



Cognitive Science (2013) 1–25

Copyright © 2013 Cognitive Science Society, Inc. All rights reserved.

ISSN: 0364-0213 print / 1551-6709 online

DOI: 10.1111/cogs.12028

## Estimating Large Numbers

David Landy,<sup>a</sup> Noah Silbert,<sup>b</sup> Aleah Goldin<sup>c</sup>

<sup>a</sup>*Department of Psychology, University of Richmond*

<sup>b</sup>*Center for Advanced Study of Language, University of Maryland*

<sup>c</sup>*University of Richmond*

Received 21 December 2011; received in revised form 3 July 2012; accepted 17 August 2012

---

### Abstract

Despite their importance in public discourse, numbers in the range of 1 million to 1 trillion are notoriously difficult to understand. We examine magnitude estimation by adult Americans when placing large numbers on a number line and when qualitatively evaluating descriptions of imaginary geopolitical scenarios. Prior theoretical conceptions predict a log-to-linear shift: People will either place numbers linearly or will place numbers according to a compressive logarithmic or power-shaped function (Barth & Paladino, 2011; Siegler & Opfer, 2003). While about half of people did estimate numbers linearly over this range, nearly all the remaining participants placed 1 million approximately halfway between 1 thousand and 1 billion, but placed numbers linearly across each half, as though they believed that the number words “thousand, million, billion, trillion” constitute a uniformly spaced count list. Participants in this group also tended to be optimistic in evaluations of largely ineffective political strategies, relative to linear number-line placers. The results indicate that the surface structure of number words can heavily influence processes for dealing with numbers in this range, and it can amplify the possibility that analogous surface regularities are partially responsible for parallel phenomena in children. In addition, these results have direct implications for lawmakers and scientists hoping to communicate effectively with the public.

*Keywords:* Psychology; Reasoning; Human experimentation; Mathematical cognition; Mathematical modeling; Number line estimation

---

“And so he went back over the sunny hills and down through the cool valleys, to show all his pretty kittens to the very old woman. It was very funny to see those hundreds and thousands and millions and billions and trillions of cats following him.”

—Wanda Ga’g, *Millions of Cats*

---

Correspondence should be sent to David Landy, Richmond Hall, University of Richmond, Richmond, VA 23173. E-mail: dhlandy@gmail.com

## 1. Introduction

In a 2010 interview, Vice President of the United States, Joe Biden, reported that an unnamed conservative had spent an unprecedented \$200 billion in political ads before the midterm election (Hunt, 2010). Were it true, the money would surely have been poorly spent: This amount would have funded the entire operation of the National Science Foundation for 30 years, or alternatively the entire budget of the U.S. military for 4 months. It is not true of course; the figure Mr. Biden intended was \$200 million (Shear, 2010). The mistake is a familiar one: Although we all know words like “million,” “billion,” and “trillion,” most of us feel only the vaguest sense of their referents. In this study, we will explore how people come to attribute meaning and value to the magnitudes of numbers with which we largely lack direct experience.

Large numbers (here, roughly those between  $10^5$  and  $10^{13}$ ) are interesting for both practical and theoretical reasons. First, experimental tasks in economics and psychology frequently ask people to make decisions involving large numbers (see Bohm & Lind, 1993; Cohen, Ferrell, & Johnson, 2002). Moreover, many arenas of public discourse involve large numbers, including debates about evolutionary biology (Dunning, 1997; Meffe, 1994), nanotechnology (Batt, Waldron, & Broadwater, 2008), and the reliability of DNA testing (Koehler, 1997). And, of course, many countries are currently involved in ongoing and heated political conversations about governmental budgets and national economies. The United States is not atypical: there, the budget, the deficit, and the debt are in the low trillions, while most proposed budgets involve changes on the order of millions and billions. Better understanding of how people make sense of numbers in this range has the potential to improve communication in these important arenas.

Large numbers are an excellent example of an abstract system: Magnitudes such as 1 billion are beyond our immediate experience and yet are clearly understood in part through generalizing the concrete process of counting (Carey, 2009; Leslie, Gelman, & Gallistel, 2008). Nearly every adult knows how to count to 1 billion and how to manipulate numbers in this range using basic arithmetic (Skwarchuk & Anglin, 2002), but for most of us, we do not directly work with large numeric magnitudes. Instead, relevant experiences are typically limited to encounters with large number representations through associations with relevant situations (usually by encountering assertions; e.g., that the U.S. budget deficit is \$1.4 trillion; Facebook has 700 million users; or the human body has 100 trillion cells).

One way we understand abstractions is by studying the properties of their concrete representations (Barsalou, 1999; Clark, 2006; Kirsh, 2010; Landy & Goldstone, 2007). For instance, Carey (2009) proposes that when learning to count, the memorized count list orients attention to appropriate features of the environment, so that the verbal label “eighteen” cues a learner that there is *something* that “eighteen” situations have in common. In addition to the simple presence or absence of labels, count lists have structural properties: They are typically stated in sequence, with accompanied rhythmic hand motions, and are constructed on a semiregular linguistic pattern. Here, we investigate how these structural components of symbolic systems impact inferences made by reasoners.

### 1.1. Structure in the numerals

A student learning the English counting system must master several different lists. In addition to the numbers from 1 to 9, for instance, one must learn the teen words and the tens words (10, 20, 30, and so on). The most important count list for our purposes is the *short scale*, commonly used in the United States and Britain (Conway & Guy, 1996). In this system, 1,000 million is “1 billion.”<sup>1</sup> This list “thousand, million, billion, trillion, quadrillion, ...” constitutes an effective count list, which after the initial “thousand,” bears an apparent sequential structure and clearly derives from Latin number words.

To represent a number in the short scale, one divides the number into an order (e.g., “thousands”) and a numeral phrase that runs from 0 to 999; one can think of this system as a standard positional notation with base 1,000; by analogy, we will treat a representation such as 340,000,000 as having the *digit* 340 and the *place* “000,000” or “million.”

In British and American numerical grammar, some words combine *additively*, and others combine *multiplicatively* (Hurford, 1987, 2007). *Forty-three* is an example of the former, as its value just  $40 + 3$ . *Seven hundred* illustrates multiplicative combination,  $7 \times 100$ . Normatively, words in the short scale combine multiplicatively with the immediately preceding term.

The short scale is unlike the other standard scales (units and tens words) in two ways. First, like “hundred,” the short scale combines multiplicatively. Second, the magnitudes captured by the short scale increase exponentially rather than linearly.

North American students typically learn the short scale up to “trillion” by around seventh grade (Skwarchuk & Anglin, 2002). Substantial research has investigated the struggles of children learning the verbal number system, although most investigations have focused on numbers under 100, with very few explorations of numbers over 1,000, let alone over 1 million (e.g., Baroody & Price, 1983; Gelman & Gallistel, 1978; Siegler & Robinson, 1982; but see Skwarchuk & Anglin, 2002).

### 1.2. Number line estimation and its relationship to magnitude processing

In the number-line estimation task (Siegler & Opfer, 2003), participants estimate the appropriate location for a presented number on a line with labeled end points. Roughly, children present one of two patterns of behavior on number-line tasks: Given a particular age range and a particular number range, students either err by compressing the terms toward the high end or correctly place terms close to linearly (Booth & Siegler, 2006). The transition from compressive to linear placement is typically extended, happening at different times for different number ranges. The shift to linear behavior for 0–100 typically occurs between kindergarten and second grade (Siegler & Booth, 2004), for 0–1,000 between second and sixth grade (Siegler & Opfer, 2003), and for ranges up to 100,000 (the largest previously tested) between fourth and sixth grade (Siegler, Thompson, & Opfer, 2009; Thompson & Opfer, 2010). Individual transitions are often quite sharp, with many participants exhibiting single-trial learning after feedback (Opfer, Siegler, & Young, 2011). Children exhibiting linear number-line behaviors also typically have better

memory for numbers (Thompson & Siegler, 2010), suggesting that number-line behaviors relate meaningfully with behavior on other numeric tasks.

Several theories have been proposed to account for children's number-line behavior. Opfer et al. (Opfer et al., 2011; Siegler & Opfer, 2003) explain the behavioral transition in terms of a discontinuous shift from logarithmic to linear representations of numeric magnitude. An alternative pair of models, reminiscent of the account pursued here for large number ranges, holds that children develop separate expectations for numbers under 10, and those between 10 and 100 (Moeller, Pixner, Kaufmann, & Nuerk, 2009; Nuerk, Weger, & Willmes, 2001), or, more generally, separate expectations for numbers they know, versus those they do not (Ebersbach, Luwel, Frick, Onghena, & Verschaffel, 2008). Barth and Paladino (2011) analyze number-line estimation as a proportion judgment task, using a power function to model the compression of large numbers (extending the cyclic power model of Hollands & Dyre, 2000). They then explain the observed behavioral shifts through a postulated continuous shift in proportion estimation and through refinements in children's understanding of larger reference numbers.

Although these models are conceptually distinct, both the segmented models of Nuerk et al. and Ebersbach et al. and the variant of the cyclic power model used by Barth and Paladino produce predictions extremely similar to the logarithmic model, making them difficult to distinguish empirically (Young & Opfer, 2011; Opfer et al., 2011). Although our experiments use a different population over a much larger number span, they may illuminate children's behavior by presenting clear-cut illustrations of adult behaviors relatable to the models that have been applied to children. These relationships are taken up again in the General Discussion.

### 1.3. *The magnitude of large number words*

Although line estimation behavior by children and adults using small number stimuli has been studied extensively, little is known about how these processes extend to large numbers and nonnumerical formats. At least three behaviors are plausible in adults, corresponding to the models that have been applied to children. The simplest possibility is that people might roughly correctly estimate the relative values of large numbers, placing the numbers linearly with numerical magnitude (we will call this the *linear* model). In this system, the subjective magnitude of a simple numeral is calculated by considering the value of the "digit,"  $d$  multiplied by 10 raised to the power of the number of zeroes in the place,  $z$ .

$$E = d \times 10^z$$

Of course, this system can be extended to account for numbers with nonzero values in multiple places, so that the subjective value of 340 million 216 thousand 530 would be

$$E = 340 \times 10^6 + 216 \times 10^3 + 530 \times 10^0$$

Second, in line with research on children's placement of small numbers and following theories of value in economics, adults might scale large numbers logarithmically or

pseudologarithmically (as in the log-linear model for small numbers presented by Siegler & Opfer, 2003).

Third, if learners use their concrete experience of small number words, which have uniformly spaced magnitudes, as the foundation for their understanding of numbers beyond immediate experience, then they may form the erroneous conclusion that the short scale is at least roughly uniformly spaced, for instance, concluding that 1 million lies about halfway between 1 thousand and 1 billion. If the short scale is interpreted multiplicatively, this system would seem to violate the basic principle that a counting system should be monotonic in magnitude (i.e., that each number in the count list should be larger than all the numbers that came before it), as 1 billion would be smaller than 3 million. However, if the short scale were also additive, like 43, the system would again be monotonic: 3 million is then three plus 1 million, and 10 billion is 10 plus billion. This is equivalent to uniformly spacing the orders (1 million, 1 billion, 1 trillion, ...), and linearly interpolating the digits between those orders. The strict uniform spacing criterion can be relaxed if the short-scale terms are a function of the number of zeroes; equivalently, if the number ranges picked out by the short scale are considered to be of slightly different sizes. We will call this the *segmented linear* account. Extended to the beginning of the short scale, the *segmented linear* model predicts that at least some children would space 10, 100, and 1,000 roughly uniformly (as proposed by Nuerk et al., 2001). If so, experience seems to push children away from the additive interpretation, as most older children and adults map numbers linearly in this range.

In short, there are three coherent, plausible ways to map the number words onto magnitudes: The intended *linear* system in which the words thousand, million, and billion scale exponentially (by factors of 1,000), and a second *segmented linear* model, in which short-scale names are uniformly distributed in magnitude, but numbers between them are linearly interpolated. Finally, the *log-linear* system involves scaling terms to the logarithm of their actual value, and adjusting to fit the resulting quantities on a line. All three systems are monotonic in counting and have precedents in the low number system. Although we assume linearity in the first two models, it is worth noting that both can be seen as special cases of the cycle power model, as is detailed in Appendix S1. We did not test the full cyclic power model in these cases only because the deviations from the three basic models above went beyond our primary interest here.

Experiment 1 explores the state of large-number knowledge in adults. Although adults generally perform linearly estimating values up to around 100,000, we predicted that its sequential structure might incline some adults to treat the short scale as an additive system. Experiment 1 uses a number-line estimation task to investigate the presence of nonlinear magnitude estimation behaviors in a typical undergraduate population. Experiment 2 examines the distribution of strategies in a different population, and explores the relationship between evaluative judgments of political situations and the use of the segmented linear strategy in number-line estimation.

## 2. Experiment 1

### 2.1. Method

#### 2.1.1. Participants

Sixty-seven participants recruited from the University of Richmond community received either partial course credit or monetary compensation. All participants had completed college or were currently undergraduates in college. Three participants gave responses that were generally nonincreasing across the number range and were extremely variable; these participants’ data were removed prior to analysis, and their data were replaced to bring the total up to 64.

#### 2.1.2. Procedure

Instructions informed participants that they would be placing numbers on a line, and a computer display presented a number line marked with each integer from 1 to 10. After this, participants sequentially selected (by mouse click) a location for 108 numbers presented in per-participant random order on a number line with end points marked with the values 1 thousand and 1 billion (see Fig. 1). No feedback was provided.

Each number was the product of an integer strictly between 1 and 1,000, and either  $10^3$  or  $10^6$ . Large numbers are written in a variety of formats: In English, numbers may be represented as numerals (5,000), as numbers words (5,000), or in what we might call the hybrid system (5,000). While hybrid notation is familiar and convenient, the numeral notation makes the magnitude information more explicit. In Experiment 1, participants were randomly assigned to either a numeral or a hybrid number condition. We hypothesized that numerals might remind some participants of the normative magnitude of the numbers, limiting use of the segmented linear strategy.

Second, the *range* of the numbers was manipulated between participants. In the *thousands-and-millions* condition, half of the numbers were distributed linearly across the range from 1 million to 1 billion, and half were between 1 thousand and 1 million. In this condition, half of the stimuli should normatively have been placed within two pixels of the left-hand end point. This might have seemed implausible to participants and inclined them to adjust their strategy to spread out these low numbers. We hypothesized that if participants were engaging in such task-specific reasoning, they might use the additive strategy only or primarily when some presented magnitudes lay in the range under

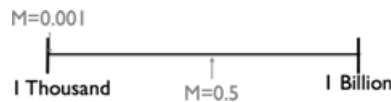


Fig. 1. A graphical depiction relating values of  $M$  to the number line. The linear model can be captured by setting  $M$  to  $(10^6 - 10^3) / (10^9 - 10^3) \approx 0.001$ , while the segmented linear error is captured by setting  $M = 0.5$ . The model then places the stimulus “1 million” at  $M$ . Responses above and below  $M$  are interpolated linearly between  $M$  and the nearest end point.

1 million. To explore whether the choice of range affected strategy selection, half of participants were placed in a *millions-only* condition in which numbers were roughly uniformly distributed across the range 1 million to 1 billion.

After completing the experiment, regardless of condition, participants filled out a paper form prompting them to generate the numerical form for each of 1 billion, 1 million, and 1 thousand. All but two participants did so correctly; one participant left the “1 billion” mark blank, while the other made significant errors. As the primary goal of this study was to understand how participants interpret given number formats, not understanding of conversions between formats, this participant was included in the primary analysis.

## 2.2. Results

The full set of responses is displayed in Fig. 2, divided by condition. Responses are coded as proportion judgments, measured from 0 (the extreme left edge of the line) to 1 (the extreme right edge). Two patterns of response are visible: one line that corresponds to linear placement, and one more diffuse group of responses following a rough line substantially above the midpoint over most of the numerical range. For half of participants, half of responses were under 1 million; it is difficult to see these responses when stimuli are plotted linearly. With what we hope is a minimum of irony, we provide plots with the  $x$ -axis log scaled after Fig. 2, to provide a clear view of the responses to submillion stimuli. To model these patterns, we fit all individual subjects’ data points simultaneously using a multilevel Bayesian model fitting approach. For the purposes of model fitting, each data point is coded as the proportion of the line left of the marked location. The individual subject-level model consists of a function from stimuli to expected line proportions, while the group-level model (probabilistically) governs the parameters of the individual-level models.

### 2.2.1. Analysis

The primary analysis compared linear and segmented models to log-linear models. As both linear and segmented linear models are linear above and below 1 million, the primary variable that distinguishes the models is the estimated position of 1 million on the line ( $M$ ), which can range from 0 (when 1 million is placed on the extreme left of the line) to 1 (the extreme right). The uniform spacing assumption predicts  $M$  values around 0.5, while the normative linear model is produced when  $M$  is 0.00999 (see Fig. 1).

The slight deviations from linearity typical in number-line estimation are often modeled using a power proportion model (Barth & Paladino, 2011; Hollands & Dyre, 2000). This model is extremely general. Although we believe that it is potentially a good model of this sort of task, to constrain our model, we assume strict linearity between the end points and  $M$  (Appendix S1 provide details on the derivation of the model). This restriction is conceptually legitimate, as we expect cyclic deviations from linear responding to be fairly small in adults, and it is pragmatically useful, as the parameters governing cyclic under- and overestimation can mimic some of the shifts in proportion judgments predicted by the segmented linear and logarithmic models (i.e., the restriction prevents problems with overfitting and model identifiability).

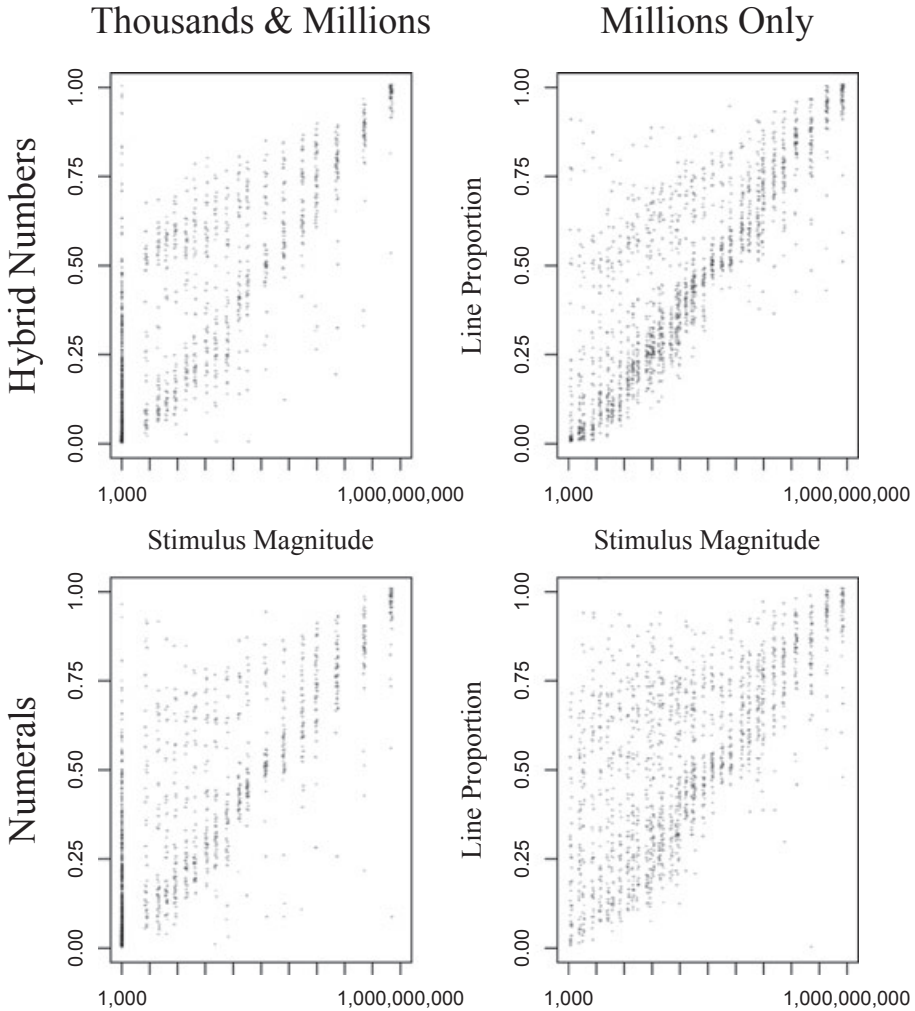


Fig. 2. The full raw data set, divided by counterbalance condition. Each point represents a single estimation of a stimulus number ( $x$  axis). The  $y$  axis represents the proportion of the line that was to the left of the estimated point.

At the individual level, in the linear and segmented linear models, the predicted proportion judgment,  $\eta(x)$ , for stimulus  $x$  is given by

$$\eta(x) = \begin{cases} \frac{x-10^3}{10^6-10^3} \times M & x < 10^6 \\ \frac{x-10^6}{10^9-10^6} \times (1 - M) + M & x \geq 10^6 \end{cases} \quad (1)$$

This model can capture both “linear” and “segmented linear” performance as special cases. In the log-linear model, the predicted proportion judgment is given by



$$\eta(x) = a \ln(x) + b \quad (2)$$

This log-linear model has two free parameters, whereas the (segmented) linear model has only one. In general, the log-linear model predicts straight-line data in log-scaled plots (such as Fig. 4, upper panels), while the (segmented) linear model predicts concave up curves in log-scaled plots (Fig. 4, middle and lower panels). It is worth noting that while the linear and segmented linear models are pinned at the extreme end points ( $10^3$  maps to 0, while  $10^9$  maps to 1), the log-linear model is not, giving it substantial extra freedom to fit variable data patterns. In particular, participants who do not comply with instructions, but just place answers randomly or near the middle of the line will be fit much better by this model than by the segmented linear model.

The log-linear model has been used extensively to model number-line estimators with low knowledge (e.g., Siegler & Opfer, 2003; Booth & Siegler, 2006), and as a model of numerical magnitude understanding in general. Both models capture variability with normally distributed deviations from model predictions (truncated at 0 and 1), fitted as described in Appendix S2.

We applied a multilevel Bayesian approach to fitting the model to the data. In this approach, the segmented linear and log-linear models are fit to individual participants, while simultaneously, the model estimates characteristics of the population from which the individuals are drawn. These parameters at the individual and group level mutually constrain each other; good model fits are those in which the modeled population was likely to yield subjects with the fitted individual parameters and the individual were likely to produce the results they did.

We assumed a population that consisted of three distinct subgroups: a log-linear group (with individual  $a$  and  $b$  values distributed normally) and two segmented linear groups with beta-distributed  $M$  values (i.e., constrained to lie between 0 and 1 inclusive). Conceptually, one of these groups was intended to correspond to roughly linear behavior ( $M = 0.001$ ), while the other was intended to correspond to highly segmented behavior ( $M = 0.5$ ). Beta distributions were chosen to constrain  $M$  to values between 0 and 1. The population parameters (e.g., means and variances of the population) governing the  $M$  and  $a$  and  $b$  parameters at the individual subject level were estimated as part of the sampling procedure. In addition, the frequency of these strategies within the population was estimated.

As is standard in Bayesian model-fitting approaches, we used MCMC to sample the posterior distribution of the model parameters conditioned on the observed data (e.g., Gelman, Carlin, Stern, & Rubin, 2004; Kruschke, 2010). Point estimates of the model parameters are reported as the marginal means across the posterior samples. The uncertainty of a given parameter estimate is expressed via the 95% highest density interval (HDI)—the range of values that most frequently occur in the samples of the posterior. We use a waterline approach to estimating these intervals (Kruschke, 2010, p. 626–629). See Appendix S2 for further details of the fitting procedure.

### 2.2.2. Model fits

Individually, 59 participants were best fit by the segmented linear or linear model in the posterior distribution, while 6 (9.3%) were best fit by the log model. Of the six fit by the log model, four gave highly variable responses that did not reach the extreme ends of the line, and only two (see Fig. 4a) responded in a manner fairly consistent with a log-linear strategy. (In each case, however, the flexibility (noted above) of the log model to avoid the right-hand end point seemed to be crucial; when a log model constrained to reach one end point (by making  $b$  in Eq. 2 a strict function of  $a$ ) was compared to the linear and segmented linear models, it was never the best-fitting model for any participant). Individual trial residuals from the mean posterior model can be seen in Fig. 3.

The model fit produced two well-separated beta distributions at the population level, one with a mean at  $M = 0.003$  (HDI = [0.0004, 0.0063])—quite close to the normative linear value of 0.001; the other had a mean at  $M = 0.34$  (HDI = [0.27, 0.40])—near, but somewhat to the left of, the midpoint. Overall, the model estimated the proportion of the population fitting the roughly linear pattern as 0.56 (HDI = [0.44, 0.67]); the segmented

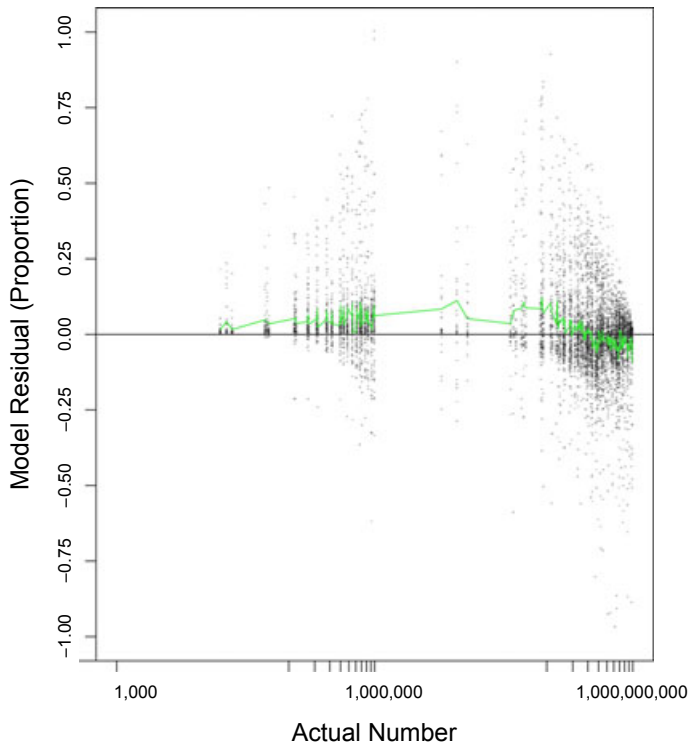


Fig. 3. Residuals of the model fits for Experiment 1. Each data point represents the residual (difference from the mean) from a single placement; the line indicates the running mean residual. The  $x$ -axis in this and all subsequent graphs has been log-compressed to make the full range easier to see. The visible curves above and below 0 reflect subjects who changed strategies across the course of the trials. It can be seen that the model systematically underestimates the estimated proportion of small values and overestimates the location of very large values.

linear pattern as 0.33 (HDI = [0.22, 0.45]); the model estimated the log-linear proportion at .11 (HDI = [0.04, 0.19]).

A histogram of the mean posterior individual subject  $M$  values for participants best fit by one of the segmented linear models is presented in Fig. 5, along with the posterior population-level density estimate. This suggests that there are two populations of participants, corresponding to the roughly linear and the segmented linear; the mean posterior weighting allocated 65% of the participants to the essentially linear bin, and about 35% to the segmented linear bin (Fig. 4 presents data from sample participants in each group).

Our control conditions were intended to evaluate whether nonlinear strategies were directly created by task demands or number format. For instance, half of the numbers estimated by participants in the *thousand and millions* condition were between 1 thousand and 1 million and should normatively have been placed in the leftmost two pixels; this might have encouraged these participants to rescale their preferred number line to fit the stimuli. Instead, segmented linear behavior occurred in all four conditions: It was numerically most infrequent in the millions-only hybrid notation condition and numerically most frequent in the millions-only numeral condition (see Table 1). An ANOVA with mean posterior  $M$  value as the dependent variable was run across the four stimulus conditions to explore whether results were highly dependent on stimulus structure. There was neither a main effect of number format ( $F(1, 57) = .44, p = .51$ ) nor of number range ( $F(1, 57) = 0.13, p = .72$ ), nor was the interaction significant ( $F(1, 57) = 2.25, p = .14$ ); thus, there is no suggestion that these factors strongly impacted behavior. Both strategies occurred in each condition, as can be seen in Fig. 2, verifying that segmented linear performance is not purely an artifact of stimulus structure or format.

### 2.3. Discussion

Experiment 1 demonstrates that in a number-line estimation task involving numbers spanning several orders of magnitudes, people frequently match the predictions of both the (incorrect) segmented linear and (correct) linear number systems; roughly 80% of all participants who provided nonnormative responses seemed to share the misconception that the short scale denotes numbers roughly uniformly spaced in magnitude, even when numbers were presented as numerals. These results are largely compatible with the idea that when experience is limited, people systematically import structures from their notational into their conceptual systems.

There does not appear to be strong evidence for influence of a logarithmic number-line representation extending to large magnitudes. Unlike with small numbers, in which segmented linear strategies and the log-linear strategy resemble each other closely (Opfer et al., 2011), here the two are readily distinguished. A small minority of responses were fit by the log-linear model; however, in over half of these cases, participant responses were highly variable and were difficult to distinguish from participants responding randomly. While logarithmic responding may occur on this task, it seems to be quite rare in this population.

Although the segmented linear account captured the data fairly well, it did not do so, as predicted, with a value of  $M$  at 0.5. Instead, the best-fitting location of 1 million was

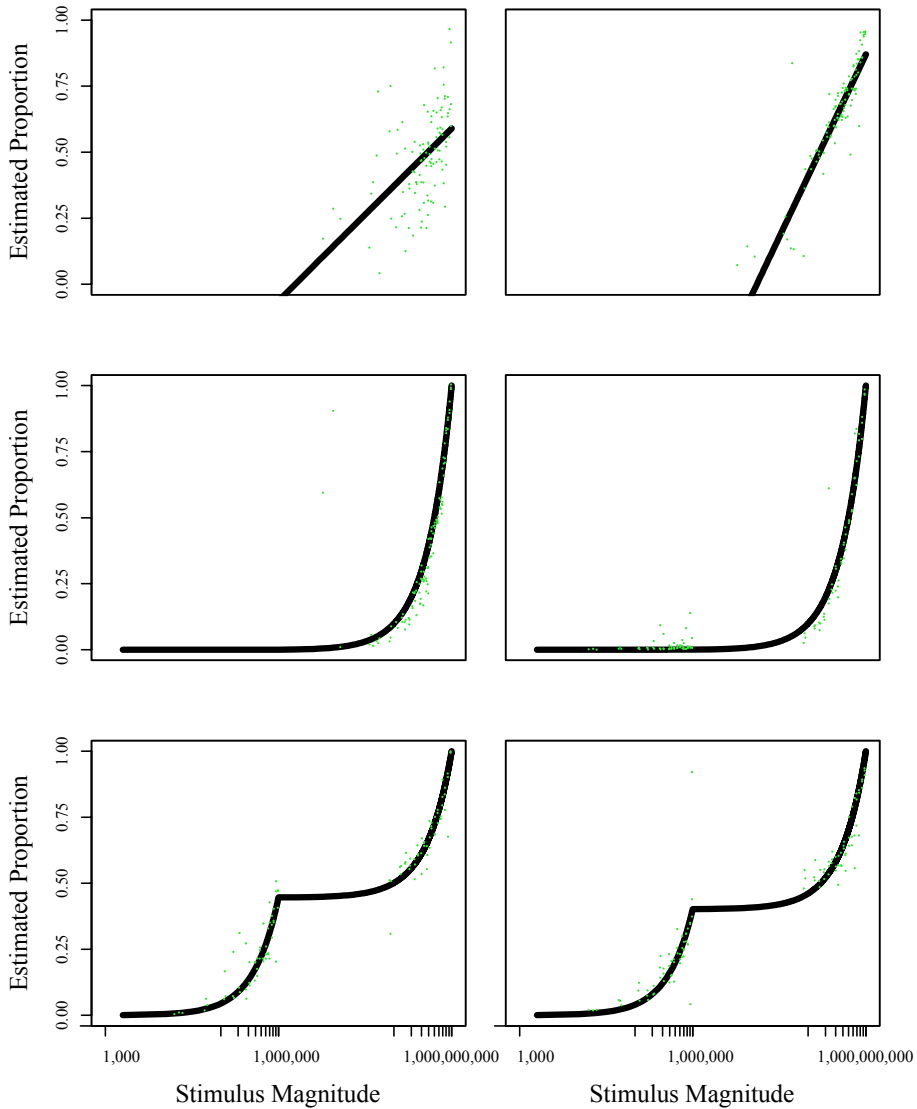


Fig. 4. Individual responses by example participants. Each point represents a single number line estimate; the  $x$ -axis is the (log scaled) value of the stimulus, while the  $y$ -axis indicates the proportion of the line to the left of the participant's mark. The black line indicates the model prediction. The top two panels result from log-linear participants; the middle two panels were fit as linear, and the bottom two as segmented linear. The full set of individual participant model fits can be downloaded from <http://davidlandy.net/DataSets/LandySilbertGoldinExperiment1DataSet.pdf>.

systematically biased to the left of center, as can be seen in Fig. 5. One possibility is that the uniform spacing assumption derived from number-writing notations may serve as a first guide, and that people combine that assumption with evidence from classrooms, media, and life experience that indicate that number words scale nonlinearly.

Table 1  
Number of participants best fit by each strategy, by condition

Condition	Linear	Segmented Linear	Logarithmic
Hybrid format			
Thousands and millions	8	7	1
Millions only	12	4	1
Numerals			
Thousands and millions	9	6	1
Millions only	7	5	3

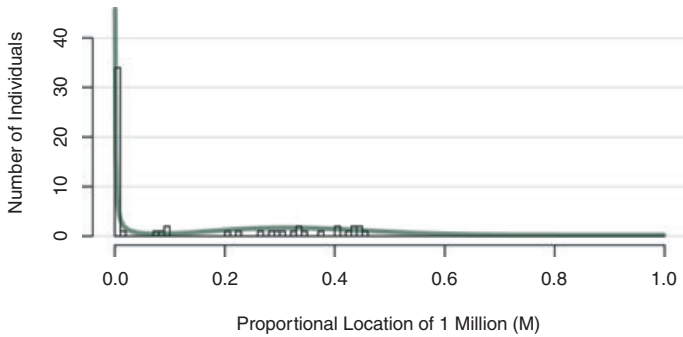


Fig. 5. The fitted values of  $M$  from Experiment 1, for all participants best fit by the segmented linear and linear models, along with the fitted distributions at the population level.

Number line estimation involves visual proportion estimation, and may induce specific reasoning processes and strategies specific to visual processing. For instance, the ease and reliability of line bisection may bias people to use the middle as a reference point (Schneider et al., 2008), which might lead people to a task-specific strategy equivalent to segmented-linear responding. Furthermore, the demands of this task may seem particularly unreasonable, at least in the *thousands and millions* condition, in which half of all the presented numbers should be placed within two pixels of the correct location.

Experiment 2 was designed to replicate Experiment 1 in a broader population and evaluate whether additive number-line behavior is reflective of general comprehension of numbers outside of line estimation tasks. The materials of Experiment 2 were designed to better capture large numbers as they arise in real situations for most Americans.

### 3. Experiment 2

#### 3.1. Method

##### 3.1.1. Participants

Participants were recruited from Amazon's Mechanical Turk. Mechanical Turk is an online website for coordinating scalable workforces, generally for short routine tasks that

can be performed online. It is used extensively by businesses and also increasingly for psychological experimentation. To register as a worker on Mechanical Turk, one must first report a home country and verify being at least 18 years old. Experiment 2 was restricted to participants reporting the United States; no other restrictions on participation were made. Participants were included in the final data analysis only if they responded to all questions.

Four hundred and fifty-six participants were recruited. Fifty-six participants who did not complete all questions were removed and replaced, yielding 200 participants in each of two counterbalanced conditions. Participants reported being over 18 (median reported age was 30) and received small monetary compensation. Participation was restricted to workers reporting living in the United States.

### 3.1.2. Procedure

The experiment consisted of two tasks performed in counterbalanced order. One was a *number-line estimation* similar to that of Experiment 1, except that participants made only five number judgments near points at which the linear and segmented models widely diverged. These values (in presentation order) were 280 million, 3 million, 670 thousand, 840 million, and 470 thousand. Numbers were always presented in the hybrid format. Numbers were presented in fixed order so that the normative location of the first number presented, 280 million, would always be reasonably far from the ends of the line. The end points were again 1 thousand and 1 billion.

In the *evaluation* task, participants read, in fixed order, three short narratives about how the governments of two fictional countries were dealing with various social challenges (see Table 2). The participants rated the quality of the attempted solutions on a 9-point scale from “very unsatisfactory” to “very satisfactory.” In each story, one number was selected as a goal, and a number to be evaluated was selected from the preceding element of the short scale. For instance, in question 3 (designed to match the U.S. budget for 2011), the goal was to eliminate the 1.1 trillion taler deficit, and the solution cut 100 billion talers. Aside from question 3, no attempt was made to make the situations numerically plausible, only to cover the relevant numerical ranges. After both tasks were completed, participants reported their age, sex, and political affiliation, and briefly

Table 2  
The stimuli and results of Experiment 2

Question	Topic/Summary	Mean Segmented—Mean Linear Satisfaction (95% CI)
1	Agrahln is trying to increase immigration from typical levels (1,000) to a goal of 1 billion. Recently, 49 million immigrated	(−0.2, 0.7)
2	Agrahln is attempting to plant 1 billion trees, and has so far planted 95 million	(0.9, 1.7)
3	Tinglan has a budget of 3.7 trillion talers, and a deficit of 1.4 trillion. Legislators cut 100 billion from the proposed budget	(0.2, 1.0)

describing their problem-solving strategy. The strategy explanations provided an extra check that participants were in fact attempting to respond reasonably to the problems.

Recapping our predictions, if the segmented linear assumptions reflect general beliefs about number magnitudes, then people who use this strategy on the number-line estimation task will report more satisfaction with partially successful government policies. On the other hand, if the additive strategy results from reasoning specific to the spatial number line, there would be no reason to expect a difference in overall evaluation conditioned on nonlinearity.

### 3.2. Analysis and results

#### 3.2.1. Number-line estimation

Results were analyzed using just the segmented component from Experiment 1; because the log model is very flexible, six data points were not sufficient to distinguish it from the other models. Fig. 6 presents the distribution of responses.

As in Experiment 1, two beta models were fitted to the population; one corresponded to roughly linear responding (mean  $M = 0.004$  HDI = [0.003, 0.007]), the other to segmented linear responding (mean  $M = 0.43$ , HDI = [0.41, 0.45]), once again slightly to the left of the midpoint. The estimated proportion of the linear group was 0.50 (HDI = [0.45, 0.55]). Fig. 7 shows the distribution of  $M$  values.

#### 3.2.2. Situation evaluation

Participants were divided into two groups based on which of the segmented linear populations they belonged to in the majority of the posterior samples. Because the distributions of responses were far from normal, separate bootstrap analyses (run using the *boot* package in R, 10,000 simulation runs) for each question evaluated the difference in mean response between groups. Results are summarized in Table 2 and displayed in Fig. 8. Across all three questions, the segmented linear group made numerically more positive evaluations than did the linear group, although the difference was significant only for questions 2 and 3. Task order did not affect the results of the evaluation substantially; bootstrap analyses were run on each condition independently, with qualitatively identical results.

### 3.3. Discussion

The number-line estimation task results were very similar to those of Experiment 1: A large minority of participants were best fit by the segmented model. Furthermore, this group responded positively to political actions involving numbers in the range from thousands to trillions, even when those numbers were several orders of magnitudes short of stated goals. Hence, this misconception does not result from reasoning specific to the number-line task. This suggests that number-line concepts may be of immediate practical importance: The form of our current political crises are shaped by a public whose views are, at least in part, driven by skewed conceptions of large numbers.

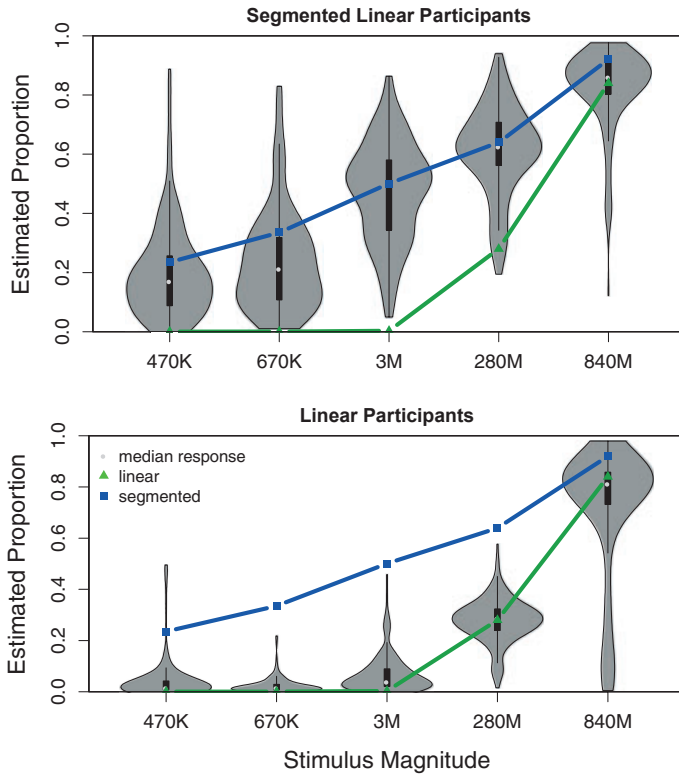


Fig. 6. Number-line estimates from Experiment 2. Each violin plot presents the data from a single stimulus number. The top plot presents data from participants categorized by the model as segmented; the bottom from those categorized as linear. The widths of the gray curves reflect the frequency of responses; the central black region represents 1.5 times the interquartile range. The lines across the data set indicate the predictions of the linear and segmented models, using the fitted population estimates of  $M$ .

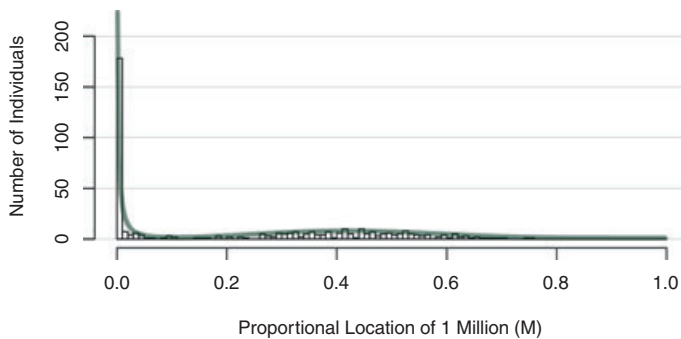


Fig. 7. The fitted values of  $M$  from Experiment 2 for all participants, along with the fitted distributions at the population level.



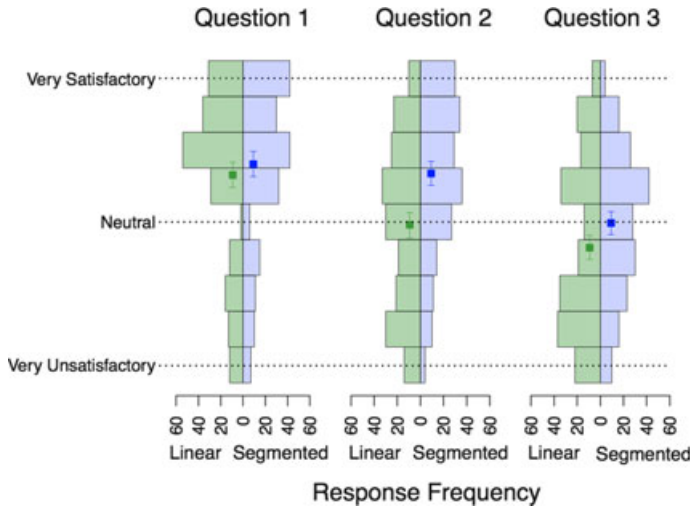


Fig. 8. Situation evaluations by participants who were classified as linear (left) and segmented linear (right). Bars indicate the frequency of responses on a 9-point scale. Boxes represent mean values, with 95% confidence intervals of the mean given by the bootstrap analysis.

Question 1, in which the government tried to increase immigration, failed to elicit large differences between the two groups. Immigration is heatedly debated in the United States, and the question content may have affected the results; post-hoc analyses within self-reported political affiliation indicated that while self-identified Democrats and independents exhibiting the segmented model had significantly more positive evaluations of the immigration action than did the linear group, Republicans fitting the segmented model tended to have *less* positive evaluations than did linear ones (95% CI of the interaction = [0.25, 1.2]). Although unpredicted ahead of time, these results are consistent with the interpretation that Republicans were by-and-large unwilling to entertain a hypothesis in which large-scale immigration has a positive value.<sup>2</sup> On the other hand, this question involved the especially ludicrous idea that immigration of 1 billion could possibly be achieved or desired. It is possible that some participants found this question difficult or impossible to entertain seriously.

There are several mechanisms that might potentially explain the relationship between number-line estimation and situation evaluation. Quite simply, it might be that people who apply the segmented linear strategy are more easily satisfied than those who apply the linear strategy, and so are happier with ineffective actions. Alternatively,<sup>3</sup> it could be that different groups of people take the political goals more literally. Political goals are often optimistic and overly ambitious; indeed, the reported level of achievement in our stories might actually be quite good by the standards of political promises. It might be that people who form linear number-line estimates had better discerned our intended question, and so responded appropriately.

More generally, it is very likely that the specific nature of the narratives chosen in this experiment impacted the judgments participants made on this task. Moreover, it is at least possible that the differences between linear and segmented-linear responders would differ in quality depending on the story material, or the particular type of questions asked. For instance, while each of the narratives we used involved cases in which actions were largely unsuccessful in absolute terms, Question 1 reported a 50,000-fold increase in immigration, which seems quite successful. It is possible that, for instance, segmented-linear responders were more likely to consider the relative increase rather than the degree to which the goals were achieved.

Although more complex reasons for the relation cannot be definitively ruled out, the results reported here are entirely compatible with a simple, coherent story: People who treat a number as being large on a line estimation task also treat that number as being large when it appears in a fictional, but realistic context.

#### **4. General discussion**

Despite appearing frequently in educational contexts and public discourse, numbers picked out by the short scale do not seem to be robustly understood by much of the population. While roughly half of our participants treated large numbers linearly, two experiments indicate that a large portion of the population—around 35% in the studies reported here—seems to evaluate large numbers based on the assumption that the number labels pick out roughly equally spaced magnitudes as the numbers increase, but correctly linearly interpolate numbers between these values. Furthermore, people who rely on segmented linear heuristic when placing numbers on a line are happier with ineffective resolutions to political problems involving comparable scales.

Although it has long been suggested that large values are represented logarithmically (Bernoulli, 1738; Dehaene, Izard, Spelke, & Pica, 2008; Siegler & Opfer, 2003), strict logarithmic behavior was rare or nonexistent on our experiments. This is surprising, given that substantial prior research has supported the hypothesis that unfamiliar number ranges are initially represented logarithmically (Siegler & Opfer, 2003; Dehaene, 2003). On the other hand, the behavior we report, in which number words and numeral representations structure magnitude estimations, is not an accepted account of children's behavior. How might we explain this discrepancy? One possibility is that large numbers fall beyond the upper range of the approximate magnitude system expected to be logarithmically scaled. If so, then children might rely on logarithmic representations of numerosity for smaller numbers (yielding log-linear behavior). For larger number ranges, no such representations are available, and people fall back on representation-specific (short-scale) information.

Another possibility is that children actually do use number representations including the scale words “tens,” “hundreds,” and “thousands” to guide their behavior, and that this strategy at least partially accounts for apparently logarithmic or power-proportional processes. Within low number ranges, up to 100 or 1,000, it has been difficult to distinguish

predictions of the log-linear model from other possible accounts, largely because the close spacing of number words in this range (e.g., 10 is one-tenth of 1,000) and noisiness of the data from children lead to extensive model mimicry (Young & Opfer, 2011). However, an increasingly sizable literature challenging the log-linear shift for small numbers has developed (e.g., Barth & Paladino, 2011; Barth, Slusser, Cohen, & Paladino, 2011; Cohen & Blanc-Goldhammer, 2011). One advantage to the approach taken here is that complex modeling approaches or indirect arguments are not needed to see that behavior is not logarithmic. Furthermore, the reasoning process involved is quite similar to the segmented linear strategy proposed by Nuerk et al. (2001). It is thus at least possible that young children use number representations for 10 and 100 in the same way adults appear to use representations of much larger numbers, but that experience with the referents of these numbers serves as a corrective, leading learners to the linear model. Although experiences with magnitudes on the order of 1,000 are rare, they are much more common than experiences involving magnitudes on the order of 500 million. For example, students assigned to write a 1,000-word essay will quickly learn that it is not twice the length of a hundred-word essay.

Of course, the frequency of the segmented linear strategy does not directly demonstrate the use of the analogous strategy in younger children, but it does demonstrate that under some circumstances, people unfamiliar with a number range might use the structure of notation to make an equal-spacing assumption (P. G. Matthews and D. Chesney, 2011). This suggests that the segmented-linear model is compatible with the kinds of cognitive strategies employed in understanding numbers, making it more plausible that children use scale terms such as tens and hundreds in understanding magnitude.

#### 4.1. *Limitations*

The model we applied to the data has several limitations. First, the model was constrained to fill the line entirely, while subjects often avoid placing marks at the extreme ends of the line. One result is that when  $M$  is far from 0.5, the model fails to exactly match uniform spacing: The space between 4 and 5 thousand will be different than that between 4 and 5 million (see Appendix S1 for more details). Second, the assumption of strict linearity between reference points, although convenient, is likely too simplistic to capture the data fully. The appendix describes the full power proportion model as applied to our data. Third, our assumption that each individual matches a single strategy is likely incorrect. Indeed, in Experiment 1, a small proportion of individuals appeared to suddenly shift between the segmented linear and the linear strategy. Finally, casual observation suggests that people may occasionally incorrectly encode “million” as “thousand” and vice versa: These errors are not captured in the current model. In general, the model captures only an important subset of the full range of behaviors exhibited in number-line estimations of large numbers.

While the segmented linear model predicts error strategies quite well over the range from 1 thousand to 1 billion, it is not clear that it applies across the full range of the short scale. On the low end, our current experiments do not indicate whether roughly uniform

spacing assumptions extend into to the thousands, or whether they begin at 1 million (past research indicates that adults are uniformly linear over the hundreds-to-thousands range). On the high end, most people are probably not familiar with terms in the short scale beyond perhaps trillion or quadrillion; it is not clear how high uniform spacing practices extend. However, over the range from 1 thousand to 1 billion, the segmented assumption appears to apply fairly well for participants who do not reflect normative linear placement.

#### *4.2. Conclusions*

Currently, the people of the United States, along with many other countries, are deciding how best to handle economic and national debt crises. These conversations crucially involve the accurate assessment of numbers across the range of  $10^6$ – $10^{13}$ . The current results suggest that a substantial fraction of Americans are ill equipped to engage in these conversations. Taken together with past work on the linearity of low numbers, the current results suggest a direct practical moral: Putting all references to comparable values into a common scale will improve clarity. For instance, public figures asserting that a budget bill will cut \$100 billion from a \$1,200-billion deficit (rather than a \$1.2-trillion deficit) are more likely to be understood by a large fraction of the population. More generally, thoroughly understanding specific misconceptions with respect to large numbers may lead to the development of interventions that successfully enable the public to work with large numerical magnitudes.

Prior work demonstrates the role that representations play in math misconceptions in understanding decimals (Vamvakoussi & Vosniadou, 2004) and fractions (Opfer, Thompson, & DeVries, 2007). The current results are distinctive in that the misconceived system—the natural numbers and their referents—are not generally considered to be as challenging as the rationals are (e.g., Carey, 2009), and the place-value system and grouping words needed to understand this task are used identically in low-number ranges. A simple induction would suffice to suggest the referents of the large number words studied here, rather than a conceptual restructuring, as has been implicated in rational-number learning. Thus, these results emphasize that even when dealing with basic abstract material, accessible concrete structures play a key role in guiding the development of concepts and strategies (Carey, 2009; Goldstone & Landy, 2010; Landy & Goldstone, 2005). Similarly, when understanding algebraic concepts, the physical structure of the typical notation system plays a huge role, even when the relevant rules are extremely simple (Kirshner, 1989; Landy & Goldstone, 2007).

More precisely, when dealing with large numbers, people rely heavily on number-naming structures to fix the meaningful properties of particular number words. Instead of using the number labels as placeholders to an independently existing world, accessed via number principles, many people attend to the surface properties of number nomenclature to determine numerical properties. These surface properties suggest that the number words in the short-scale scale linearly. As the 4-year-old daughter of the first author (who was at the time learning to read two-digit numbers) put it, “100 is just one more than 10.

It's three: one, two, three!" On this interpretation, number meanings are constructed by borrowing structure from the symbol systems used to represent them.

## Acknowledgments

This research was partially funded by Department of Education, Institute of Education Sciences grant R305A110060, as well as an undergraduate research grant from the University of Richmond to the third author. Thanks to Lisa Byrge, Iris Van Rooij, Lydia Nichols, Erin Ottmar, Dan Navarro, and three anonymous reviewers for helpful comments and suggestions.

## Notes

1. France, Germany, and many other countries use the "long scale," in which 1,000 million is "1 milliard," while 1 million million is "1 billion." We report exclusively about U.S. participants and consider only the short scale.
2. Supporting this interpretation, a follow-up experiment used a similar story, but one in which jobs were being created to support a high birth and immigration rate. Participants rated their satisfaction with the number of created jobs, which ought to be consistently positive across political ideology; in line with our predictions, roughly uniform number-line estimators with each ideology rated the performance as more satisfactory than did linear responders, and in the group as a whole, the difference was significant (95% CI = [0.16, 1.37], on a 9-point scale).
3. We thank an anonymous reviewer for bringing this possibility to our attention.

## References

- Baroody, A. J., & Price, J. (1983). The development of the number-word sequence in the counting of three-year-olds. *Journal for Research in Mathematics Education*, 14 (5), 361–368.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22 (04), 577–660.
- Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift. *Developmental Science*, 14, 125–135.
- Barth, H., Slusser, E., Cohen, D., & Paladino, A. M. (2011). A sense of proportion: Commentary on Opfer, Siegler, & Young. *Developmental Science*, 14, 1205–1206.
- Batt, C. A., Waldron, A. M., & Broadwater, N. (2008). Numbers, scale and symbols: The public understanding of nanotechnology. *Journal of Nanoparticle Research*, 10 (7), 1141–1148. doi:10.1007/s11051-007-9344-1.
- Bernoulli, D. (1738). "Specimen theoriae novae de mensura sortis," *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 5, St. Petersburg 175–92. German translation by A. Pringsheim, *Versuch einer neuen Theorie der Wertbestimmung von Glücksfällen*, Duncker & Humblot: Leipzig 1896. English translation by L. Sommer, "Exposition of a New Theory on the Measurement of Risk." *Econometrica* 22 (1954), 23–36.

- Bohm, P., & Lind, H. (1993). Preference reversal, real-world lotteries, and lottery-interested subjects. *Journal of Economic Behavior & Organization*, 22 (3), 327–348.
- Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, 41, 189–201.
- Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.
- Clark, A. (2006). Material symbols. *Philosophical Psychology*, 19 (3), 291–307. doi:10.1080/0951508060068987.
- Cohen, D. J., & Blanc-Goldhammer, D. (2011). Numerical bias in bounded and unbounded number line tasks. *Psychonomic Bulletin and Review*, 18, 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.
- Conway, J. H., & Guy, R. K. (1996). *The book of numbers*. New York: Springer Verlag.
- Dehaene, S. (1997). *The number sense*. New York: Oxford University 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, 1217–1220.
- Dunning Jr., J. B. (1997). The missing awareness, part 2: Teaching students what a billion people looks like. *Conservation Biology*, 11 (1), 6–10.
- 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.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.
- Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.
- Goldstone, R. L., & Landy, D. (2010). Domain-creating constraints. *Cognitive Science*, 34(7), 1357–1377.
- Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review*, 107 (3), 500–524.
- Hunt, A. (2010). Political capital with Al Hunt. Retrieved from Bloomberg News Web site: <http://www.bloomberg.com/video/63925814/>. Accessed January 23, 2013.
- Hurford, J. R. (1987). *Language and number*. Oxford, England: Basil Blackwell.
- Hurford, J. R. (2007). A performed practice explains a linguistic universal: Counting gives the packing strategy. *Lingua*, 117, 773–783.
- Kirsh, D. (2010). Thinking with external representations. *AI & Society*, 25 (4), 441–454. doi:10.1007/s00146-010-0272-8.
- Kirshner, D. (1989). The visual syntax of algebra. *Journal for Research in Mathematics Education*, 20(3), 274–287.
- Koehler, J. J. (1997). One in millions, billions, and trillions: Lessons from *People v. Collins* (1968) for *People v. Simpson* (1995). *Journal of Legal Education*, 47 (2), 214–223.
- Kruschke, J. K. (2010). *Doing Bayesian data analysis*. Oxford, England: Elsevier Academic Press.
- 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.
- Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33 (4), 720–733.
- Leslie, A. M., Gelman, R., & Gallistel, C. R. (2008). The generative basis of natural number concepts. *Trends in Cognitive Sciences*, 12 (6), 213–218. doi:10.1016/j.tics.2008.03.004.
- Matthews, P. G., & Chesney, D. L. (2011). Straightening up: Number line estimates shift from log to linear with additional information. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1936–1941). Boston, MA: Cognitive Science Society.

- Meffe, G. K. (1994). Human population control: The missing awareness. *Conservation Biology*, 8 (1), 310–313.
- 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.
- 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., & Young, C. J. (2011). The powers of noise-fitting: Reply to Barth and Paladino. *Developmental Science*, 14, 1194–1204.
- Opfer, J. E., Thompson, C. A., & DeVries, J. M. (2007). Why children make “better” estimates of fractional magnitude than adults. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th annual cognitive science society* (pp. 64–70). Austin, TX: Cognitive Science Society.
- Schneider, M., Heine, a., Thaler, V., Torbeyns, J., Desmedt, B., Verschaffel, L., Stern, E., & Jacobs, A. M. (2008). A validation of eye movements as a measure of elementary school children's developing number sense. *Cognitive Development*, 23 (3), 409–422.
- Shear, M. (2010). \$200 billion in ads, Mr. Biden? That's real money. *New York Times Blogs*. Available at: <http://thecaucus.blogs.nytimes.com/2010/10/22/200-billion-in-ads-mr-biden-thats-real-money/?ref=politics>. Accessed January 23, 2013.
- Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development*, 75 (2), 428–444.
- 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., & Robinson, M. (1982). The development of numerical understanding. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior*, Vol. 16 (pp. 242–312). New York: Academic Press.
- 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.
- Skwarchuk, S. L., & Anglin, J. M. (2002). Children's acquisition of the English cardinal number words: A special case of vocabulary development. *Journal of Educational Psychology*, 94 (1), 107.
- Spence, I. (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 683–692.
- Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64, 153–181.
- 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.
- Thompson, C. A., & Siegler, R. S. (2010). Linear numerical-magnitude representations aid children's memory for numbers. *Psychological Science*, 21 (9), 1274–1281. doi:10.1177/0956797610378309.
- Vamvakoussi, X., & Vosniadou, S. (2004). Understanding the structure of the set of rational numbers: A conceptual change approach. *Learning and Instruction*, 14 (5), 453–467. doi:10.1016/j.learninstruc.2004.06.013.
- Young, C. J., & Opfer, J. E. (2011). Psychophysics of numerical representation. *Zeitschrift für Psychologie/ Journal of Psychology*, 219 (1), 58–63. doi:10.1027/2151-2604/a000047.

## Appendix A:

### Derivation of the general model

The model used to describe the linear and segmented linear number line strategy can be derived as a special case of a Steven's power law (Stevens, 1957), applied to a propor-

tion judgment task (Spence, 1990). First, assume that the subjective magnitude,  $p$  of an objective stimulus magnitude,  $x$ , has the general form

$$p = \alpha x^\beta$$

When  $\beta = 1$ , the subjective magnitude is linear in the objective stimulus magnitude. When it deviates from 1, the subjective magnitude is compressed or expanded. Now one must place that number on a line, in a way that corresponds with the subjective magnitude. Number line estimations may be seen as an instance of basic proportion estimation, in which the stimulus is compared to some small reference point ( $R_0$ ) and some large reference point ( $R_1$ ), and is placed so that the proportional distance of the placement between from the left-hand and right-hand locations of the reference points on the line matches the subjective proportion of the gap between  $R_1$  and  $R_0$  occupied by the stimulus,  $x$ . This is easiest to picture when the left-hand and right-hand locations are the left and right boundaries of the line, and the reference values are the marked values of the endpoints. In this simple case, the proportion,  $P$ , is predicted to be

$$P = \frac{\alpha(x - R_0)^\beta}{\alpha(x - R_0)^\beta + \alpha(R_1 - x)^\beta} = \frac{(x - R_0)^\beta}{(x - R_0)^\beta + (R_1 - x)^\beta}$$

Now, if  $\beta = 1$ , this model reduces to simple linear placement. This is what is assumed in the models described in the main text. In general, when  $\beta$  is not 1, the  $\beta$  parameter produces cyclic deviations from an otherwise linear response in the form of over- and underestimation of proportions between reference points. When  $\beta$  is one, judgments are linear interpolations between reference points, while values of  $\beta$  less than one produce overestimation immediately above and underestimation immediately below reference points, and values of  $\beta$  greater than one produce the opposite pattern.

Hollands and Dyre (2000) observed that people may use accurately known reference points other than the extreme values to estimate proportions; points such as the midpoint of a line, or even explicit tick-marks, are very useful in estimating proportions. Assuming ordered reference magnitudes  $R_0, R_1, R_2, \dots, R_n$  whose proportions are accurately known, and assuming that proportion of stimulus values between reference points are estimated using only the nearest references, Hollands and Dyre conclude that the predicted proportion would be

$$P = \frac{(x - R_{i-1})^\beta}{(x - R_{i-1})^\beta + (R_i - x)^\beta} \times \frac{R_i - R_{i-1}}{R_n - R_0} + \frac{R_{i-1} - R_0}{R_n - R_0}$$

where we have adapted Hollands and Dyre's model to allow the smallest (left-hand) reference point to be non-zero. Because the deviations from linearity caused by the power-functional representation cycle between reference points in this model, it is called the *cyclical power model*.



However, the segmented linear model presented here does not assume that reasoners employ accurately known reference points. Instead, that model can be seen as a power-proportion interpolation between (possibly) incorrectly known reference points. Assume, then, ordered reference magnitudes  $R_0, R_1, R_2, \dots, R_n$  whose subjective line locations are  $K_0, K_1, K_2, \dots, K_n$ , with all  $K_i$  between 0 and 1. Then the predicted estimated proportion for any point is

$$P = \frac{(x - R_{i-1})^\beta}{(x - R_{i-1})^\beta + (R_i - x)^\beta} \times \frac{K_i - K_{i-1}}{K_n - K_0} + \frac{K_{i-1} - K_0}{K_n - K_0} \quad (\text{A1})$$

Equation A1 provides a general case of proportion estimation with potentially inaccurate reference points. In our case, we assume that the only internal reference point is the point at one million,  $K_1$  (called  $M$  in the main text) We further assume that people accurately represent the endpoints, so that in the experiments reported here,  $R_0 = 1,000$  and  $K_0 = 0$ , while  $R_2 = 1,000,000,000$  and  $K_2 = 1$  (and  $n = 2$ ). Finally, we fix  $\beta = 1$ . In this case, the above equation can be simplified to the equation given in the text:

$$\eta(x) = \begin{cases} \frac{x-10^3}{10^6-10^3} \times M & x < 10^6 \\ \frac{x-10^6}{10^9-10^6} \times (1 - M) + M & x \geq 10^6 \end{cases}$$

In this case, the only free parameter is  $M$ . If  $M$  has the normative value, the model essentially reduces to the Spence power model. If  $M = 0.5$ , it implements the ideal uniformly spaced model.

### Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Appendix S1.** Landy Silbert Goldin Appendix B.