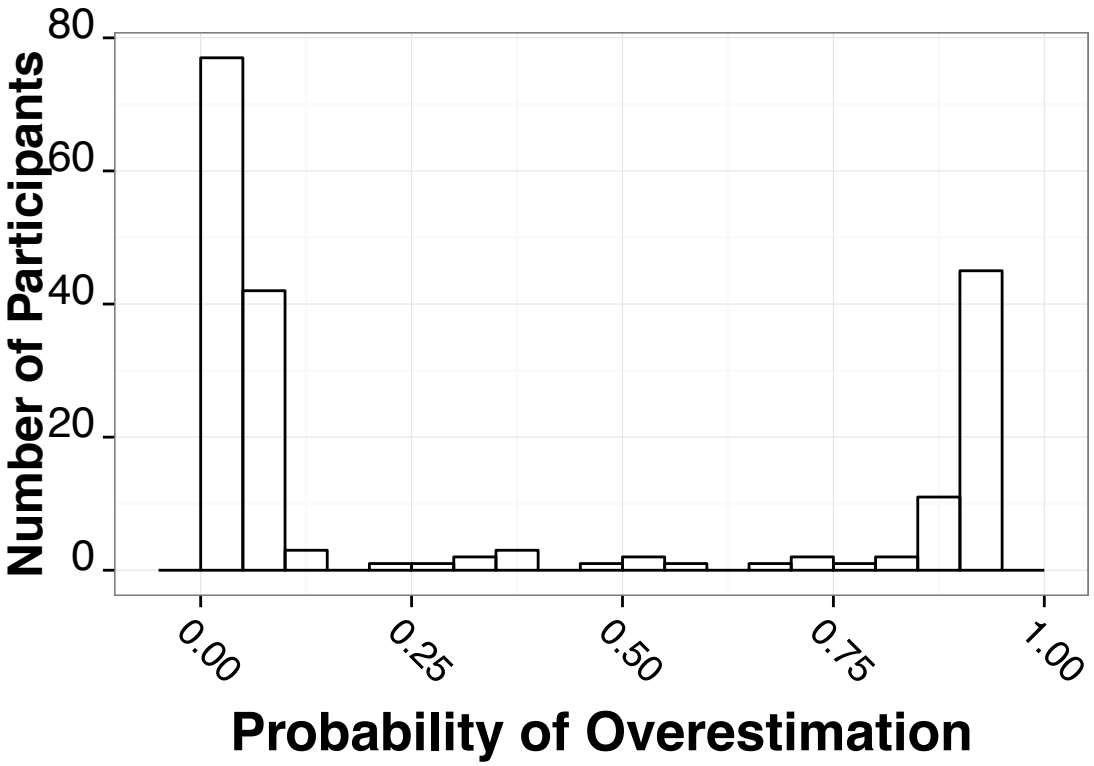


Figure 4: Histogram of participants' estimated probability of being produced by the overestimation (vs. linear) model.

The model was fitted using Stan (Carpenter et al.), via the rstan module (R Development Core Team, 2008). The model is available at dlandy.psych.indiana.edu/CuttingInLine/. The model was run on 6 chains with 2000 iterations per chain, and a thinning of 1. High R-hat values may suggest a failure of chain mixing. In this case, all R-hat values were less than 1.1. Mean responses of the fitted groups are presented in Figure 5, along with model residuals.

Evaluations of Model Predictions

Finally, we turn to the direct test of the predictions of the model. We start with predictions 1, 2, and 3: the relationships among overestimation, response variability, and gap discontinuities.

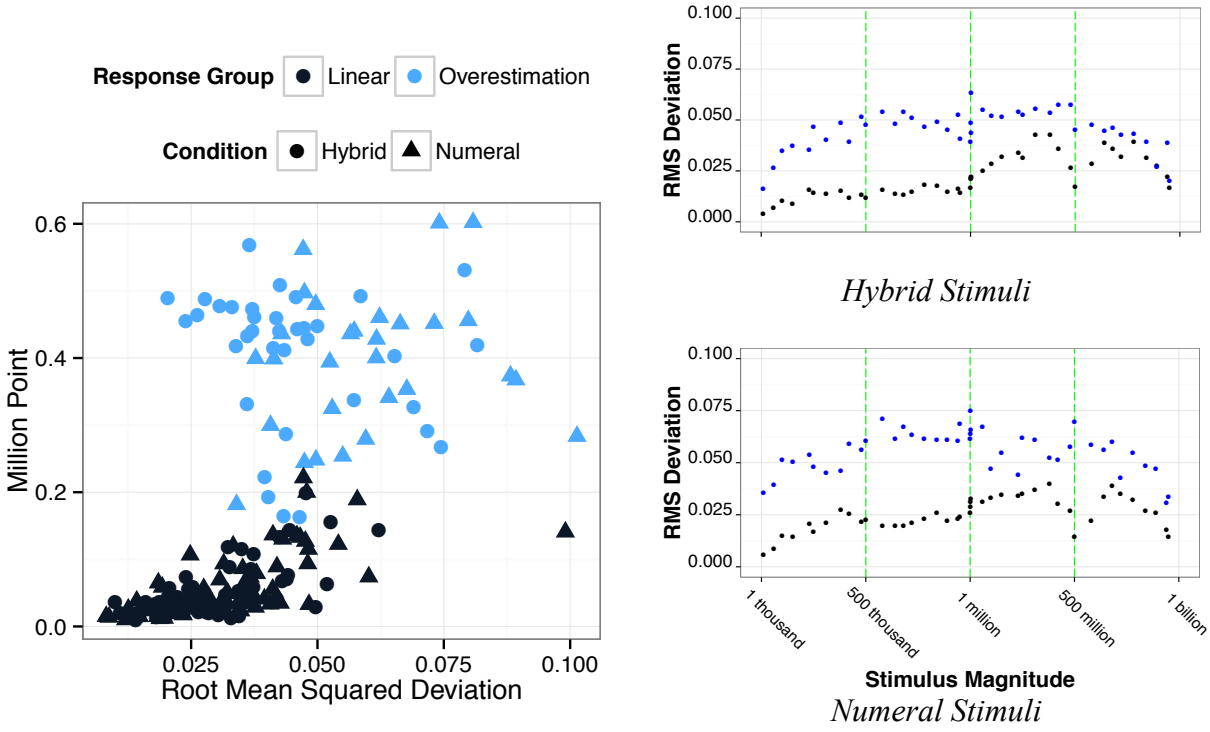
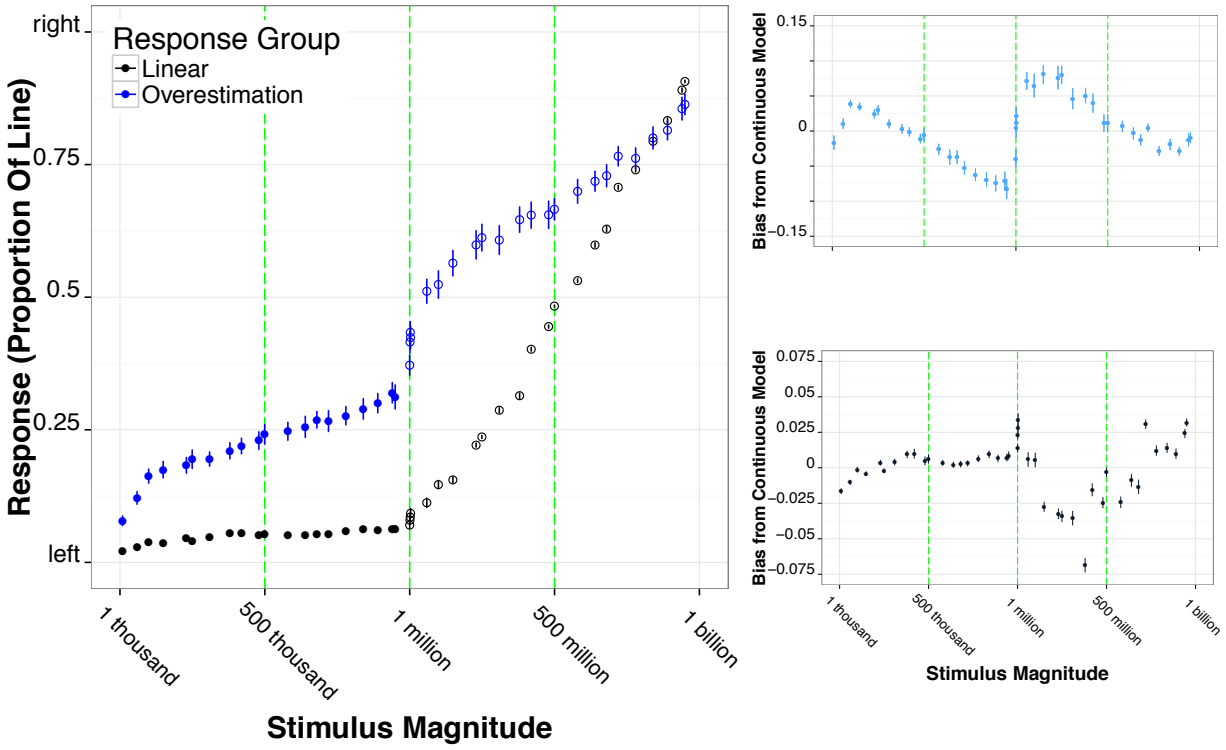


Figure 5: Relationship between estimated response variability and overestimation of 1 million. The left panel plots the mean estimated variability from the model fits against the estimated location of 1 million. The right panels show, for each stimulus set, the empirical deviation from the mean predicted response.

Prediction 1 states that because overestimates result from low quality number-line mapping processes, response variability should be relatively high among people who strongly overestimate 1 million. In line with this idea, the mean estimated posterior of standard deviation

of the response (from the best-fitting model) for the overestimating group was significantly higher (M(sd)=0.15; HDI(sd)=[.107, .200]) than that of the linear group (M(sd)=0.086; HDI(sd)=[.068, .105]).

As can be seen in the right panels of Figure 5, overestimating participants were more variable in their responses across the stimulus and response ranges, not just for a limited range of values. This suggests that these patterns result from intrinsic uncertainty, rather than differences in how the two groups placements relate to the line edge.

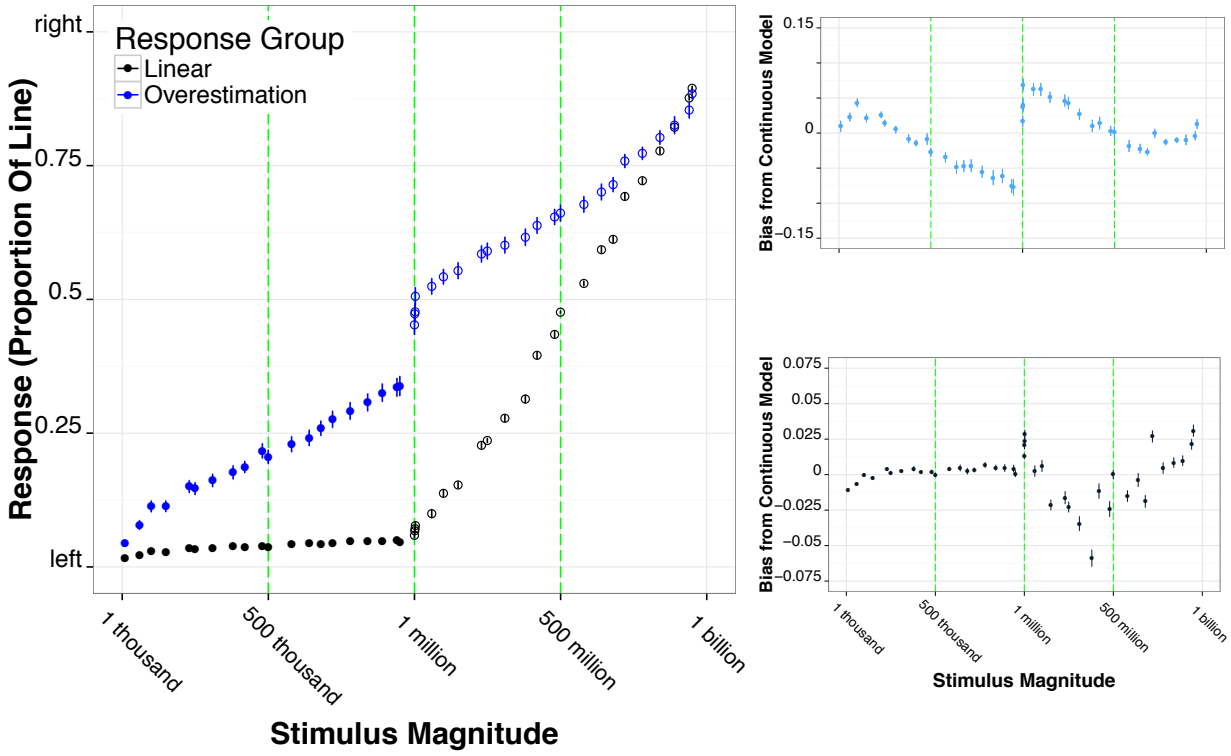


Numeral Stimulus Group Responses

Continuous model responses

Figure 6: Mean participant responses, divided into the linear (lower line) and overestimation (upper line) groups, for those participants shown numerals. The right panels plot

the residuals for each group from the best-fitting continuous model (left two panels of Figure 2). Error bars represent standard error of the mean estimate for each point.



Hybrid Stimulus Group Responses

Continuous model responses

Figure 7: Mean participant responses, divided into the linear (lower line) and overestimation (upper line) groups, for those participants shown hybrid numbers. The right panels plot the residuals for each group from the best-fitting continuous model (left two panels of Figure 2). Error bars represent standard error of the mean estimate for each point.

Of primary interest are the location estimates for 1 million and the degree of 'prototype use' fitted by the model. In line with prediction 2, participants with larger estimates of 1 million also showed large gaps around 1 million. Figures 6 and 7 show the mean estimates of each

group, in which the gap around 1 million is clearly visible in the overestimating group for both participants shown numerals and those shown hybrid stimuli; Figure 8 plots the posterior estimate for each individual. Both groups showed significant bias away from strict linearity in their prototype use. As predicted by category adjustment, the overestimation group showed substantial adjustment toward the 'millions' and 'thousands' prototypes, in that estimated bias of the population was significantly greater than 0 (M(*b*)=0.24; 95%HDI = [0.21, 0.29]). While it is more difficult to directly observe, the highly linear group showed a small but statistically highly significant overlap-style discontinuity—similar to the rightmost panel of Figure 2 (M(*b*)=0.98; 95% HDI=[-0.04,-0.01]), and visible in the residuals of the continuous model fit. This is a significant shift in the direction *away* from the prototype, and toward the edges of the screen. In both the linear and the overestimation group, the region very close to 1 million was not well fit by the prototype model, a point to which we will return in discussing response times.

Relationship between gap discontinuity and edge avoidance

As can be seen in Figures 6 and 7, participants in the overestimation group showed both larger gap discontinuities and more endpoint avoidance. This is predicted by the assumption that both phenomena are, at least in part, produced by prototype bias (prediction 3). This is because on the prototype bias account, responses that are farthest from the prototypes (here assumed to be the center of the *thousands* and the center of the *millions* categories) are most biased by the categorical information. Gaps at the location of 1 million result as people bias extreme values such as 900 thousand and 2 million in toward the center of the *thousands* and *millions* categories. Edge avoidance results from the same process with stimuli at the other extreme, such as 10 thousand and 920 million. The extreme values (1 thousand and 999 million) are biased

symmetrically with the 'central' values (999 thousand and 1 million), and so gaps at the center are expected to correlate with avoidance of the edges.

However, this prediction cannot be evaluated by looking at high-density regions of the posterior (Kruschke, 2010). This is because the model used above is constrained to show the phenomena of interest: the model automatically avoids edges to the degree it shows a gap discontinuity! For simplicity this prediction was evaluated separately using a very simple estimation procedure, in which two straight lines were fitted for each participant, one to each

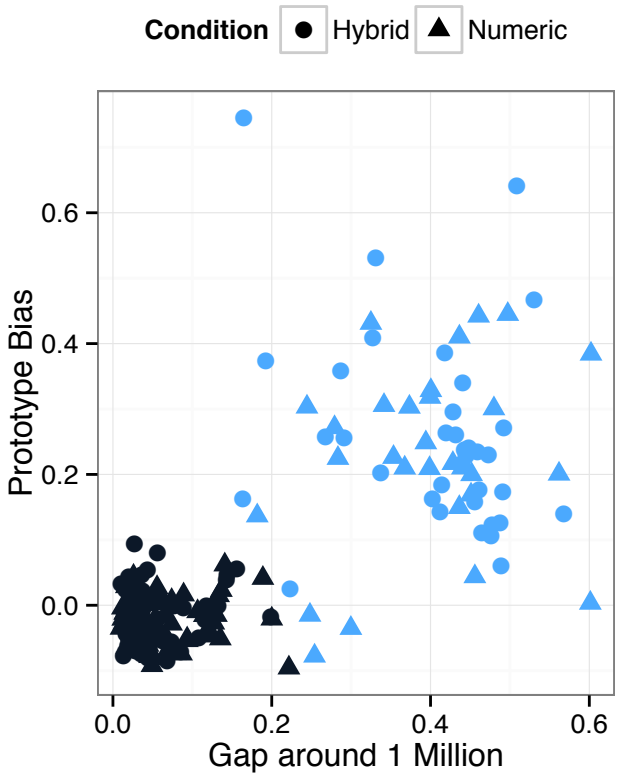


Figure 8: Prototype bias vs. mean estimated location of 1 million, for each individual. Each dot indicates the mean posterior fitted for the individual from the MCMC fit. The result

shows a group differentiation, with a few border cases. The linear group shows slight but robust expansion of the prototype range; the overestimation group contracts toward the prototype.

range: each line was parameterized by its endpoints (i.e., the location of 1 thousand and 1 million for the first line, and the location of 1 million and 1 billion for the second line). . These lines were fitted using a maximum likelihood technique. The gap indicated by the interior ends of the line was correlated with the gap indicated by the exterior ends (see Figure 9). The distance from the actual edges of the line of the predicted linear judgments at 1 thousand and 1 billion was compared to the distance between the two interior points. Indeed, there was a relationship between leaving a gap at 1 million and avoiding the edges ($r=0.32$, $t(193)=4.7$, $p<0.0001$).

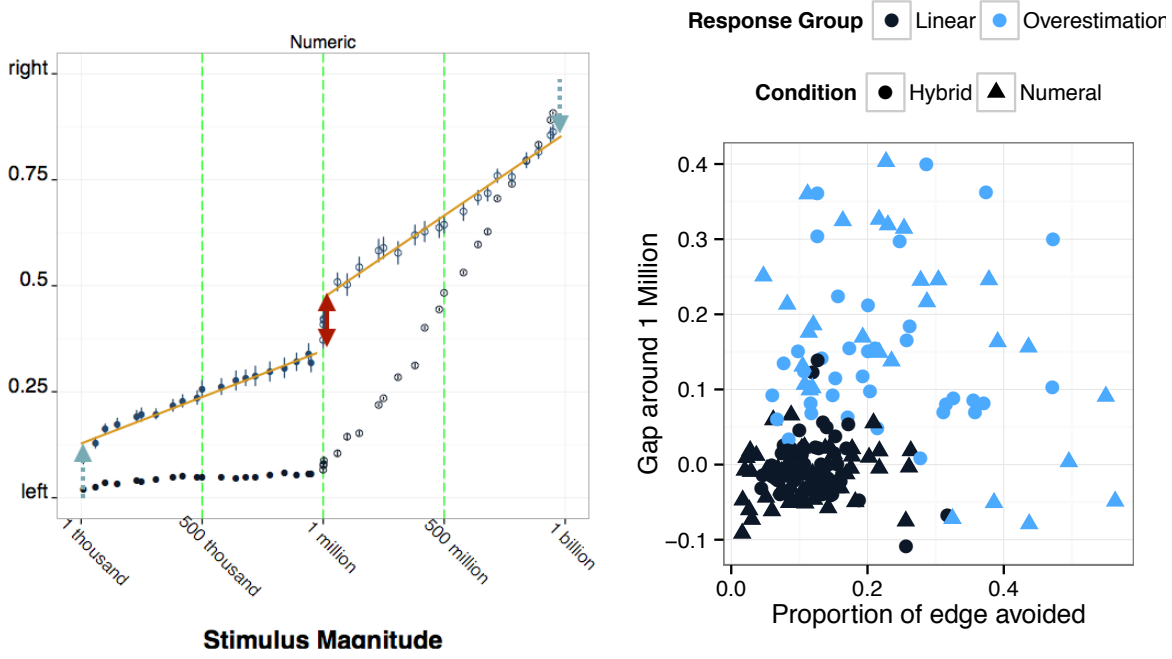


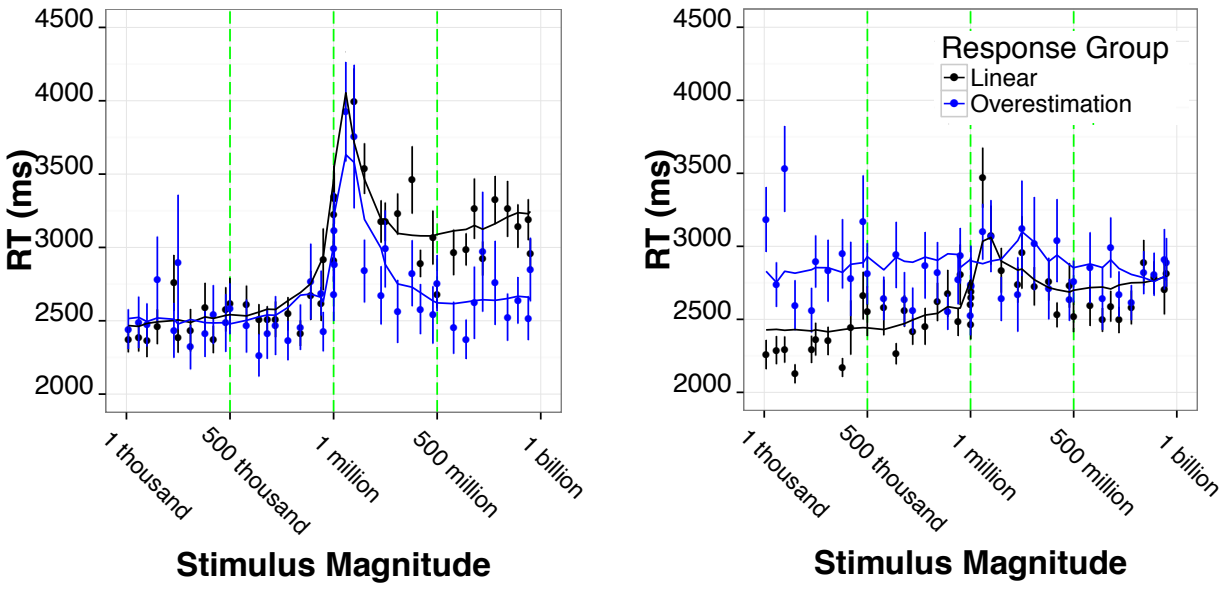
Figure 9: The left panel indicates the method used to estimate "edge avoidance". First, two lines were fitted, one to stimuli above and one to stimuli below 1 million (yellow solid lines; fitted for each group). Then, the difference between the actual edges, and the minimal and maximal predictions (exterior dotted arrows) was compared with the difference between the two predictions for 1 million (central solid arrow). The right panel shows the results, indicating a significantly higher edge avoidance proportion when the gap in the middle was larger.

Response Time Analysis and Results

A very small number of response times (0.08%) were longer than 50 seconds, and were removed. The remaining response times were fitted using the parameterized ex-Gaussian model, discussed above and in the appendix. The category adjustment account predicted that near the 1 million boundary, a response decision would cause a strong slowing of response times, to the degree that people are making categorical distinctions between numbers above and below 1 million (prediction 4). Response times indeed showed a large increase near 1 million for three of the four groups (see Figure 10): only the participants who overestimated when shown hybrid stimuli did not show a strong increase in response time in the vicinity of 1 million.

Because response times are not as well understood as response structures, we consider any definite conclusions preliminary. However, to give some estimate of robustness, we conducted bootstrap linear models predicting each of the major variables by condition and response behavior (linear or overestimate). This analysis of the model parameters suggested a statistically robust categorization cost maximum near 1 million ($M(b_2)=13100$; 95% CI = [9200, 15100], $p<0.001$); the total size of this peak did not vary across condition significantly (all p's > 0.5), but the analysis did suggest a condition by response pattern effect in the distance scaling

parameter (α), such that the peak was more diffuse or absent for people in the hybrid condition (β=-0.02, p~0.01), for the overestimation group (β=-0.03, p<0.001), but these two effects were subadditive in the interaction (β=0.04, p~0.02). A fuller breakdown of the patterns in response time variables is presented in the appendix.



Numeral Stimuli *Hybrid Stimuli*

Figure 10: Mean response times across participants for each of the two groups of subjects. The lighter (blue) lines and dots represent the overestimation group, while the darker (black) lines represent the linear group.

Discussion

We often speak as though *natural number* is a singular concept, and as though the processing of aligning number names with implicit magnitude and individuation representations

gives us access to the entire structure. Here we have argued that not only does such an alignment typically appear only as a result of a lengthy developmental process (though often quickly for a given range of numbers; Opfer & Siegler, 2007; Siegler, Thompson, & Opfer, 2009), it is never fully completed. Larger numbers whose magnitude can be successfully mapped onto a line are not mapped through a process of systematically integrating into a common linear scale. Instead, it seems that both linear and non-linear responders on the task share a common approach consisting of dividing the scale into culturally given regions, and applying linear responses over those regions. These subscales must then be coordinated with each other to approximate a single line. More broadly, studies of mathematical thinking have tended to assume that the cognitive constituents of number varieties match the mathematical constituents. We suggest that, for the naturals at least, representation and processing of large numbers do not have a coherent, uniform structure, but a hierarchical one, at least for numbers into the range of 10^{12} or so.

To our knowledge, this is the first paper to empirically demonstrate the hypothesis that participants adjust their estimation in response to scale categories. Several novel observations support the interpretation that participants divide the unified numerical scale into separate "thousands" and "millions" categories, with linear scales within each category. The primary piece of evidence is the strong discontinuities apparent around 1 million: the 'gap' discontinuity exhibited by the overestimation group, and the 'overlap' discontinuity exhibited by the "linear" group. This interpretation is supported by two novel supplementary predictions regarding data not usually explored in low-number line production task: response variability and 'edge avoidance'. In the current data set, all three of these values were positively correlated, as predicted by the category adjustment account. Crucially, this was true even in the "linear" group,

suggesting that despite their close approximation to linear responding (and the label we gave them), this group is also using the categorical structure of numeral representations as a guide to production.

Previous work (including Landy, Silbert, and Goldin 2013) treated linear responses as non-categorical; the current work demonstrates that these participants also treat 1 million as a special categorical boundary. The special role of 1 million for the linear group is further supported by the response time patterns, which show a large peak near 1 million. Across strategies people seem to pick out 1 million as a special reference location, forming an interesting comparison with previous studies finding that people have difficulty imposing a category boundary unsupported by external structure (Simmering & Spencer, 2007).

An interesting characteristic of the task is that not only did more accurate responders weight the categorical information less strongly than the item-specific information, at the limit the accurate responders biased slightly away from the category line. We do not have a strong theoretical account for why this response structure should obtain, and it may be a result of particular demands on the task (e.g., to differentiate marks which are objectively in indistinguishable locations); however, it is consistent with other findings in the literature, which have found category contrast effects in cases where knowledge is high due to the perceptual nature of the task (Goldstone, 1995) or a short memory delay (Crawford, Landy, and Presson, 2014). Interpretation of this pattern must be taken with caution at this point, but one interpretation is that this stretching allows increased discrimination of within-category items. This may be partially a response to task demands, which suggest that very proximal stimuli (such *100 thousand* and *700 thousand*) should be differentiated.

Response time patterns have not been extensively studied on the number line task; no previous study has examined possible categorical effects on response time, nor has any study to our knowledge examined the response time patterns in large numbers. Three of four groups showed a peak near 60 million. This peak was even more pronounced for the linear than the overestimation group, but was in roughly the same location across participants, particularly when viewing numeral stimuli. We speculate that this peak might reflect three factors: a category decision boundary at 1 million, combined with a probability of categorizing the stimulus at all, which decreases sharply near the decision boundary, and an asymmetry where 1 million is 'in' the million range (and as a result, the psychological distance between *900 thousand* and *1 million* is potentially larger than that between *1 million* and *2 million*). Although these are all well motivated theoretical assumptions, they are just that, and must be explored in future research.

The number line estimation task has recently experienced increased interest and scrutiny due to both novel interpretations of the mechanisms subserving linear and logarithmic behavior in children (Chesney & Matthews, 2013, Moeller et al., 2009) and its clear importance in predicting future mathematical success (Siegler, 2009; Thompson & Opfer, 2010). The data presented here do not speak directly to the mechanisms involved in line estimation in young children over relatively small ranges, but the theoretical perspective relates to positions taken in that domain. In particular, the category adjustment account presented here is entirely compatible with the suggestion that some number line behaviors result from category-based reasoning processes. Thompson & Opfer (2010) found that participants learned linear behavior over numbers in the tens of thousands by aligning them with numbers in the range from 1-100. Participants were specifically taught to treat scale words as categories by swapping them with

nouns (e.g., “15 thousand is like 15 cherries”); in this case, the authors interpret this in terms of an analogy that allows the extension of linear resources, rather than an online processing strategy. Even more directly related to this study is Laski & Siegler (2007), who find that young children’s qualitative categorization of numbers into categories such as “small”, “medium” and “large” improves over time, and relates to the linearity of their line placements. Indeed, Laski & Siegler suggest that young children use divisions of the number line into qualitative categories to ‘divide and conquer’ the number line; although they do not produce a response model, our approach is entirely consonant with the perspective they articulate. Cicchini, Anobile, & Burr (2014) noted that in many cases compressive (“logarithmic”) responding can be explained as adjustment toward the previous set of stimulus items; though the mechanism is different from the category adjustment approach used here, both assume that bias-variance tradeoffs significantly impact number line productions. In the context of large numbers, Resnick et al. (2013) suggested that scientifically and experientially meaningful categories are used when responding to large numbers contextualized as distances or time. In Resnick’s proposal the relevant categories were, for instance, geological periods; we find that in the absence of such contextually given categories, the number system itself suffices to provide relevant categorical structures. The prominent role that numbers across wide scales play in math and science education (Newcombe, 2013) and political decision-making (Guay et al., under review A) together with promising evidence that relatively short training can impact understanding (Guay et al., under review B; Landy, Silbert, and Goldin, 2012; Resnick et al., 2012) make understanding the structure and malleability of numerical processing an important practical challenge facing the learning sciences.

The task participants were asked to perform was unreasonable—putting half the marks within a pixel of the left-hand end of the line. It may be tempting to dismiss the observed patterns as 'task demands'. Such an explanation would overlook the nature of the experimental situation, and the nature of scaling. In the case of large-scale comparisons, no particular task is ever 'reasonable', and yet comparisons are often important in practical situations (e.g., balancing national budgets by cutting particular, often quite small, programs). Furthermore, in any experimental context, participants are asked to engage in particular, often unusual behaviors. The ways people grapple with task requirements are informative about the resources available to them (Stenning & Van Lambalgen, 2008). Particularly in the case of large numbers, for which a unified 'number sense' is less likely, a description of the constraints and processes that govern the construction of task-specific responses constitutes a good theory of numerical understanding. In this case, there are many ways to respond to this 'unreasonable' task, but a systematic pattern is evident in how people actually accomplish it: people appear to construct "small" linear ranges of around 3 orders of magnitude; beyond that, people make use of culturally available and visually salient reference points. Of course, numerical magnitudes do not stop at trillion: although roughly the response strategy we suggest may work for a few orders of magnitude, it too must be altered as numbers increase beyond a few orders of magnitude (Rips, 2012). That said, it may be the case that a different distribution of numbers would impact the behavior of, particularly, linear responders on this task; previous research suggests that participants show the overestimation pattern even with quite different distributions, or after the first trial (Landy et al, 2013; Guay et al, under review A).

A natural limit point of the task-specific approach is that interactions with task materials constitute cognitive and perceptual functions involving abstractions (Zhang & Norman, 1995; Landy & Goldstone, 2005; Landy, Allen, & Zednik, 2014; Clark, 1998, 2006). From this perspective, it is perhaps surprising that so few differences are seen between numeral and hybrid versions of the task. One might expect that—because hybrid notations render much more salient the distinction between, say 5 million (5,000,000) and 500 thousand (500,000)—processing approaches would be quite different. That they are so similar may indicate that the processing of these distinct forms is closely linked, or involves internal conversion into a common form (say, by mentally 'reading' the numeral as hybrid words; see Dehaene, Molko, Cohen, & Wilson, 2004; McCloskey, 1992), or simply that the structures are analogous enough to support similar processing strategies. For instance, the placement of the commas may be used in a way analogous to the scale words when dealing with numerals.

Even for natural numbers just barely beyond the range of common experience, rather than directly extending core conceptual tools, people engage in processes of construction: they coopt existing perceptual-cognitive systems (Anderson, 2014; Goldstone, Landy, & Son, 2010) that work well to form linear mappings of smaller-number ranges (not accurate numerosity counts), and compose and iterate them to create new perceptual-cognitive systems. New complex structures are built by coordinating older systems—by aligning preexisting perceptual and cognitive resources so that the joint activity of these systems mediates the normative behavior; in short, new systems are routines built of components that include older systems (Anderson, 2010; Landy & Goldstone, 2005). We have found that 1 million is a location for a discontinuity—it might have been the case that familiarity or psychophysical factors created a boundary in

strategy at any arbitrary number (Ebersbach et al., 2008). The observed pattern suggests that people use the culturally provided numeral system to select appropriate magnitudes at which to recycle cognitive resources. In this case, well-designed cultural-cognitive systems function as metaphorical telescopes: just as telescopes extend the bounds of perception at the cost of some distortion, large-number representations extend the natural bounds of numerical perceptions by connecting them to new contents—at the cost of somewhat distorting those contents.

The cardinalities of sets of objects (the major experiential basis for the natural numbers) have amazing properties that derive entirely from the successorship function. It appears, however, that human representations of natural numbers, at least beyond a paltry few hundred thousand iterations, rely on resources quite distinct from successorship or even a metric "number line", and impose substantial additional structure given by the numeral systems. A fundamental mistake made by classical empiricism was to assume that the inner representations were iconic—that they were *like* the outer represented. When reasoning about large numbers, we appear to rely on representations that are fundamentally unlike the numbers themselves.

References

Ashcraft, M. H., & Moore, A. M. (2012). Cognitive processes of numerical estimation in children. *Journal of experimental child psychology*, *111*(2), 246-267.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, *33*(04), 245-266.

Anderson, M. L. (2014). After Phrenology: Neural reuse and the interactive brain

. MIT Press.

Ashby, F.G. & Townsend, J.T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154-179.

Ashby, F. G. & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology 38,* 423-466.

Barth, H.C., & Paladino, A.M. (2011). The development of numerical estimation: evidence against a representational shift. *Developmental Science, 14,* 125–135.

Barth, H., Lesser, E., Taggart, J., & Slusser, E. (2014). Spatial estimation: a non-Bayesian alternative. *Developmental science*.

Campbell, J. I. (1994). Architectures for numerical cognition. *Cognition*, *53*(1), 1-44.

Campbell, J. I., & Clark, J. M. (1988). An encoding-complex view of cognitive number processing: Comment on McCloskey, Sokol, and Goodman (1986). *Journal of Experimental Psychology: General 117*(2) 204-214.

Carpenter, B., Lee, D., Brubaker, M. A., Riddell, A., Gelman, A., Goodrich, B., Guo, J., Hoffman, M., Betancourt, M., & Li, P. *Stan: A Probabilistic Programming Language.*

Chesney, D. L., & Matthews, P. G. (2013). Knowledge on the line: Manipulating beliefs about the magnitudes of symbolic numbers affects the linearity of line estimation tasks. *Psychonomic bulletin & review*, *20*(6), 1146-1153.

Cicchini, G. M., Anobile, G., & Burr, D. C. (2014). Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences*, *111*(21), 7867-7872.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends Cogn. Sci*. 10, 370–374. doi: 10.1016/j.tics.2006.06.012

Clark, A. (1998). "Magic words: how language augments human computation," in *Language and Thought: Interdisciplinary Themes*, eds P. Carruthers and J. Boucher (Cambridge: Cambridge University Press), 162–183.

Cohen, D. J. (2009). Integers do not automatically activate their quantity representation. *Psychonomic bulletin & review*, *16*(2), 332-336.

Cohen, D. J., & Blanc-Goldhammer, D. (2011). Numerical bias in bounded and unbounded number line tasks. *Psychonomic bulletin & review*, *18*(2), 331-338

Cohen, D. J., Ferrell, J. M., & Johnson, N. (2002). What very small numbers mean. *Journal of Experimental Psychology: General, 131* (3), 424–442.

Cohen Kadosh, R. (2009). Numerical representation in the parietal lobes: abstract or not abstract? *Behav. Brain Sci*. 32, 313–328. doi: 10.1017/S0140525X09990938

Coren, S, & Enns, J. T. (1993). Size contrast as a function of conceptual similarity between test and inducers. *Perception & Psychophysics 54*(5), 579-588.. http://dx.doi.org/10.3758/BF03211782

Crawford, L.E., Landy, D., & Presson, A.N. (2014). Bias in spatial memory: prototypes or relational categories? Poster presented at the 36th Annual Conference of the Cognitive Science Society. Quebec City, Quebec.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410.

Dehaene, S., & Brannon, E. (Eds.). (2011). *Space, time and number in the brain: Searching for the foundations of mathematical thought*. Academic Press.

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science, 320,* 1217-20.

Dehaene, S., Molko, N., Cohen, L., and Wilson, A. J. (2004). Arithmetic and the brain. *Curr. Opin. Neurobiol*. 14, 218–224. doi: 10.1016/j.conb.2004.03.008

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive neuropsychology*, *20*(3-6), 487-506.

Ebersbach, M., Luwel, K., Frick, A., Onghena, P., & Verschaffel, L. (2008). The relationship between the shape of the mental number line and familiarity with numbers in 5-to 9-year old children: Evidence for a segmented linear model. *Journal of experimental child psychology*, *99*(1), 1-17.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*(7), 307-314.

Gelman, R. (2011) The case of continuity. *Behavioral and Brain Sciences, 34*(3), 127-128.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8-38.

Goldstone, R. L. (1995). Effects of categorization on color perception. *Psychological Science*, 298-304.

Guay, B.M., Davis, Z.J., DeLaunay, M., Charlesworth, A., & Landy, D. (Under Review A). Number Comprehension Impacts Political Judgments.

Guay, B., Chandler, C., Erkulwater, J., & Landy, D. (Under review B). Testing the Effectiveness of a Number-based Classroom Exercise.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin, 109*(2), 340-347.

Hovland, C. I., Harvey, O. J., & Sherif, M. (1957). Assimilation and contrast effects in reactions to communication and attitude change. *Journal of Abnormal and Social Psychology, 55*, 244–252.

Huttenlocher, J., Hedges, L. V., Corrigan, B., & Crawford, L. E. (2004). Spatial categories and the estimation of location. *Cognition*, *93*(2), 75-97.

Huttenlocher, J., Hedges, L.V., & Duncan, S. (1991). Categories and particulars: prototype effects in estimating spatial location. *Psychological Review, 98*, 352–376.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of experimental psychology: General*, *129*(2), 220.

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

Ji, L. J., Zhang, Z., & Nisbett, R. E. (2004). Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of personality and social psychology*, *87*(1), 57.

Kanayet, F., Opfer, J. E., & Cunningham, W. A. (2010). Electrophysiological evidence for multiple representations of number in the human brain. In *Proceedings of the XXXII Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum*.

King, D. L. (1988). Assimilation is due to one perceived whole and contrast is due to two perceived wholes. *New Ideas in Psychology*, *6*(3), 277-288.

Kruschke, J. K. (2010). *Doing Bayesian data analysis*. Oxford, England: Elsevier Academic Press.

Landy, D., Allen, C. & Zednik, C. (2014). A perceptual account of symbolic reasoning.

Frontiers in Psychology, 5, 275. doi: 10.3389/fpsyg.2014.00275

Landy, D. & Goldstone, R. L. (2005). How we learn about things we don't already understand. *Journal of Experimental and Theoretical Artificial Intelligence, 17*, 343-369. doi: 0.1080/09528130500283832

Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. Cognitive Science, 37, 775–799.

Landy, D., Silbert, N., & Goldin, A. (2012). Getting Off at the End of the Line: The Estimation of Large Numbers. *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Laski, E. V., & Siegler, R. S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development, 78(6), 1723-1743.*

LeFevre, J. A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 216.

Leslie, A. M., Gelman, R., & Gallistel, C. R. (2008). The generative basis of natural number concepts. *Trends in cognitive sciences, 12*(6), 213-218.

Lochhead, J. (1988). *Some pieces of the puzzle* (pp. 71-82). Hillsdale, NJ: Lawrence Erlbaum.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077-1083.

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.

McCloskey, M. (1992). Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition, 44,* 107–157.

Moeller, K., Pixner, S., Kaufmann, L., & Nuerk, H. C. (2009). Children's early mental number line: Logarithmic or decomposed linear?. *Journal of experimental child psychology*, *103*(4), 503-515.

Newcombe, N. S. (2013). Seeing Relationships: Using Spatial Thinking to Teach Science, Mathematics, and Social Studies. *American Educator*, *37*(1), 26.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 282.

Nuerk, H. C., Weger, U., & Willmes, K. (2001). Decade breaks in the mental number line? Putting the tens and units back in different bins. *Cognition, 82* (1), 25–33.

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, *55*(3), 169-195.

Opfer, J. E., Siegler, R. S., & Young, C. J. (2011). The powers of noise-fitting: Reply to Barth and Paladino. *Developmental Science,14,* 1194–1204.

Resnick, I., Newcombe, N. S., & Shipley, T. F. (2014). Dealing with Big Numbers: Representation and Understanding of Magnitudes Outside of Human Experience. Manuscript under review.

Resnick, I. & Shipley, T.F. (2013). Application of the Category Adjustment Model in Temporal, Spatial, and Abstract Magnitude at the Billions Scale. In Knauff, M., Pauen, N., Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Resnick, I., Shipley, T., Newcombe, N., Massey, C., Wills, T. (2012). Examining the Representation and Understanding of Large Magnitudes Using the Hierarchical Alignment model of Analogical Reasoning. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*

R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rips, L. J. (2012). How many is a zillion? Sources of number distortion. *Journal of Experimental Psychology: Learning, Memory, and Cognition 39*(4), 1257-1264.

Siegler, R. S. (2009). Improving the Numerical Understanding of Children From Low-Income Families. *Child Development Perspectives, 3*(2), 118-124.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*, 237–243.

Siegler, R. S., Thompson, C. A., & Opfer, J. E. (2009). The Logarithmic-To-Linear Shift: One Learning Sequence, Many Tasks, Many Time Scales. *Mind, Brain, and Education*, *3*(3), 143-150.

Simmering, V. R., & Spencer, J. P. (2007). Carving up space at imaginary joints: Can people mentally impose arbitrary spatial category boundaries?.*Journal of Experimental Psychology: Human Perception and Performance*,*33*(4), 871.

Stenning, K., & Van Lambalgen, M. (2008). *Human reasoning and cognitive science.* MIT Press.

Sullivan, J., & Barner, D. (2012). How are number words mapped to approximate magnitudes?. *The Quarterly Journal of Experimental Psychology, 66*(2), 389-402.

Tanasescu, V. (2007). Spatial semantics in difference spaces. In *Spatial Information Theory* (pp. 96-115). Springer Berlin Heidelberg.

Thompson, C.A., & Opfer, J.E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development, 81,* 1768-1786.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic bulletin & review*, *7*(3), 424-465.

Young, C. J., & Opfer, J. E. (2011). Psychophysics of numerical representation: A unified approach to single-and multi-digit magnitude estimation. *Zeitschrift für Psychologie/Journal of Psychology*, *219*(1), 58.

Zhang, J., & Norman, D. A. (1995). A representational analysis of numeration systems. *Cognition*, *57*(3), 271-295.

Appendix

Response Model.

We develop a model response based on the Category Adjustment theory proposed by Huttenlocher et al. (1991). This model is conceptually displayed for the hypothetical stimulus 200 million in Figure A1.

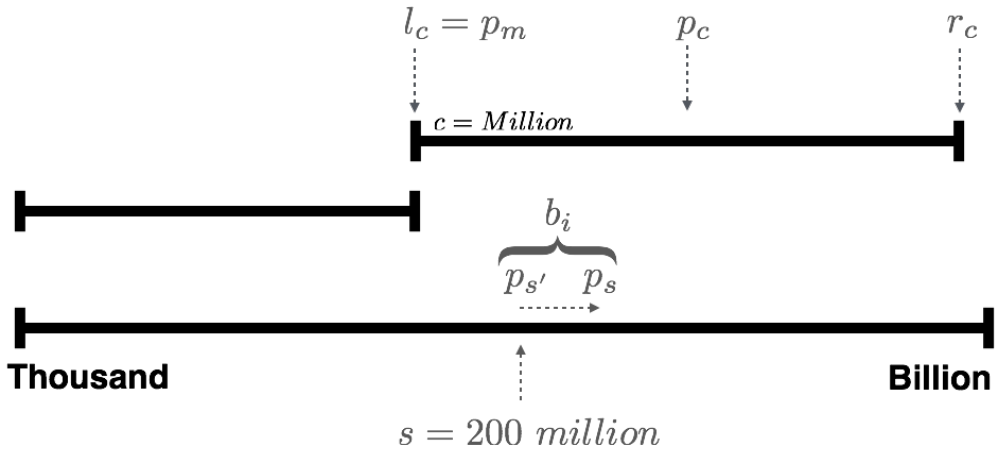


Figure A1: The response model. In this parameterization, b_i>1, indicating systematic inward bias toward the category center.

We assume that people first assign a location for the category boundary at 1 million, p_m. Then, when a stimulus number, s, is presented, they decide which scale category that number belongs in: *millions* or *thousands*. We'll call that category *c* (for simplicity, we ignore the possibility that people may only weakly categorize elements near the boundary, as in Huttenlocher, Hedges, Corrigan, and Crawford, 2004). Then, people determine an initial

stimulus location, $p_{s'}$, linearly based on the category endpoints, l_c and r_c, and the numerical values min_c and max_c.

$$p_s = l_c + (r_c - l_c)\frac{s - min_c}{max_c - min_c}$$

For instance, say that a participant puts the category boundary at 0.3, and extends their responses to the far right hand edge, i.e., $r_c=1$. Then the initial location for *200 million* is determined as follows: 200 million is in the millions category, which has its left end at p_m=0.3. The maximum value is 1 billion and the minimum 1 million, so

$$p_{s'} = 0.3 + (1 - 0.3)\frac{200 \times 10^6 - 10^6}{10^9 - 10^6} \approx 0.44$$

This initial location is associated with a participant-specific noise parameter, σ_i, accounting for variability in participant responses. This initial location is then scaled toward or away from the midpoint by category-based adjustments. Following Crawford, Landy, & Presson, 2014 (see also Barth, et al, 2014), we allow for two sorts of category-based bias: inward adjustment toward category exemplars, and outward contrastive bias away from them. Inward bias is typically taken to serve a noise-decreasing function: by biasing responses toward the center of the *millions* category, one is sure to at least be approximately right: for participants with large noise relative to category size, this approach can reduce overall error (Huttenlocher, Hedges, & Vevea, 2000). The functional basis for outward bias, if any, is less clear, but it may serve to increase within-category differences; this is perhaps especially important for categories such as *thousands*, which normatively take up very little space on the line in the current task. Because here category adjustment and outward bias are both presumed to be linear in distance from the category center,

the net effect can be captured in a single participant-level parameter, b_i, which we will call the *prototype bias*. Higher values of b_i indicate a greater inward tendency, while lower values indicate an outward bias. Thus, if the physical midpoint of the category on the line is p_c, the resulting predicted response location is

$$p_s = (1 - b_i)(p_{s'} - p_c) + p_c$$

When b_i=0, responses are not biased by the category, and so the model exhibits piecewise linearity (or full linearity, if $p_m \sim 0.001$). b_i<0 corresponds to a contrast pattern (and an overlap discontinuity), while b_i>0 corresponds to inward assimilation toward category prototypes (leading to a gap discontinuity). b_i has sensible values in the range from -1 to 1.

We developed this final model through a process of model expansion (Gelman & Shalizi, 2013). We began by observing residuals from the simple segmented model reported in Landy et al. (2013), noting that the model provided only a poor fit to the data, particularly around 1 million (see Figures 6 and 7). Several models in the categorical vein were fitted with various structures and prior parameters to capture the patterns reported here. Each participant had several fitted parameters, including an estimated location for each of 1 thousand, 1 million, and 1 billion, and a degree of prototype bias.

Response Time Model

$$pdf(x; \mu, \sigma, \tau) = \frac{1}{\tau} e^{\left(\frac{\mu - x}{\tau} + \frac{\sigma^2}{2\tau^2}\right)} \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma} - \frac{\sigma}{\tau}} e^{-\frac{t^2}{2}} dt$$

ExGaussian models divide responses into exponentially distributed decision times and normally distributed non-decision times (Van Zandt, 2000). In this model, each of these was

parameterized by the stimulus factors listed above, and the result was fitted using maximum-likelihood to produce parameters for each individual.

The non-decisional time corresponds to μ in the exGaussian model; both decisional components were summed to yield an estimate of τ. τ was therefore the sum of a decision about the category (roughly treated as taking time exponential in psychological proximity to 1 million) and about where in a given category to place the element in the line:

$$\tau = cat(d_s)\beta_2 e^{-(1+d_s)^{\alpha}} + \beta_1 s + \beta_0$$

The psychological distance between the stimulus and the boundary was considered to be a combination of the numerical distance d_n (scaled by whether the stimulus was larger or smaller than 1 million) and the categorical distance d_c (whether the item fell into the thousands or millions range). This allows the model to capture the relative role of surface dissimilarity (Cohen, 2009) and numerical magnitude in accounting for overall dissimilarity—for instance, whether 1.5 million is more similar to 1 million than is 500 thousand.

$$d_s = \delta_0 d_n(s) + \delta_1 d_c(s)$$

In order to adequately capture the data, it was necessary to include a new idea, familiar from category adjustment theory: the idea that elements sufficiently near the boundary are not categorized at all. This is sensible, since those elements are not clearly members of one category or the other, and therefore no increase in accuracy is expected by taking their ambiguous categories into account. For instance, the exact element ‘1 million’ would simply not be

categorized (Huttenlocher, Hedges, Corrigan, & Crawford 2004). This is especially plausible on examination of the response distribution (Figure 10), which suggests that the elements in the low millions were treated ambiguously. In the response time model, we captured this with a simple sigmoid function, which modulated the degree to which categorization was presumed to occur.

$$cat(d) = \frac{1}{1 + e^{\delta_1(\delta_0 - d)}}$$

Several patterns emerged from the model fits, though it should be clear that the response time model itself has 10 parameters (fitting 172 data points per person), and is quite flexible. Furthermore, statistical significances were not corrected for multiple comparisons. Non-decisional response time (μ) was not different in any conditions, nor was the variability in non-decisional time (σ) (all p's >0.25); there was a marginally different slope in the decision time (τ), such that linear responders took longer for trials involving larger numbers (β~3e-07; 95% CI= [1.9e-07, 7.0e-07], p~0.08) more so than did overestimaters. No other patterns reached even uncorrected significance.

Table 2: Parameters estimated for each individual.

Var	*Range*	*Description*	*Distribution*
M	[0..1]	Location on line of "1 million"	Beta
p	[-1..1]	Inward bias toward prototypes	Beta, Rescaled from [0..1]
K	[0..1]	Expected position of 1 thousand	Beta
B	[0..1]	Expected position of 1 billion	Beta
P(M)	[0..1]	Probability that a participant is drawn from the overestimation group	Beta
D	[0.. ∞]	Response Deviation	Gamma

Table 3: population level prior and posteriors for the linear group.

Var	Range	Prior distribution	95% HPD
M_{linear}			
α	[0..∞]	*Gamma(20, 50)*	[0.92, 1.3]
β	[0..∞]	*Gamma(60, 2)*	[18, 27]
$M_{overestimate}$			
α	[0..∞]	*Gamma(60, 20)*	[3.9, 5.4]
β	[0..∞]	*Gamma(150, 15)*	[7.2, 9.4]
K			
α	[0..∞]	*Gamma(2, 1.5)*	[0.57, 0.91]
β	[0..∞]	*Gamma(15, 1.5)*	[14, 23]
B			
α	[0..∞]	*Gamma(15, 1.5)*	[10, 15]
β	[0..∞]	*Gamma(2, 1.5)*	[0.9, 1.5]
D_{linear}			
Shape	[0..∞]	*Gamma(5, 1.5)*	[0.48,0.71]
Scale	[0..∞]	*Gamma(7, 8)*	[5.0, 9.0]
$D_{overestimate}$			
Shape	[0..∞]	*Gamma(5, 1.5)*	[0.44, 0.75]
Scale	[0..∞]	*Gamma(7, 8)*	[2.5, 5.5]
P_{linear}			
α	[0..∞]	*Gamma(5, 0.1)*	[100, 180]
β	[0..∞]	*Gamma(5, 0.1)*	[110, 190]
$p_{overestimate}$			
α	[0..∞]	*Gamma(5, 0.1)*	[17, 38]
β	[0..∞]	*Gamma(5, 0.1)*	[10, 22]
P(M)			
α	[0..∞]	*Gamma(5, 50)*	[0.08, 0.13]
β	[0..∞]	*Gamma(5, 50)*	[0.12, 0.23]