# Getting off at the end of the line: the estimation of large numbers

**David Landy (dlandy@richmond.edu)**
Department of Psychology, 28 Westhampton Way
University of Richmond, VA 23173 USA

**Noah Silbert (nsilbert@umd.edu)**
Center for the Advanced Study of Language
University of Maryland. College Park, MD

**Aleah Goldin (aleah.goldin@richmond.edu)**
Department of Psychology, 28 Westhampton Way
University of Richmond, VA 23173 USA

**Abstract**

Despite their importance in public discourse, numbers in the range of one million to one 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. Common conceptions of the number line suggest a logarithmic compression of the numbers (Dehaene, 2003). Theories of abstract concept learning suggest that in situations where direct experience is unavailable, people will use the structure of notation systems as a proxy for the actual system. (Carey, 2009; Landy & Goldstone, 2007).

Evaluations across two subject populations largely matched the predictions of the latter account. Approximately 40% of participants estimated *one million* approximately halfway between *one thousand* and *one 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. Very brief training procedures proved partially successful both in correcting number line placement and in shifting participants' judgments of geopolitical situations. These results reinforce notions of abstract concepts as grounded in external notation systems, as well as having direct implications for lawmakers and scientists hoping to communicate effectively with the public.

**Keywords:** number cognition, mathematical cognition, formal reasoning, human subjects experimentation

## Introduction

Large numbers[1] are interesting for both practical and theoretical reasons. Many arenas of public discourse rely on an understanding of large numbers, including debates about evolutionary biology, nanotechnology, and the reliability of DNA testing. The United States is currently involved in a heated conversation about the national budget and economy. The budget, the deficit and the debt are in the low trillions, while most proposed budget changes are in the millions and billions. Americans generally exhibit poor knowledge about spending on specific programs by the federal government (Gilens 2001), and it is likely that poor understanding of large numbers contributes to this ignorance.

Number systems covering this range are also an excellent example of an abstract system: magnitudes such as *one billion* are beyond our immediate experience and yet are clearly understood in part through abstracting the concrete process of counting (Carey, 2009; Leslie, Gelman, & Gallistel, 2008). We experience large numbers primarily syntactically, and through associations with situations (e.g., claims that the U.S. 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 (Clark, 2006; Landy & Goldstone, 2007; Kirsh, 2010). 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, however, count lists have other structural properties: for instance, counting numbers are typically stated in sequence, with accompanied rhythmic hand motions, and are constructed on a semi-regular pattern. Here, we wonder how structural components of symbolic systems impact inferences made by reasoners.

### Structure in the numerals

A student learning the English counting system must master several different lists. In addition to the numbers from 1-9, one must learn the teen words, the tens words, and –most importantly for our purposes—is the *short scale*, used in the United States and Britain. In this system, one thousand million is "one billion". 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. North American students typically learn the short scale up to "trillion" by around 7$^{th}$ grade (Skwarchuk and Anglin, 2002).

There are several common notations for understanding large numbers. In this paper, we focus on perhaps the most common on, which we will call the *hybrid notation*, because it combines number words and numerals. Examples of numbers in this from include "324 million", "426", or "5 thousand."

We model large number understanding by combining two conceptually separate steps: the first involves the interpretation of a number word into an abstract numerical quantity ("abstract" because we are agnostic with respect to how people would actually estimate perceived quantities in the range of millions and billions—here we mean merely the interpretation can be treated as a metric), and the mapping

[1]Here, roughly those between $10^5$ and $10^{13}$.

of a quantity into a response. For brevity, we blur these distinctions here.

Of the many plausible ways that people might extract quantities from number representations, the simplest is that people might roughly correctly estimate the relative values of large numbers. We will refer to this as the *linear* or *normative* model of large number understanding.

Second, if learners use the structure of the number notations—especially the short scale—as a guide to numerical size, then a different pattern is expected. Since the number words—millions, billions, trillions, are similar and uniformly spaced in their count list, people might evenly distribute the referred quantities. Since adults generally linearly estimate numbers from 1-1000 (Seigler & Opfer, 2003), this suggests a piecewise-linear pattern, in which (roughly) values like 1 thousand, 1 million, and 1 billion are separate units, which are spaced evenly on the line, and other values (such as 500 million) are linearly interpolated between these points. We will call this the *uniform spacing* or *piecewise linear* model.

Another plausible approach is based on developmental studies of line estimation with small number ranges (Siegler & Opfer, 2003). These studies have repeatedly demonstrated that number estimation errors tend to be highly compressed at the large end of the line. Traditionally, this compression has been modeled using a logarithmic function, and a fitted linear mapping from quantities to line positions (Booth & Siegler, 2008; Siegler & Opfer, 2003). We will call this combination the *log-linear* model.

Finally, it is naturally plausible that some people would either have no interpretation of the large numbers, or highly variable or non-monotonic interpretations.

## Empirical Methodology

We used two tasks to explore the number word interpretation: Number line estimation, and situation evaluation.

In typical *number line estimation* tasks, a participant is presented a line with labeled endpoints, and a stimulus numeral. The participant makes a mark indicating their estimate of the proportion of the line that corresponds to the proportion relating the stimulus number to the specified range. In the experiments reported here, the left end was always *1 thousand*, and the right end was *1 billion.* Prior to performing estimations, participants were shown a marked number line ranging from 1 to10, and were instructed to likewise place their numbers in a linear manner.

In *situation evaluations*, participants made qualitative judgments about attempted government actions involving short-scale quantities. 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 one question a fictional country's government had a goal to eliminate their 1.1 trillion "taler" deficit, and proposed the solution cut 100 billion talers. Participants rated the quality of the attempted solutions on a 9-point scale from "very unsatisfactory" to "very satisfactory".

# Experiment 1

## Method

**Participants & Procedure** Partial course credit or monetary compensation was given to 67 participants recruited from the University of Richmond community. Three participants gave responses that were generally non-increasing across the number range, and were extremely variable; these participants' data were removed and replaced to yield our goal of 64 participants.

Participants made 108 number line estimates, on a line ranging from 1 thousand to 1 billion. Each stimulus number was the product of an integer strictly between one and one thousand, and either $10^3$ or $10^6$.

Two between-groups differences were used to rule out possible confounds in our approach. First, in Experiment 1 half of all participants viewed numbers in the hybrid notation; for half all stimuli and endpoints were presented in the pure numeral format. Second, the range of the stimulus numbers was manipulated between participants, so that we could evaluate whether people shifted their placement to fit the distribution of observed numbers. Half of the participants saw numbers only in the millions; half estimated numbers which were evenly divided between those above and below 1 million. Neither manipulation affected results qualitatively or altered significance of contrasts; similar patterns were observed across all four groups; the slight differences will not be discussed here.

After completing the experiment, regardless of condition, participants filled out a paper form prompting them to generate the numerical form for each of one billion, one million, and one thousand. All but two participants did so correctly; one participant left the "one billion" mark blank, while the other made significant errors.

### Analysis

Our primary analysis compared linear and uniform spacing model fits with the log-linear, using a hierarchical Bayesian model fitting approach.

Since both models are linear above and below one million, the primary variable distinguishing the linear and uniform spacing models is the estimated position of one million on the line (*M*). M was fitted at the individual subject level; since M ranges from 0 (extreme left) to 1 (right), the population was fitted as a uninformative beta distribution. Within this framework, the linear model is the special case when M = 0.001, pure uniform spacing is produced when M=0.5. The prior on M was uniform between 0 and 1, and 0 elsewhere.

This *segmented linear* model was compared to a log-linear model, $y = a\ln(x)$; this model also has one parameter, fixing the shape of the linear component. The left intercept of both models was fixed at 0. To capture variability in responses, both models assume truncated (at 0 and 1) normal distributed deviations from the model prediction.

A hierarchical mixture model mixing both components at the group level was fit to the data using JAGS through the

rJAGS package. In this model, each subject has some probability $\pi$ of producing split linear responses and some probability $1 - \pi$ of producing a log-linear responses. The model thus categorizes individuals as part of the fitting procedure. The model was simulated using MCMC, with 4 chains with 100 samples per chain, a burn-in of 30,000 iterations, and a thinning of 250 iterations per sample.

## Results

Figure 1 shows the fitted values of M. Qualitatively, nearly all participants were captured very well by the segmented linear model. The logarithmic model was selected as better fitting by the model for only one participant. Three other participants produced non-monotonic fits with wide variability, and were poorly fit by both models. The remaining 61 participants matched well the predictions of the segmented linear models. Figure 2 illustrates two typical patterns of response: one group of participants (n = 36) were fit very well by the linear model, and thus had low M values; the other group had high values of $\pi$ with typical M values centered around 0.4 (n=19). The few participants with intermediate M values (n=6) between 0.1 and 0.3 seemed to switch strategies, producing responses which were sometimes close to linear, and at other times very close to the uniform spacing model.

## Discussion

Experiment 1 demonstrates that there is not a general misunderstanding of large numbers, nor a logarithmic scaling of these numbers. Instead, a single, specific misconception of large numbers predominates errors: at least 85% of substantial deviations from linear responding involved a piecewise linear behavior, in which each of the ranges of "millions" and "billions" are linearly constructed, but are each of approximately identical size. Despite the prevalence of smooth, log-like functions in theories of economic and psychological utility functions and psychological magnitudes, evaluations of large numbers appear to no more than rarely approach logarithmic scaling.

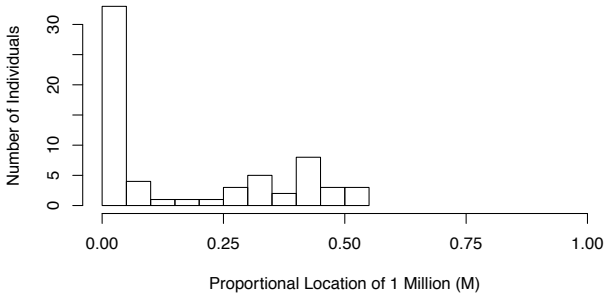


Figure 1: Histogram of individual fitted values of the position of 1 million (*M*). The normative value is 0.001.

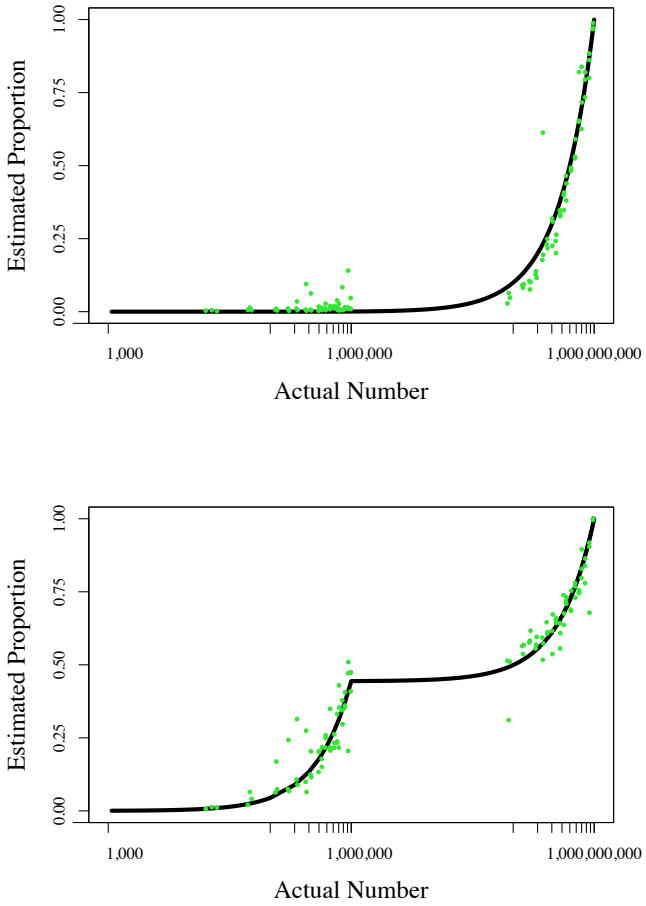


Figure 2: Number line estimates for a linear (top, M=.0004) and piecewise linear (bottom, M=.44) sample subject, along with predictions of the segmented linear model. The log-linear model predicts straight line responses on a log-scaled x-axis.

It is possible that participants in our study misconstrued the nature of the task, believing, for instance, that a segmented linear graph was requested. We believe this is unlikely for two reasons. First, although both linear and logarithmic number lines are fairly commonplace (for instance, as graph axes), segmented linear number lines—lines in which one linear number range lies adjacent to a linear range with a different unit—are vanishingly rare. Second, while piloting these materials we interviewed many individuals completing this task. While many made the error, none gave evidence having misunderstood the task. On the contrary, these individuals seemed very surprised when they realized or were told the normative location of one million.

# Experiment 2

The number line is an idiosyncratic task, involving visual and spatial components as well as number processing per se. It might be that the results of Experiment 1 result from idiosyncratic reasoning, and would not generalize well to

other kinds of number judgments. One purpose of Experiment 2, then, was to explore whether piecewise linear number line estimation would generalize to other tasks.

In studies involving smaller number ranges, learning of linear behavior can be strikingly sudden—with participants often becoming linear across an entire range from the presentation of just a single point (Opfer & Siegler, 2007). A second purpose of Experiment 2 was to explore whether similar approaches could lead to sudden reductions in misconceptions about large numbers, and such shifts in line estimation would generalize to evaluative judgments.

## Participants and Procedure

300 participants were recruited from Amazon's Mechanical Turk in exchange for small monetary remuneration. Mechanical Turk is a scalable workforce solution frequently used by psychologists to recruit subjects for online experiments (Mason & Suri, 2012). All tasks were completed remotely through a web interface.

Each participant first performed eight number line estimations (the *pretest*), followed by an intervention. Half of all participants saw an *encouragement* intervention, which simply thanked them for their hard work, and asked them to do their best on the rest of the experiment. The other half of participants saw a *training* screen, which reminded them that 1 billion was equal to 1,000 millions, and showed them the normative placement of 10 million on the number line from 1 thousand to 1 billion. Participants then completed eight more number line estimates (the *posttest*), followed by three situation evaluation questions.

In the *situation evaluation* task, participants read, in fixed order, three short narratives about how the governments of two fictional countries were dealing with various social challenges. 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". After both tasks were completed, participants reported their age, sex, and political affiliation, and briefly describing their problem-solving strategy. The strategy explanations provided an extra check that participants were in fact attempting the problems.

## Analysis and Results

Number Line Estimation. Estimates were modeled using a version of the model described in Experiment 1. Because the unimodal beta model at the family level did not capture the pattern of observed behaviors, in Experiment 2 data was fit only at the level of the individual. Further, the logarithmic model was not tested. Thus, the single model parameter was the estimated location of one million, *M*. Separate models were fit to the data before and after the intervention.

Figure 3 illustrates the shift in number line behavior before and after the intervention. An ANOVA evaluating M values as a dependent measure over time of estimation (pre vs. post intervention) and condition, indicated a significant interaction between the two ($F(1, 298)=15.8$, $p<.01$). There was also a main effect of condition ($F(1, 298)=4.4$, $p<.05$); considering only the pretest data, the difference was not significant ($F(1,298)=.21$, $p>0.5$).

As in Experiment 1, the empirical values of the M parameter were contrary to the predictions of the uniform spacing model. While the model predicts a mean value around 0.5 among the piecewise linear group, the actual mean fitted value was around 0.40.

**Situation Evaluations.** Evaluations were averaged across the three situations for analysis. These average responses were moderately normally distributed. An ANOVA of mean evaluation against pretest M and condition revealed significant effects of both ($F(1, 298)=11.3$, $p<0.01$, and $F(1, 298)=4.3$, $p<0.05$, respectively). Once behavior at posttest was included, however, it was the only significant predictor of situation evaluations ($F(1, 298)=15.8$, $p<0.001$; see Figure 4); condition was no longer significant ($F(1, 298)=2.4$, $p\sim.12$), suggesting that some of the effect of training on situation evaluation resulted from shifts in processes involved in number line estimation.

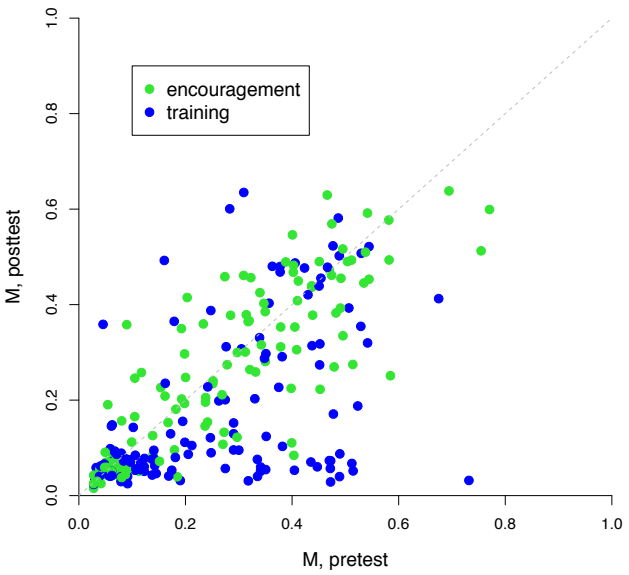


Figure 3: Best-estimated location of 1 million (M) at pretest and posttest. The normative location is 0.001. The large preponderance of blue circles in the bottom right represents the efficacy of the training.

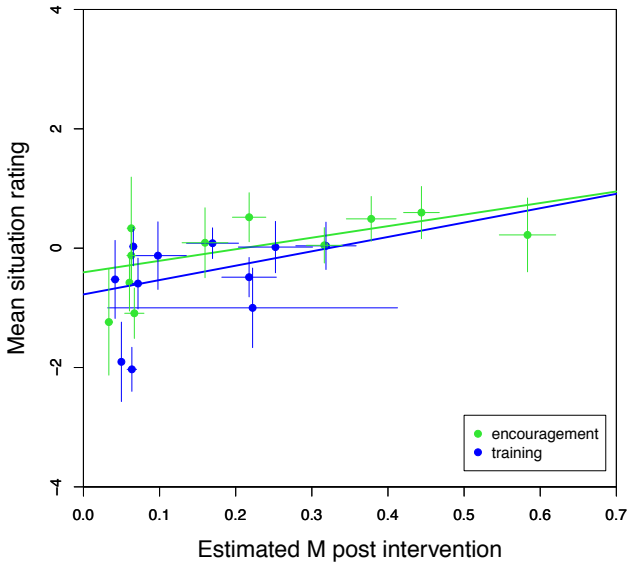


Figure 4: Situation evaluation against post-intervention placement of 1 million (M), binned into quantiles based on pre-intervention number line estimates. Errors reflect standard errors in the estimate of post-intervention values.

### Discussion

Experiment 2 demonstrated that the same strategies found in Experiment 1 are employed in a substantively different population: 45% of judgments were compatible with the piecewise linear account. However, participants were readily educable: just a single example of a normatively placed term sufficed to correct number line estimations in nearly half of error-prone estimators. Training seemed to be largely all-or-none: participants who shifted strategies halfway through the task gave responses nearly indistinguishable from those who had been following the final strategy from the beginning, judging both by number line estimations and situation evaluations.

Furthermore, Experiment 2 demonstrated that number line strategies are closely related to political judgments involving numbers in this range. People who estimate the number lines normatively are less optimistic about political situations involving numbers in this range. Furthermore, participants trained on the number line shifted their evaluations of political situations. Hence, the uniform spacing misconception does not result from reasoning specific to the number line task.

## General Discussion

Numbers picked out by the short scale—despite appearing frequently in educational contexts and public discourse—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 40 percent in the studies reported here—seems to evaluate large numbers based on the assumption that the number labels are roughly equally spaced as the numbers increase. Furthermore, people who rely on an equal spacing heuristic when placing numbers on a line are more satisfied with poor resolutions to political problems involving comparable scales.

Currently, the people of the United States, along with many other countries, are deciding how best to handle economic debt and deficit 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. This conversation is of direct relevance to the practice of scientific research, which is often funded by grants in the low millions of dollars. Detractors of the government spending on science research and other programs often present funding information by contextualizing these amounts within the overall budget using short-scale labels.

Logarithmic number line behavior was rare or non-existent on this task, despite substantial prior research that has supported the hypothesis that unfamiliar number ranges are initially represented logarithmically (Siegler & Opfer, 2003; Dehaene, 2003). One possibility is that large numbers fall beyond the upper range of the approximate magnitude system (Izard & Dehaene, 2008). Another possibility is that the reasoning processes we find adults employing when estimating large numbers account for apparently logarithmic behavior in young children (Nuerk et al, 2001).

Although the hypothesis that people infer spacing on the number line from the structure of the short scale labels predicted the basic pattern or responses, it does not predict the observed structure perfectly. In particular, most people who erred in their estimate of the relative values of 1 thousand, 1 million, and 1 billion did not put 1 million halfway between the other two, but substantially close to 1 thousand. Anecdotally, people we have observed often placed 1 million more or less exactly in the middle, then 'correct' to approximately the 40% mark. One possibility is that this positioning reflects a compromise between uniform spacing and normative number knowledge, but the nature of such a compromise remains speculative.

These results are striking in that the actual numerical system of short scale words and place value notation is formally extremely simple, and the referent system—the natural numbers—is acquired fairly early in mathematical development. A simple induction suffices to suggest the referents of the large number words studied here, rather than a conceptual restructuring, as has been implicated in rational-number learning. 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). 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. As

the four-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!" Magnitudes in this range are constructed by borrowing structure from the symbol systems used to represent them.

## Acknowledgments

Partial funds for this research came from an University of Richmond undergraduate research grant to the third author and Department of Education, Institute of Education Sciences grant R305A110060. . Thanks to Lisa Byrge, Iris Van Rooij, Erin Ottmar, and the Cave Lab for suggestions.

Experiment 1 is reported as Experiment 1 of Landy, Silbert, and Goldin (under review).

## References

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

Booth, J. L., a& Siegler, R. S. (2008) Numerical magnitude representations influence arithmetic learning. *Child Development, 79,* 1016-1031.

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.

Dehaene , S. (2003). The neural basis of the Weber-Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, *7*, 145-147. *Psychology*, *99*(1), 1–17.

Gilens, M. (2001). Political ignorance and collective policy preferences. *American Political Science Review, 95*(2), 379-396.

Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review*, *107*(3), 500-524

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

Kirsh, D. (2010). Thinking with external representations. *AI & Society*, *25*(4), 441-454.

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.

Mason, W., Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1-23.

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.

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

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.