Analogue Magnitudes and Knower-Levels: Re-Visiting the Variability Argument

James Negen and Barbara W. Sarnecka ({jnegen,sarnecka}@uci.edu)
Department of Cognitive Sciences, 2201 Social & Behavioral Sciences Gateway Building
Irvine, CA, 92697-5100 USA

Abstract

What cognitive system(s) initially provide the numerical content that defines the cardinal number words for young children? Le Corre and Carey (2007) argued that the answer cannot be the analogue magnitude system. Here we re-examine the most powerful of their arguments, which concerned the system's signature *scalar variability* (the standard deviation of answers grows linearly with the mean). Using adult data, we explore a nuance of this signature: that while it is certainly true of the continuous, underlying activation in the brain, it may not always be true of the number-word responses that people produce. With this in mind, we re-examine the aforementioned variability argument; contrary to Le Corre and Carey, we conclude that young children's estimates of small set sizes (up to and including their number-knower-level) *do* show scalar variability.

Keywords: Cognitive Science, Psychology, Cognitive Development, Perception, Bayesian Modeling, Developmental Experimentation, Human Experimentation

How children acquire the meaning of number words is a question of great interest in cognitive development. In infancy, the child can create exact representations only of very small sets (up to 3 items) and can create only approximate representations of larger sets (for review, see Feigenson, Dehaene, & Spelke, 2004). It is from these foundations that the child must build an understanding of integers.

There are at least two families of theories about the role that the analogue magnitude number system plays in number-word learning. One says that (a) the numerical content of number words initially comes from the analogue magnitude system (i.e., the representations of numerosity are only approximate) and (b) number-word meanings are learned in the same way as any other words for antecedently available percepts (e.g., Dehaene, 1997; Gelman & Galistell, 1978; 2004). On this view, the child essentially understands the logic of numbers from the beginning, but must learn the *exactness of number words*. That is, the child must learn that "seven" means exactly 7 – not approximately 5 to 9.

The other family of theories contend that (a) the numerical content initially comes from enriched parallel individuation representations, (which are exact, but cannot go higher than 3 or 4) and (b) number-word meanings are learned one at a time, and in order (e.g., Carey, 2009; Le Corre & Carey, 2007; Sarnecka & Lee, 2009; Lee & Sarnecka 2010). On this view, the child figures out the principle of cardinality and the logic of the integer system by extrapolating from the first three or four exemplars—the words *one, two, three* and possibly *four*.

Our proposal takes something from both views. We consider it adequately demonstrated that children do learn the first few cardinal number-word-meanings one at a time, and in order (Carey, 2009; Le Corre, Van de Walle, Brannon & Carey, 2006; Lee & Sarnecka, 2010; Sarnecka & Lee, 2009; Wynn, 1990, 1992). Thus, in this paper we will use the knower-levels framework, wherein a *one-knower* is a child who just knows "one", a *two-knower* knows "one" and "two", etc., and a *CP-knower* understands how counting works and what it is used for. Though this view has historically been paired with the view that the numerical content of these words comes exclusively from enriched parallel-individuation representations, that linkage is not logically necessary.

This issue has been examined before by Le Corre and Carey (2007), who made a forceful argument against the analogue magnitude system being involved in number-word learning. The argument was based on variability signatures: Noise in analogue magnitude representations grows at the same rate as the number of objects being estimated. In other words, if you briefly show people 10 dots, and ask them how many are there, they will respond with variability σ. If you then show them 20 dots, they will respond with variability 2σ. More formally, the standard deviation of a person's responses, divided by the mean of their responses, will be a constant (the 'coefficeint of variance', hereafter COV). This *scalar variability* is a defining characteristic of the analogue magnitude system throughout the lifespan.

Le Corre and Carey (2007) argued that children's estimates of the number of items in a picture did not show scalar variability. Le Corre and Carey used a Give-N task (where children give a certain number of items to the experimenter) as well as a Cards task (where children quickly estimate the number of items on a card). Le Corre and Carey first identified children who were one-, two-, three-, or four-knowers as measured by the Give-N task. Then they analyzed the variability of children's estimates (on the Cards task) for sets of 1, 2, 3 and 4 items. The mean COV grew from 0 (for estimates of 1 item) to 0.4 (for estimates of 4 items), leading Le Corre and Carey to conclude that early on, the cardinal meanings of these words are not defined in terms of the analogue magnitude system.

However, there is a problem with this argument. Logically, scalar variability could be found in two places: in the underlying activations in the brain, or in the distribution of number-word responses. Most studies of the analogue magnitude system in adults treat these as the same. That is, they assume that number-word responses transparently

reveal the underlying activations. This is a reasonable assumption when the numbers involved are large. But with small numbers, the available words (i.e., *one*, *two*, *three* and *four*) may not be sufficient to express the variation in the underlying activations. Below, we describe these ideas in more detail, then support our argument with some new data from adults.

Scalar Variability of Activations versus Responses

Creating and using an analogue magnitude representation is a multi-step process. One place that scalar variability might be found is in the actual number-word responses that people produce. For example, if you show someone 10 and 20 dots in alternation, over and over again, and they say something like: 9, 22, 11, 18, … their responses could show scalar variability. That is, the standard deviation of the responses could grow linearly with the number of items to be estimated.

Another place you could look for scalar variability would be in a latent, real-valued variable – possibly the logarithm of activation in certain 'number neurons' – that describes the person's perception of how many dots are present. Thus, when shown 10 and 20 dots, the person may experience something like 8.82, 21.58, 11.21 and 18.44 units of activation. But of course, participants do not say that there are 11.21 dots on the screen – they just say "eleven"[1]. Scalar variability of the activations means that the standard deviation of this latent variable (as opposed to the responses) grows linearly with the number of dots shown.

For large numbers, these two ideas make almost exactly the same predictions. However, the predictions differ for smaller numbers. Imagine that a person says "four" when shown four dots, on a large number of trials. The variability in responses is zero, but the variability in activations could be much larger. The person might be experiencing anywhere from 3.5 to 4.5 units of activation on each trial. If we add the constraint that activation must be normally distributed, then the continuous standard deviation could be as large as 0.1 while the discrete standard deviation stays very near zero (out to several decimal places).

An analogy may be helpful: Imagine you are about to play a game with a friend. You will see a number of dots on a screen. Your friend has to say how many there are, but can't see the screen. You have to tell your friend how many dots there are, but you're not allowed to speak – only to draw a single line on a piece of paper. Your partner can then measure the line accurately with a ruler. Before the game starts, you agree on a simple code: 1 inch = 1 item. It's easy to see that the average amount of error in your line length will be proportional to the number of items.

Let's imagine that, when you see 2 items, all of your lines fall between 1.25 and 2.75 inches; you are accurate to ± .75 inches. Now cut that in half for when you see 1 item, and

you are accurate to ± .375 inches. As such, your line lengths show scalar variability. But what about the answers your partner gives? There will be some variation at 2 (when she rounds 2.75 inches to 3, for example) but none at 1 (where all lines fall between .625 and 1.375 inches, and so are rounded to 1). The zero variation in her responses for trials of 1 cannot be half of the above-zero variation for trials of 2.

Thus, the length of your lines can show scalar variability while the number-word answers after measuring do not. This happens whenever the variation is small compared to the minimum distance between values after rounding. In more traditional statistical language, the variation is very fine-grained and the rounding is very coarse-grained, leading to the phenomenon of *heaping*: all of the data gets heaped onto one point.

Le Corre and Carey (2007) argued that one could expect to see scalar variability in the verbal responses all the way down to cards with 1 item. This claim was based on work by Cordes, Gelman, Galistell, and Whalen (2001). In that paper, the experimenters had instructed participants to tap a space bar a certain number of times in a steady rhythm, while saying "the" with each press to suppress counting. Cordes et al. found that participants' verbal responses showed scalar variability for numbers down to 2. However, they did not test on the number 1, and participants showed a relatively high amount of noise in their answers (a standard deviation of around .4 at minimum). As such, Cordes et al's data don't tell us where the variability was located: at such high levels of noise, the two sources of variability make much the same predictions.

The Present Study

In experiment 1, we examine the two sources of variability by extending the experiment by Cordes et al. (2001). Like them, we showed adult participants a numeral on a screen, and the participants' task was to tap the space bar that number of times. They had to do this in a steady rhythm while repeating "the" with each tap. We made one modification to reduce the amount of noise in the data: participants' responses were slowed to 2 Hz by having them tap along with a metronome.
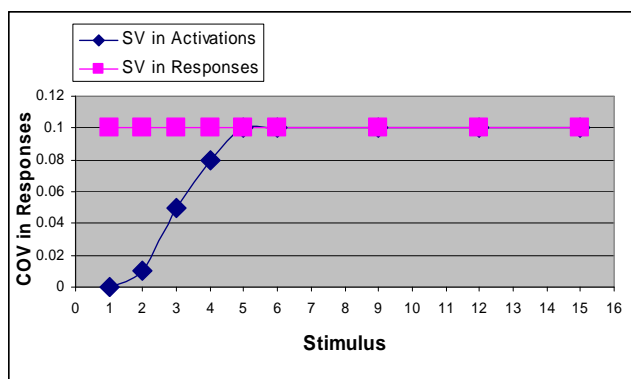

Figure 1: Different behavioral predictions for scalar variability in activations vs. responses.

---

[1] It may be more satisfying to imagine that this is the center of a confidence distribution. In other words, the participant would experience 11.21 units of activation, and be most confident in the integer 11 (which is closest), followed by 12, then 10, then 13, etc.

The two predictions from the two sources of scalar variability are illustrated in Figure 1. Conveniently, the pattern predicted by scalar variability in the responses is identical to the null hypothesis for a repeated-measures ANOVA.

In Experiment 1, we show that small-number estimation data from adults is consistent with the assumption of scalar variability in the activations. In experiment 2, we develop a simple model that incorporates number-knower-levels and the analogue magnitude system, and we test it using data from children's performance on the Give-N and Cards tasks. We find that the model provides a good fit to these child data, with no problems of variability.

Experiment 1

Methods

Participants Eleven undergraduates were recruited from the University of California, Irvine and successfully completed the task. Two additional participants were excluded because they could not keep a steady rhythm during training trials. Participants were each given a half point of extra credit in an introductory Psychology course.

Procedure On each trial, participants saw a numeral on the screen. They were asked to tap the space bar that many times, along with a 2Hz metronome, saying "the" with each tap to suppress counting. Participants were asked to avoid trying alternate strategies, like chunking, and to just keep tapping until they felt the number had been reached. The numerals used were 1, 2, 3, 4, 5, 6, 9, 12, and 15. For training, there were two trials with each number, and all participants declined the offer to go through training again. For data recording, there were 40 trials with the number 1, and 20 trials with each of the other numbers, for a total of 200 trials. Participants were allowed as many brief breaks as they wanted, though none took more than two. All participants finished the testing session in under 20 minutes.

Results and Discussion

All but two participants were able to keep a steady rhythm throughout the experiment (the two who couldn't were excluded). The mean coefficient of variance (COV) for each stimulus is shown in Figure 2. A repeated-measures ANOVA with the Greenhouse-Geisser correction suggests that the mean COV across stimuli is not constant, $F(2.8, 28.2) = 8.16$, $p = .001$. This discredits the idea of scalar variability in responses: if such variability were true of adult performance, then we should expect the mean COV in responses to be the same for all the numbers we tested.

On the other hand, these data are very consistent with the assumption of scalar variability of activations, with a correlation between prediction and observation of over $r = .8$. There appears to be a section where the rounding-off occurs (numbers 1 to 3), and a section with full observed variance (numbers 5 to 15), as predicted. It's a little surprising that there was so much variability in responses to

1, but this variability comes from only 2 of the participants, representing 3 errors out of 440 trials, and still appears to be lower than the section with full observed variance. It is clear that these data are better explained by scalar variability in the activations than in the responses.
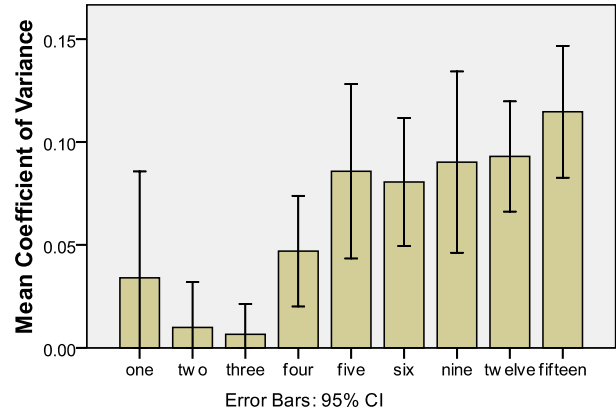

Figure 2: Observed COVs (standard deviation divided by mean) in the responses were lower on average for lower numbers of requested taps, consistent with scalar variability in the activations (but not in the responses).

The amount of variability at numbers 1 to 4 suggests that we were successful in repressing subitization, replicating Cordes et al., 2001. Also, subitizing is usually characterized by a hard limit at 4, but the lowest number where each participant made a mistake ranged from 1 to 6. In addition, if the effect was due to subitizing, we would be able to eliminate it by subtracting out all of the jumps from 4 to 5 (across the limit of what people can subitize). However, this is not the case, $F(7,4) = 7.546$, $p = .035$. Finally, all participants reported that they found it impossible to track individual taps during the task. (Tracking individual taps would show that participants were using the parallel individuation system). All of these facts point away from subitizing as an explanation.

Next, we test these predictions against a corpus of child behavioral data, to re-examine the question of whether the analogue magnitude number system may be involved in early number-word learning.

Experiment 2

We made a simple model of how children would solve the Cards task using the analogue magnitude system. In this task, children are shown a certain number of items on a card and are asked to guess the number of items. This is the same as Le Corre and Carey's (2007) Cards task, but without the time limit. (The purpose of the time limit is to prevent counting, but children who do not yet understand the cardinality principle do not use counting to answer anyway.)

This is how the model works: the child looks at the card once and forms an analogue magnitude representation of the number of items present. This representation could be measured in units of activation. The child then rounds off

this activation level to the nearest positive (non-zero) whole number. If the child knows this number word, she says it. If the child does not know the word for that number, she says a number from among the number words whose meanings she does not know. (For a discussion of how the child chooses *which* undefined number to guess, see Sarnecka & Lee, 2009; Lee & Sarnecka 2010.)

For example, suppose we show a two-knower 2 items on several trials. On the first, she experiences 2.3 units of activation, rounds it to 2, and says "two". On the second, she experiences .38 units, rounds it to 1, and says "one". On the third, she experiences 3.3 units, rounds to 3, and picks "six" from among the words that she does not know ("three" and up). These activation levels are being drawn from a normal distribution with a mean of 2.

The computations involved are based on the normal cumulative density function. Say that this child has a COV of α when shown a card with 2 items on it. The chances of saying "one" are $\Phi((1.5-2)/\alpha)$. The chances of saying "two" are $\Phi((2.5-2)/\alpha)-\Phi((1.5-2)/\alpha)$. The chances of saying "three" and up (beyond the knower-level) are proportional to $1-\Phi((2.5-2)/\alpha)$. The lower α is, the more often the child will be correct.
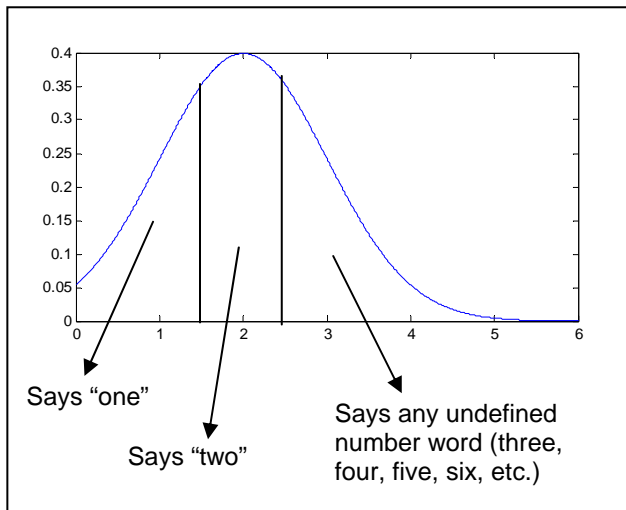

Figure 3: An example of the model: a two-knower, shown 2 items, draws an activation level from a normal distribution N(2,1) and responds accordingly.

Furthermore, we assume that for each stimulus γ (each number of items shown on the card, up to 4) and knower-level τ, a particular child's COV is drawn from a normal distribution with mean $\mu_{\gamma,\tau}$ and standard deviation σ. A variable δ describes the constant difference between each μ_γ and the next one up. So $\mu_{2,2} = \mu_{1,2} + \delta$, etc. We planned to use Bayesian inference on these latent variables: if $\delta \neq 0$, then this method of describing variation does not show scalar variability – i.e., a constant COV – and our model is wrong.

It may bother the reader that this model allows for the analogue magnitude system to output a negative number.

However, there is no empirical argument against this– assuming the participant knows to round to the nearest positive whole number. A log-normal distribution may be substituted, which would limit the outputs to positive numbers, but this generally has little effect; having low outputs rounded to 1 accomplishes the same basic goal as making the distribution of outputs stay above zero. Since a normal distribution is symmetric, we assume that the variation to the higher side is sufficient to estimate total variation, e.g., the variation around 1 item is well-represented by the number of times the child says "two" or "three". Put another way, a two-knower has an equal probability of saying "two" as experiencing between -.5 and .5 units of activation, when shown 1 item. In any case, the probability of negative activation is often very small.

Methods

These data were taken from a longitudinal study testing children on both the Give-N task and the Cards task. The dataset thus offers a way of independently assessing the child's knower-level (using Give-N) and then estimating δ (using the Cards task).

Participants A total of 97 monolingual English-speaking children participated. Children were tested once every two weeks for twenty weeks, resulting in a total of 454 sessions. Because little is known about the week-to-week consistency of children's knower-levels and/or analogue magnitude acuity, each session was modeled as a new child. We included all children who were two-, three-, or four-knowers (as determined by Give-N), and who completed at least 15 trials of both the Give-N and the Cards task. This resulted in a total of 161 sessions (45 sessions with two-knowers, 51 with three-knowers, and 65 with four-knowers).

Give-N The purpose of this task was to determine what number-word meanings each child knew (i.e., to determine the child's number-knower-level.) The experimenter began the game by bringing out a stuffed animal (e.g., a lion), a plate, and 3 bowls, each containing 15 small identical rubber toys (e.g., toy bananas, approx. 3 cm long). The experimenter said to the child, "In this game, you're going to give something to the lion, like this [experimenter pantomimes putting an item on the plate and sliding it over to the lion]. I'm going to tell you what to give him." Instructions were of the form, "Can you give the lion TWO bananas?"

Trials were in pseudorandom order, always starting with a request for one item. There were a total of eighteen trials: six trials each asking for one, two, three, four, six, and eight items. Children were given generalized positive feedback after each trial (e.g., "Thank you!"), regardless of their responses.

The child's knower-level was estimated using Lee & Sarnecka's (2010) model. In this model, the child has a prior probability of giving each number of objects: small handfuls are very likely, as is giving the entire bucket. When the

child is asked for a specific number, the response probabilities are updated according to her knower-level. Every child is given a uniform prior chance to be each knower-level, and a posterior distribution is calculated based on the data. We sampled over this posterior distribution with 3 independent chains, 2,000 burn-in samples and 10,000 data collection samples. We then assigned each child's knower-level based on the expected value[2] from this posterior distribution.

Cards Task This was our comparison task, which provides a set of responses independent from Give-N, so that one can be used to determine knower-level and the other to model the analogue magnitude response. The stimuli were photographs of the same stimuli used in Give-N, on a white background. The difference is that children were asked to label the numerosity instead of generating it. For example, if a child was asked for three red trains in Give-N, she was later shown a picture of three red trains and asked the number on it. As a check that the child was paying attention, trials were thrown out if the child did not produce the correct object name (e.g., trains) along with the number word. Before the first trial, there was a check that the children knew the words for the objects being used, which did not seem to pose a problem for our participants.

Again there were eighteen trials: three trials each for one, two, three, four, six and eight items, in pseudorandom order. We asked children the question "What's on this card?" because questions that start with "How many …" are often interpreted as commands to count (Sarnecka & Carey, 2008). Children were given generalized positive feedback after every trial. Order of tasks was counterbalanced across sessions.

Results and Discussion

There were a total of 1,903 usable trials from the Cards Task. We decided to give the model a unit normal prior on effect size δ/σ, as a way to make the prior dimensionless and reasonably vague (Jeffreys, 1961; Rouder, Speckman, Sun, Morey & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, in press). We then ran the model with MATLAB and WinBUGS, which uses a form of Gibbs sampling to describe the posterior distribution. We ran five independent chains, with 10,000 burn-in samples and 25,000 collection samples. The within-chain variability matched the between-chain variability, even though the chains were initialized with different effect sizes, which is a good indication of proper convergence.

At the point of the null hypothesis ($\delta = 0$; i.e., the average COV is the same for every number of items shown), there was a prior density of .39. At the same point, there was a posterior density of 5.72, estimated with a normal kernel density method. By the Savage-Dickey Theorem (Dickey &

Lientz, 1970), the Bayes factor is 5.72/.39 = 14.34, meaning that the data were 14.34 times more likely to be generated by the null hypothesis than the alternate. This is very strong evidence in support of the null hypothesis. In other words, Children *do* show scalar variability in at least one task that taps number-word knowledge – but it is scalar variability of activations, not of responses.
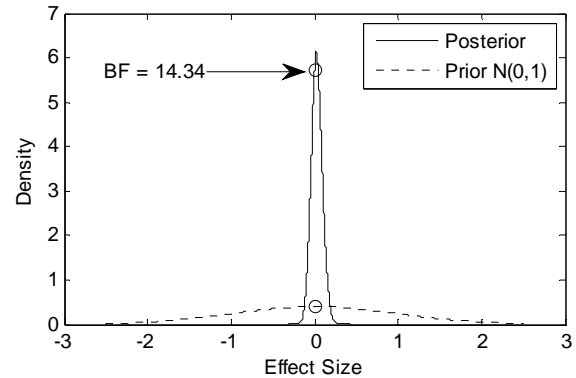

Figure 4: Prior and posterior distributions of δ/σ. There is a great deal of support for the null hypothesis ($\delta = 0$; i.e., the mean COV for every stimulus is the same within each knower-level) in the child data.

General Discussion

In Experiment 1, we showed that scalar variability of adult small-number estimates may be located in the activations, but not in the responses. Our analysis suggests that scalar variability is not always present in the answers that people give when you ask them to estimate a number of items. Instead, scalar variability is present in the latent continuous variable that serves as input, and which gets rounded off in people's answers. In Experiment 2, we modeled how children respond when asked how many items are on a card. Our model included (a) the use of analogue magnitudes assuming scalar variability in the activations and (b) knower-levels, which are stages of number-word knowledge (e.g., Sarnecka & Lee, 2009). We showed that, contrary to previous reports, there is no problem fitting the well-known signature of scalar variability to real child data. In fact, we found strong evidence that children's responses *do* show scalar variability, with a Bayes factor of 14.34.

This reverses one of the most powerful arguments against the involvement of the analogue magnitude system in early number-word learning. Scalar variability is a very reliable feature of the analogue magnitude system. We have presented evidence that this signature is found in children's verbal responses to a small-number estimation task, even before the children have figured out the cardinal principle of counting. The statistical method we used has a very useful feature: it can quantify support for the null hypothesis, rather than simply rejecting or failing to reject it. Thus, we can report positive evidence that children's responses do show scalar variability. This suggests that children do imbue the first few number words (up to and including their

---

[2] Another reasonable method would be to take the mode instead of the expected value. This method has high agreement in this case, r = .992.

knower-level) with numerical content from the analogue-magnitude system.

Of course, evidence for the involvement of analogue magnitudes is not evidence against the involvement of enriched parallel individuation; it's possible that a parallel-individuation-based model could also be fit sensibly to these data. There are still many reasons to think that parallel individuation is involved. In particular, we do not dispute Le Corre and Carey's finding that new CP-knowers do not connect numbers above 5 to the analogue magnitude system (Le Corre & Carey, 2007; see Sarnecka & Lee, 2009 for a convergent finding). In addition, children seem to become three-knowers or four-knowers, but not five-knowers or six-knowers. The limit after 4 coincides with the set-size limit on the parallel individuation system (e.g., Feigenson & Carey, 2003).

Moving forward, we think tasks that tap more than just the child's number-word knowledge should be part of the debate. For example, there is new interest in the question of whether number-word knowledge is related to estimation acuity, after it was shown that acuity at age 14 retroactively predicts math grades back to first grade (Halberda, Mazzocco & Feigenson, 2008). In particular, the present model predicts that estimation acuity should correlate with the number of within-knower-level errors on the Cards task. This kind of argument could be very useful in determining which systems are important to number-word learning and how they contribute.

We hope that this line of work will lead to increasingly accurate descriptions of how children acquire integers, and will help to, resolve debates over the roles of various cognitive systems in number-word learning. The test case of number may then inform more general theories of how new representational resources are acquired.

Acknowledgments

References

Carey, S. (2009). *The Origin of Concepts*. Boston: MIT Press.

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal and nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review, 8*(4), 698-707.

Dehaene, S. (1997). *The Number Sense*. New York: Oxford University Press.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics, 42,* 204-223.

Feigenson, L., & Carey, S. (2003). Tracking individuals via object files: Evidence from infants' manual search. *Developmental Science, 6*(5), 568-584.

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

Gelman, R. & Gallistel, C. R. (1978). *The Child's Understanding of Number*. Oxford: Harvard University Press.

Gelman, R. & Gallistel, C. R. (2004) Language and the origin of numerical concepts. *Science, 306*(5695), 441-443.

Halberda, J., Mazzocco, M.M.M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455*(2), 665-669.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Le Corre, M., Van de Walle, G., Brannon, E. M., Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive Psychology, 52*(2), 130-169.

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

Lee, M.D., & Sarnecka, B.W. (2010). A model of knower-level behavior in number-concept development. *Cognitive Science, 34,* 51-67

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225-237.

Sarnecka, B. W. & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108,* 662-674.

Sarnecka, B. W. & Lee, M. D. (2009). Levels of Number Knowledge in Early Childhood. *Journal of Experimental Child Psychology, 103*(3), 325-337.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (in press). How to quantify support for and against the null hypothesis: A fexible WinBUGS implementation of a default Bayesian t-test. *Psychonomic Bulletin & Review*.

Wynn, K. (1990). Children's understanding of counting. *Cognition, 36*(2), 155-193.

Wynn, K. (1992). Children's acquisition of number words and the counting system. *Cognitive Psychology, 24*(2), 220-251.