Available online at www.sciencedirect.com
ScienceDirect
COGNITION
Cognition 106 (2008) 1221–1247
www.elsevier.com/locate/COGNIT

Calibrating the mental number line ☆

Véronique Izard [a,b,c,*], Stanislas Dehaene [b,c,d]

[a] *Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA*
[b] *INSERM, U562, Cognitive Neuroimaging Unit, F-91191 Gif/Yvette, France*
[c] *CEA, DSV/I2BM, NeuroSpin Center, F-91191 Gif/Yvette, France*
[d] *Collège de France, F-75005 Paris, France*

Received 21 August 2006; revised 30 May 2007; accepted 1 June 2007

Abstract

Human adults are thought to possess two dissociable systems to represent numbers: an approximate quantity system akin to a mental number line, and a verbal system capable of representing numbers exactly. Here, we study the interface between these two systems using an estimation task. Observers were asked to estimate the approximate numerosity of dot arrays. We show that, in the absence of calibration, estimates are largely inaccurate: responses increase monotonically with numerosity, but underestimate the actual numerosity. However, insertion of a few inducer trials, in which participants are explicitly (and sometimes misleadingly) told that a given display contains 30 dots, is sufficient to calibrate their estimates on the whole range of stimuli. Based on these empirical results, we develop a model of the mapping between the numerical symbols and the representations of numerosity on the number line.

Keywords: Numerical cognition; Estimation; Modeling

☆ This manuscript was accepted under the editorship of Jacques Mehler.

* Corresponding author. Address: Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA. Tel.: +1 617 384 7900.
E-mail address: veronique.izard@m4x.org (V. Izard).

0010-0277/$ - see front matter
doi:10.1016/j.cognition.2007.06.004

1. Introduction

Even in the absence of symbols and language, infants and animals are able to perceive the number of items in a set and engage in simple calculations (Brannon & Terrace, 2000; Flombaum et al., 2005; Hauser, Tsao, Garcia, & Spelke, 2003; McCrink & Wynn, 2004; Wynn, 1992; Xu & Spelke, 2000). It is thought that they use a non-verbal system (*number sense*) which represents numbers analogically and approximately (Dantzig, 1967; Dehaene, 1997). Human adults exhibit the same behavioral characteristics as infants and animals when confronted with non-symbolic stimuli (Barth, Kanwisher, & Spelke, 2003; Barth et al., 2006; Cordes, Gelman, & Gallistel, 2001; Whalen, Gallistel, & Gelman, 1999). For example, when adults, infants or animals have to discriminate numerosities, their performance improves when the distance between the numerosities to be discriminated increases, and more precisely follows the Weber's law: namely, the extent to which two stimuli can be discriminated is determined by their ratio (Cantlon & Brannon, 2006; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Pica, Lemer, Izard, & Dehaene, 2004). Adults may still rely on their number sense to solve certain tasks involving symbolic stimuli (Dehaene & Marques, 2002; Gallistel & Gelman, 2005; Moyer & Landauer, 1967; Spelke & Tsivkin, 2001). This implies the existence of an interface between the system of verbal numerals and the non-verbal analog representations of numerosity.

However, the question of how this interface works has not received much attention. Suppose that one is presented with a large set of dots and is asked to estimate verbally how many dots are present. Using the analog numerosity representation alone, one might know that the quantity is 30% larger than on the previous trials, but how does one know that there are around 20, 40 or 80 dots? The analog code must be calibrated before a given quantity can be named.

Previous studies on numerosity estimation indicate that spontaneous estimates are poorly calibrated: for example, Minturn and Reese (1951) report responses diverging from the true numerosity by a factor 4 (e.g., response 50–700 for a stimulus containing 200 dots). Even if overestimated responses can be observed in some participants, most of the studies report a tendency to underestimate the actual value (Indow & Ida, 1977; Krueger, 1982, 1984). The amount of underestimation increases in the course of the experiments, but is also present from the very first trial (Krueger, 1982). Estimates are modulated by the non-numerical parameters of the stimuli, as more dense dot arrays tend to be more underestimated (Hollingsworth, Simmons, Coates, & Cross, 1991), and arrays are judged to contain more dots when dots are regularly spaced (Ginsburg, 1978). Furthermore, the estimation pattern seems to be influenced by the range of stimuli tested. For example, Hollingsworth et al. tested a wide range of values and found that participants overestimated the least numerous arrays (less than 130 dots), and underestimated the most numerous ones (containing up to 650 dots). Arrays of less than 130 dots, which were overestimated in this last study, have been reported to be underestimated elsewhere, when the range of stimuli tested was more narrow. This possible range effect illustrate the general non-linear shape of the response function, which has been described as a power function (Indow & Ida, 1977; Krueger, 1972, 1982, 1984, 1989; but see Masin, 1983 for an alternative view).

Besides this general tendency to underestimate, responses are also very variable, both within and between participants, and follow a similar pattern of variability in both cases: the larger the numerosity to be estimated, the more variable the responses. Krueger (1982) computed mean estimates of numerosity for several participants, and reported a larger dispersion of these mean responses across participants for large numerosities than for small ones. Focusing on the variability of the responses within each participant, Logie and Baddeley (1987) observed the same phenomenon: responses of each participant become more variable as numerosity increases. The law describing the variability of responses was characterized later and called *property of scalar variability* (Whalen et al., 1999): the mean responses and standard deviation of the responses are actually proportional to each other as the numerosity to be estimated varies, hence the coefficient of variation ($CV = \frac{\text{standard deviation}}{\text{mean}}$) is constant across all numerosities. The property of scalar variability appears to be a fundamental signature of numerical estimation tasks. It has been observed not only in humans, but also in animals, such as rats, when required to produce a certain number of presses on a lever (Gallistel & Gelman, 2000). In humans this property has been identified in various estimation contexts, involving non-symbolic stimuli (Whalen et al., 1999) as well as symbolic stimuli: in a price estimation task, where both encoded numbers and responses were given in symbolic format, Dehaene and Marques (2002) showed that the responses still follow the property of scalar variability.

However, it has been noted that providing participants with some information about the numerosities tested could reduce the amount of variability, at least between participants, and also improved the accuracy of the responses. At the mid-course of one of his experiments, Krueger (1984) showed to the participants an example of an array containing 200 dots. This information considerably reduced the level of variability between participants in subsequent blocks of trials. Lipton and Spelke (2005) designed a simplified estimation task, where participants were systematically shown two arrays at each trial, were told how many dots the first one contained, and were asked to estimate the numerosity of the second array. Average responses matched perfectly the numerosity of the second array in this task, in adults as well as in children aged 5 years selected for their proficiency in counting. Even less explicit information seems sufficient to increase accuracy in the estimates: for example, Whalen et al. (1999) had participants perform an estimation task after they did a numerosity production task, where they were explicitly told which numbers to produce. The same range of numbers were used in the stimuli of both experiments. Responses in this task showed a limited amount of underestimation, certainly because participants had inferred the range of numbers from the previous experiments.

These observations suggest that it should be possible to calibrate the mapping between the analog representations of numerosity and the verbal numerical labels, and enhance participants' accuracy in estimation tasks. However, this calibration process has not been systematically investigated to date. How fast does calibration occur? How much information is needed? Furthermore, would calibration be a global or a local process? One possibility is that the feedback given for one numerosity would exert only a local influence, modifying the mapping only within a small interval around the reference numerosity. Alternatively, calibration could be a global

process, in the sense that giving feedback on only one numerosity would affect the estimates for all numerosities.

Here, we use the numerosity estimation task to study the structure of the interface between the non-verbal and verbal numerical representation systems. Participants were asked to estimate the quantity of dots in rapidly flashed arrays. In the absence of calibration, estimates were inaccurate, although they consistently increased with numerosity. A calibrated estimation task was then used to investigate how calibration affects the estimates on a wide range of test numerosities. Before they gave numerical estimates, participants were presented with one inducer array, which they were told contained 30 dots, but in reality contained either 25 (overestimated inducer), 30 (exact inducer) or 39 dots (underestimated inducer). Participants were able to calibrate their responses on this inducer trial, even after a very short learning period, and their responses were calibrated not only for the numerosity showed as reference, but for all the numerosities tested.

Our procedure is deeply related to the classical magnitude estimation procedure (Stevens, 1957). The magnitude estimation procedure has been introduced to explore internal psychophysical continua such as brightness, loudness, pain or even abstract sensations such as anxiety. In most cases, participants are given a reference stimuli, and told to associate this reference with a given number, e.g., 100. Sometimes, participants have to find their reference by themselves, a situation close to our non-calibrated estimation experiment. Then, they are presented with different stimuli varying along the continuum of interest, and asked to use numerical responses to rate their perception of these stimuli with respect to the reference stimulus. Just as our procedure, the magnitude estimation method involves the definition of a numerical response scale. However, in magnitude estimation, the response units can be arbitrarily scaled, although in our numerosity estimation task, there is an objectively true answer, and participants are trying to approach this answer. Both numerosity estimation studies and magnitude estimation studies typically report that the participants' responses vary monotonically, as a power law of the input stimulus (for a review, see Krueger, 1989).

2. Experiment 1: Non-calibrated estimation

2.1. Introduction

In order to later evaluate the effects of calibration, we first studied how participants estimated the numerosity of our stimuli in the absence of any indication of number by the experimenter. In this first experiment, the instructions given ensured that participants would not be able to infer the range of numerosities presented.

2.2. Experimental methods

2.2.1. Stimuli

Dot arrays containing 1–100 dots were flashed on a computer screen for a duration of 100 ms. Stimuli were yellow on a black background. Arrays covered a visual angle of

11.5°, with a minimal visual angle of 0.3° for each dot. To prevent observers from using strategies based on low-level continuous variables, we generated three sets of arrays controlling for total luminance, size of dots, density of the array and occupied area on the screen (Fig. 1). In the first set, the sum of the area of all the dots (total luminance of the array) as well as occupied area on the screen were kept constant over all numerosities. Thus in this set, the size of the dots decreased with increasing numerosities, and the density of the arrays increased with numerosity. In the second set, the size of the dots was held constant, along with total occupied area. As a consequence total

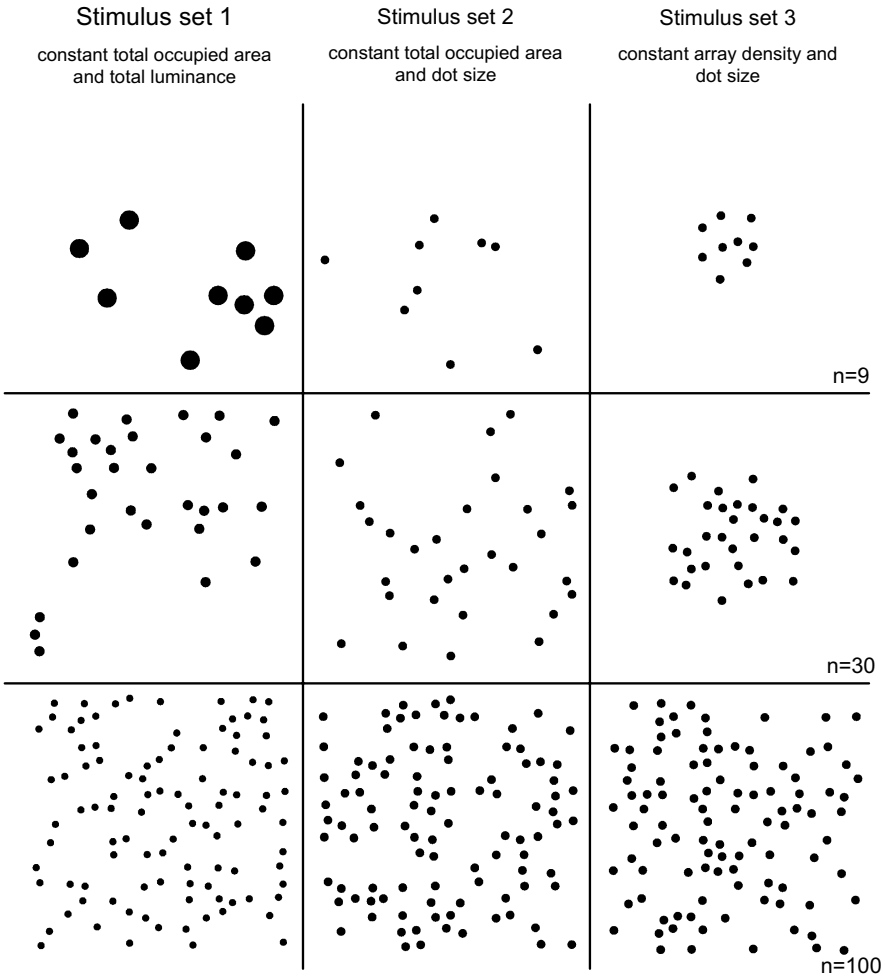

Fig. 1. Examples of stimuli of numerosity 9, 30 and 100 belonging to the 3 stimulus sets. Stimuli were originally yellow on a black background. Stimulus set 1: total occupied area and total luminance are constant, thus dot size decreases with numerosity and array density increases with numerosity. Stimulus set 2: total occupied area and dot size are constant, total luminance and array density increase with numerosity. Stimulus set 3: array density and dot size are constant, total luminance and total occupied area increase with numerosity. In the three sets the mean dot size and mean total luminance are equal. Mean density is about twice larger in the 3rd set. Consequently, total occupied area is about twice smaller in the 3rd set.

luminance and density increased with number. Finally, in the third set, size of the dots was constant as well as density of the array. Total luminance and total occupied area thus increased with number. The mean total luminance and the mean size of dots were kept constant over all sets. However, to keep the size of dots within a perceptible range, we had to choose a different mean density value as well as mean occupied area in the third set than in the first and second sets.

Once the four parameters had been determined with the above constraints, a square grid of positions was generated from their values. To draw the array, the software chose at random a set of n positions on the grid and added to them a random jitter to make the grid non-obvious to the participants. Density was defined as the proportion of occupied positions on this grid divided by the total number of available positions. For each numerosity, we presented one stimulus from each set. By doing this, we were able to detect whether participants were using a strategy relying on one of the low-level variables controlled. Indeed, if a participant used such a strategy, its responses would not vary with numerosity in at least one of the three stimulus sets.

2.2.2. Procedure

Participants were required to estimate the numerosity of the arrays. Responses were unrestricted, since giving a set of restricted responses would necessarily have provided information about the magnitude of the numbers involved. Participants were instructed to provide the number that seemed the most adequate to them, both in terms of its magnitude, and of its level of precision. Instructions did not include any numerical value. Participants typed their responses in Arabic format on the numerical pad of the keyboard.

Because a high level of variability was expected across participants, we elected to collect a high number of trials per participant. Each of them participated in five sessions of 600 trials each, distributed over 12 blocks of 50 trials. Each session lasted approximately one hour and the five sessions were performed over five consecutive days. During each session, each numerosity ranging from 1 to 100 was presented 6 times, with two arrays belonging to each stimulus set.

2.2.3. Participants

Five french speakers (3 males, 2 females; aged 23–26) participated in the first experiment for a payment. All of them were naive to the experimental conditions. They all performed five whole sessions, but the data of two sessions of subject ML were lost.

2.3. Results

All the participants strongly underestimated the numerosity of the arrays (Fig. 2). However, their responses remained consistent: for all the participants, the responses increased monotonically with numerosity over the whole range of numerosities considered (linear regressions between mean responses and numerosity for each participant: slopes ranged from 0.20 to 0.43, differing significantly from 0 for all the participants, all ps < 0.0001). While all the participants underestimated the numeros-

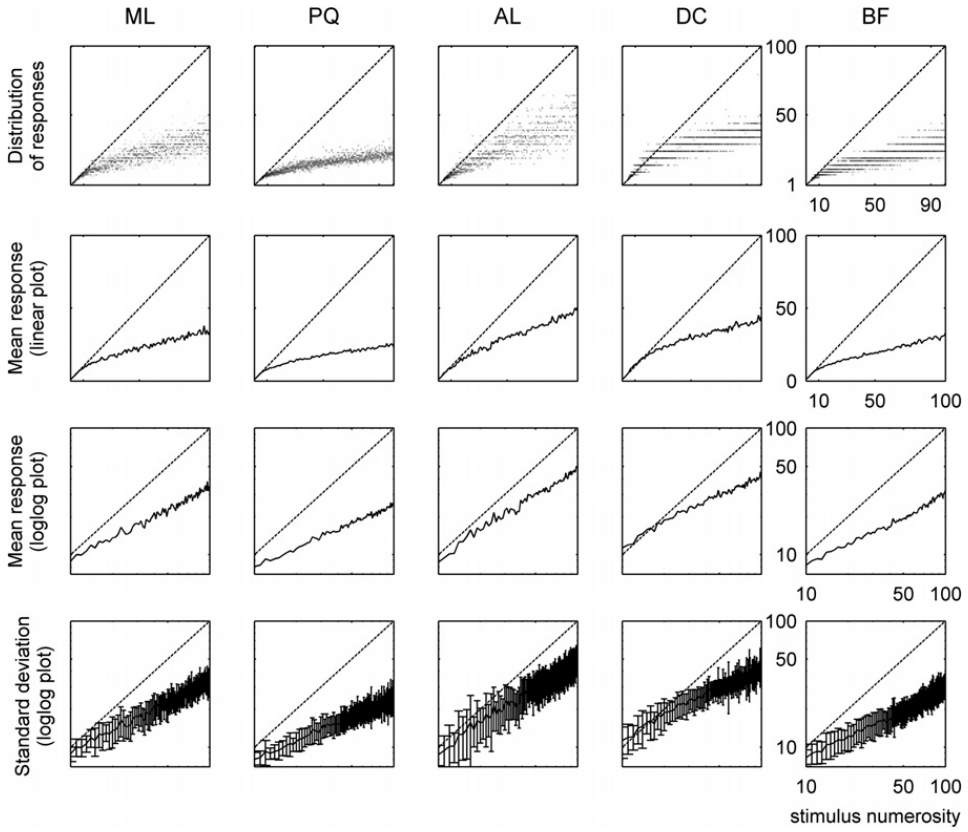

Fig. 2. Responses of the five non-calibrated participants. Horizontal axis: stimulus (n); Vertical axis: response (R). The top row shows the distribution of the responses, using a gray color scale to represent the logarithm of the frequency of occurrence of the response R for each stimulus n. On the second and third lines, mean responses are presented, first on a linear plot, and then on a log–log plot, where both horizontal and vertical axes have been log-transformed. If the response function are power shaped as predicted in the theory part, they should appear as a straight line on a log–log plot. On the last row, error bars representing standard deviation of the responses are added to the log–log mean response plot. The property of scalar variability predicts that standard deviation bars should all appear of the same length on this graph.

ity of the stimuli, there was considerable variability between participants in the amount of underestimation. For example, the mean stimulus numerosity leading to the response '30' was, respectively, 75.4 for subject ML, 71.7 for PQ, 60.8 for AL, 58.6 for DC and 82.1 for BF.

In magnitude estimation experiments, the shape of the response function is generally a power function. To test whether this property extended to the present data set, we ran regression analyses on the log-transformed data: if the response functions were power-shaped ($R = \alpha\, n^{\beta}$), then the log-transformed response function must be a linear function of the log-transformed numerosity ($\log(R) = \log(\alpha) + \beta\log(n)$). A slope of 1 in the log–log regression analysis would indicate that the responses were actually linear ($R = \alpha n$). For all the participants, the log–log regression fitted the data remarkably well ($R^2 = 0.93$–0.97 depending on the participant), and the

exponents were all significantly smaller than 1 (ranging from 0.57 to 0.76 depending on the participant), indicating that responses were shaped as a power function, and were not linear.

Another well-established fact about estimation data is that responses follow the law of scalar variability: the coefficient of variation (standard error of the responses/mean response) is constant. Two separate analyses were completed to test for scalar variability in our data. First, for each participant, we computed the coefficients of variation (CVs) for numerosities above 25 and submitted these to a linear regression with numerosity. The levels of fit were poor ($R^2 = 0.01$–0.07), indicating that CVs did not tend to vary with numerosity. The slopes indicating the dependence of CVs from numerosity were all close to 0 (range -0.0004 to 0.0006; non-significantly different from 0 except in participants PQ ($p = 0.023$) and ML ($p = 0.037$)). As a second more sensitive test of the strict linearity between standard errors and mean response, we used a log–log analysis. If scalar variability is verified, the logs of the standard error and of the mean estimates should form a straight line with a slope of 1.0 ($\log(\mathrm{Sd\_}R(n)) = \log(R(n)) + \gamma$). A slope different from 1.0 would indicate that the relation between the standard errors and the means is a more complex power function. However, the slopes were not significantly different from 1.0 for 4 out of the 5 participants (slopes for these observers: 0.86–1.31, $p > 0.08$; for the last observer PQ the slope was 1.39, $p = 0.0048$; $R^2 = 0.42$–0.74 for the different participants). While scalar variability in not strictly verified in participant PQ, it still provides an adequate description of these data, since the coefficient of variation did not vary dramatically in this participant: the slope of the linear regression was 0.0006 for this participant, which represents only 0.3% of the mean coefficient of variation (0.23). We conclude that in general, the property of scalar variability gives an adequate description of these data, even if it is not strictly verified in one of our participants.

2.4. Discussion

We recorded non-calibrated estimates in 5 participants, over numerosities ranging from 1 to 100. For each participant, the responses displayed internal coherence, as they were increasing with numerosities. However, the actual numerosity of the stimuli was strongly underestimated. Participants differed from each other in how much they underestimated the stimuli. The estimates followed two classical laws of estimation experiments. First, responses functions were power shaped, as in magnitude estimation experiments. Second, the variability of the responses followed the law of scalar variability within 4 of our 5 participants (as numerosity increased, standard deviation of the responses increased proportionately to the mean response).

Is it possible to calibrate the participants, and have them produce more accurate responses? How much information would be needed before participants calibrate their responses? Would this calibration be a local process, having an effect only around the numerosity calibrated, or would it be global, in the sense that all the numerosities would be calibrated as a whole? In order to address these questions,

we designed a second experiment, where we presented participants with a reference array before they estimated numerosities.

3. Experiment 2: Calibrated estimation

3.1. Experimental methods

3.1.1. Stimuli

Stimuli contained 9–100 dots. They were generated with the same procedure as in Experiment 1.

3.1.2. Procedure

As in Experiment 1, participants were still required to estimate the numerosity of dot arrays; however, as we wanted to pool data across groups of participants in this experiment, allowed responses were restricted to a set of round numbers only, i.e., the numbers 1 through 10 and the decade numbers above 10. This type of numbers corresponded to the most frequent responses produced by the participants in the first experiment (39% of the responses).[1] Responses were collected in the same way as in the first experiment.

The experiment consisted of two sessions, each session containing six blocks of 46 trials. Before each block participants were presented with a reference dot array (inducer) and were told that it contained 30 dots. However, depending on the experimental condition, the inducer actually contained either 25 dots (overestimated inducer), 30 dots (exact inducer) or 39 dots (underestimated inducer). The numerosity of the reference trial stayed the same through each experimental session. Displays of the inducers always belonged to the third stimulus set (constant array density and dot size).

After six blocks, there was a break lasting at least 20 min, after which participants performed a second experimental session with a different inducer. Participants who had seen overestimated or underestimated inducers were presented with exact inducers in this second part of the experiment. Amongst the participants who had seen exact inducers, half were assigned to the overestimated inducer condition, and the other half to the underestimated inducer condition. Thus, for half of the participants, the inducer presented in the second session was larger than in the first session, and for the other half, the inducer was smaller in the second session than in the first. Gender was balanced in all conditions.

[1] One might wonder whether giving additional information by restricting the responses might have contributed to calibrate the participants in the second experiment. To check for a possible effect of response instructions, we conducted an informal survey where 10 participants were presented with a single dot array containing 58 dots, in the same conditions than in the first experiment, except that they were instructed to use only small numbers or decade numbers in their response (same instructions as in Experiment 2). The participants still underestimated the array (mean estimate 42 ± 11.3). Participants of Experiment 1 showed the same amount of underestimation for their first estimate of a stimulus of numerosity 55–58 (mean estimate 40.6 ± 34.5; corresponding to trials number 1–21 for the different participants).

Unless explicitly specified in the results section, the analyses presented here concern only the first experimental session. For this session, three groups are considered: underestimated inducer, exact inducer, overestimated inducer.

3.1.3. Participants

Twenty-four french speakers (12 males, 12 females; aged 19–33) were paid for their participation. Six participants were assigned to each of the underestimated and overestimated conditions, and 12 participants to the exact inducer condition. In the second experimental session, 12 participants saw an inducer larger than in the first session, and 12 participants saw a smaller inducer. Participants were naive to the purpose of the experiment.

3.2. Results

3.2.1. Influence of inducers on numerosity naming

After calibration, responses still increased monotonically with numerosity (linear regression on mean responses against numerosity: slopes ranged from 0.40 to 2.3, significantly different from 0 for all the participants, all ps < 0.0001; Fig. 3). Furthermore, induced participants adapted their responses quite accurately to the inducer trial. The mean stimulus numerosity leading to the response '30' were 27.4, 31.5, and 40.1, respectively, in the overestimated, exact and underestimated inducer conditions, where the participants were induced to associate response 30 with stimuli containing, respectively, 25, 30, and 39 dots. The difference between groups approached significance ($F(2,21) = 3.32$; $p = 0.056$). All these distributions differed significantly from the (much higher) numerosity associated with the response '30' by the non-induced participants (t-test; all ps < 0.01). Although restricting allowed responses to decade numbers only might have contributed to compensate the participants' bias in Experiment 2, by helping them to figure out the range of stimuli we were presenting, this effect cannot account for the difference in calibration between the groups of Experiment 2. These differences can only be imputed to the different inducers presented.

The value of the inducer had an effect on the whole scale of numerosities presented, thus allowing us to reject the hypothesis of a local calibration. We grouped stimuli into 10 numerosity levels ([9 14], [15 24], ..., [85 94], [95 100]) and ran an ANOVA with factors of Numerosity and Inducer. Numerosity had a significant effect ($F(9,189) = 105.47$; $p < 0.0001$) and interacted with inducer ($F(18,189) = 2.17$; $p = 0.0054$). Numerical responses grew more slowly in the 'large inducer' condition (Fig. 3). This effect was not restricted to the neighborhood of the inducer value, but extended to the whole scale: even in the range [95 100] the responses were significantly smaller in the underestimated inducer group than in the overestimated inducer group ($F(1,10) = 6.00$; $p = 0.03$).

After the break, in the second experimental session, the participants readapted their responses to the new value of the inducer. For this analysis, we separated the participants in two groups, depending on whether the inducer presented in the second session was larger or smaller than the inducer of the first session. We ran an ANOVA with factors of session (first or second), group, and numerosity (with 10

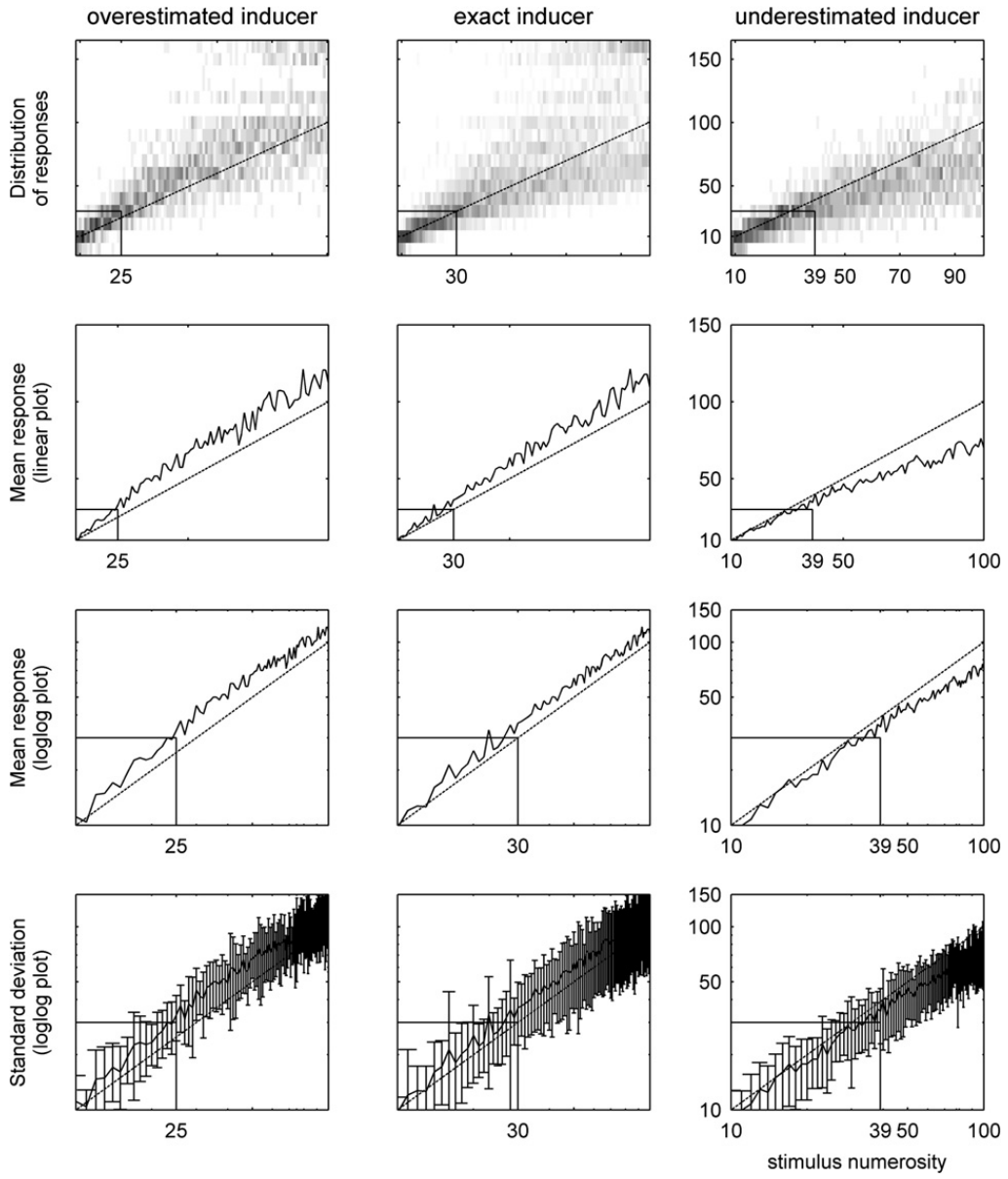

Fig. 3. Responses of the 3 groups of induced participants (same format as Fig. 2). All groups were presented with an inducer trial which they were told contained 30 dots. In the overestimated inducer group, the inducer contained 25 dots; in the exact inducer group it contained 30 dots; in the underestimated inducer group it contained 39 dots.

levels as previously). Together with an effect of numerosity ($F(9,184) = 121.3$; $p < 0.0001$), we observed an effect of the experimental session ($F(1,10) = 6.9$; $p = 0.025$), and an interaction between session and numerosity ($F(9,198) = 12.8$; $p < 0.0001$). These effects indicated that numerosities were overall more underestimated in the second session, particularly in the large numerosity range (see Krueger,

1982 for a review of similar observations in magnitude estimation experiments). More importantly, the interaction between group and session approached significativity ($F(1,10) = 3.7$; $p = 0.082$), and we also observed a strong triple interaction ($F(9,198) = 4.9$; $p < 0.0001$). The latter two effects indicate that the participants readapted differently depending on the relative size of the first and second inducers: the participants for whom the second inducer was larger than the first one decreased their responses in the second session, particularly in the large numerosity range. On the contrary, the participants who saw a second inducer smaller than the first one produced essentially equal responses in both parts. The responses in this second session are still influenced by the value of inducer presented in the first session, as responses did not fit the inducer as precisely as in the first session (mean numerosity associated with the response '30' by participants presented resp. with overestimated (25), exact (30) and underestimated (39) inducer in the second session: 31.5, 33.1 and 33.7). However, the influence of the recalibration again was extended to the whole range of numerosities, as in the first session. Hence, restricted analyses on the interval [95 100] revealed an interaction between group and session ($F(1,22) = 5.1$; $p = 0.034$), as well as an effect of session ($F(1,22) = 15.3$; $p = 0.00076$).

3.2.2. Psychophysical description of the responses and influence of non-numerical parameters

Since the calibrated estimation task corresponded to a magnitude estimation experiment, estimates were expected to form a power function of numerosity. We used a log–log regression analysis on the responses of each participant to test for the shape of the response function, as in the first experiment. The regression fitted the responses remarkably well ($R^2 = 0.80$–0.93 depending on the participant), and exponents of the power function, as indicated by the slopes in these regressions, ranged from 0.69 to 1.23. The mean exponent was 0.96, but departed significantly from 1.0 in 23 of the 24 participants, thus indicating that the response functions were not linear, but really shaped as power functions, with an exponent close to 1.0 on average.

Second, the variability of the responses was still scalar after calibration: as predicted, the standard deviation of the responses increased with the mean response, and the coefficient of variation (standard deviation/mean response) was essentially constant. Data were submitted to the same analyses as for the non-calibrated participants. First a linear regression was run between coefficient of variation and numerosity. The slopes of these linear fits were very close to zero (overestimated inducer group: slope = 0.0008, different from zero $p = 0.049$; exact inducer group: slope = 0.0013, different from zero $p = 0.002$; underestimated inducer group: slope = -5.10^{-5}, non-significantly different from zero). The level of fit was very poor in these regression ($R^2 = 0.0004$–$.12$), indicating a weak covariation between coefficients of variation and numerosity. Second, we used a log–log regression to examine the relation between the standard deviation and the mean response. The slopes of the regressions were close to 1 though they departed significantly from 1 in the overestimated and exact inducer groups (overestimated inducer group: slope = 1.17, $p = 0.02$; exact inducer group: slope = 1.21, $p < 0.0001$). In the underestimated

group however, it was not different from 1 and indicated strict scalar variability (slope = 1.02, $p = 0.7$). We conclude that scalar variability is approximately, but not always perfectly verified in our data set.

Finally, we verified that participants were basing their responses on the numerosity of the arrays, rather than on some non-numerical parameter, such as array density, or dot size. Our stimuli belonged to three intermixed sets with different types of controls. Thus, if a participant was basing his responses on a single non-numerical parameter, his responses would be flat in at least one of the three stimulus sets. On the contrary, we observed monotonically increasing responses in all of the three sets: for each participant, the slope of a linear regression between responses and numerosity was significantly higher than zero within each set (1st set: slopes, 0.4–2.4; 2nd set: slopes, 0.4–2.0; 3rd set: slopes, 0.4–2.4; all ps < 0.0001). This result rules out the possibility that the observers were relying on a single low-level parameter. However, estimates produced for the stimuli belonging to the third set were smaller than for the other sets ($F(2,46) = 9.57$, $p = 0.00034$), probably because the average density or total occupied area differed in the third set compared to the other ones. It is possible that participants were using a complex combination of two or more parameters to extract numerosity (such as multiplication of density by total occupied area). Indeed, such combinations might be one way by which the visual system extracts approximate numerosity from visual displays (Allik & Tuulmets, 1991; Frith & Frith, 1972).

4. Experiments: Discussion

We studied numerical estimation both with and without calibration by a reference trial. Although non-calibrated estimates systematically underestimate the true numerosity, it is possible to calibrate estimation, and participants adapt their responses very precisely to the indication given, even when this indication is inaccurate (e.g., 39 dots pretended to be 30). Furthermore, the influence of the calibration is not restricted to the neighborhood of the calibrated numerosity but extends to the whole range of numerosities tested. Finally, as expected, the participants' behavior follows well-established psychophysical laws, both with and without calibration: first, the mean responses form a power function of numerosity, as in magnitude estimation experiments (Krueger, 1989); second, the variability in the responses follows the law of scalar variability, as in numerosity production tasks (Dehaene & Marques, 2002; Whalen et al., 1999).

In accordance with our results, some observations reported by Minturn and Reese (1951) and Krueger (1984) indicated that observers modify their estimates when they are provided feedback about their estimates, and that even a minimal amount of feedback modifies the estimates over all numerosities tested. After recording spontaneous estimation in a first session, Minturn and Reese recorded a second session, where after each trial they indicated to their participants the real value of numerosity. In this feedback condition, the amount of variability between participants was considerably reduced. Moreover, the effect of feedback was strong

as it was still present 8 months after the initial testing. As only a restricted number of different numerosities were presented as stimuli, one could argue that the task participants performed was not an estimation task anymore in this second part: rather, participants might have been identifying a numerosity amongst different possibilities. However, the authors noted that the modification of the responses could be seen in the very first block, and even on the very first trials: whenever they had been provided information on a given numerosity, participants transferred this information to other numerosities. Later, Krueger reported a similar observation. Between two blocks of a numerosity estimation experiment, he presented a dot array to the participants and informed them that this array contained 200 dots. This isolated reference reduced considerably the variability of the responses across participants, over the whole range of numerosities tested. Our calibrated estimation experiment establishes in a more systematic way the finding that reference trials have a impact on subsequent estimates. Furthermore, by presenting different inducer conditions, we also show that the calibration is very precise: the responses of the participants differ between our groups, and fit our different inducers with a high degree of precision.

Although calibration is immediate, long lasting, and global in adults, studies using some kind of calibration have obtained mixed results when applied to children. Lipton and Spelke (2005) have shown that children aged 5 years can use repeated indications over multiple numerosities to produce accurate estimates of arrays containing up to 100 dots, provided that they can count up to this number. However, Booth and Siegler (2006) observed that giving a single indication of 1000 dots to children aged 8 and 10 years was not sufficient to fully calibrate their estimations: although the children's response increased linearly with numerosity, a possible effect of the calibration procedure, they only produced 500 dots on average when asked for an array of 1000 dots. Possibly, this failure can be explained by the high range of stimuli used. The acquisition of a mapping between the verbal symbols and the internal representation of magnitudes may occur gradually in children, and slowly extend to a wider range of numbers as familiarity with these numbers increases (see Le Corre & Carey, 2007 for a failure in the estimation of small numbers by younger children).

What is the mechanism underlying calibration in adults? It is possible that calibration is a strategic process, by which participants consciously modify their numerical responses, for instance by performing an approximate mental multiplication on them. It is also possible that calibration results from an automatic learning process, occurring unintentionally. More likely, calibration results from a mixture of controlled and automatic effects. After the end of experiment, participants were asked informally about their impressions; they reported that they consciously corrected their responses to match the inducer, but when we told them that non-calibrated participants had estimated the maximum numerosity at 50 instead of 100, they did not admit that their spontaneous responses would have been that inaccurate. Orthogonally to the question of the automaticity of calibration, one can also wonder at which processing stage the calibration occurs: in the encoding of numerosities itself, in the

process of response selection, or both. These questions will be addressed in the following section, using a theoretical model of the estimation task.

5. Theory

5.1. General description of the model

In what follows, we specify a theoretical model of the numerical estimation task, where an observer is instructed to estimate the numerosity of stimuli. Then, we validate this model by comparing its predictions to empirical data. In a third step, we use the model to characterize the calibration process in more details, and separate effects originating from numerosity encoding (sensitivity), or response selection (bias).

Our model is inspired by the Thurstonian framework (Thurstone, 1927) and by the signal detection theory (Wickens, 2002). The Thurstonian framework, originally developed to describe the perception of continuous variables such as weight, brightness, or loudness, has been applied to numerosities by van Oeffelen and Vos (1982), who proved its efficiency in predicting performances in a numerosity identification task, where participants had to identify the numerosity of dot arrays amongst two possibilities. Since then, this model has been applied successfully to several tasks, all requiring binary judgments: comparison of numerosity, matching to sample (Piazza et al., 2004), addition and subtraction of numerosities (Barth et al., 2006; Pica et al., 2004). Here, we extend this model to the numerical estimation task, which requires participants to give a numerical response.

Our model is based on five hypotheses. First, we assume that each perceived numerosity is encoded on an internal continuum, called the *number line* (hypothesis 1). Second, following Fechner's law, we assume that the number line is compressive and logarithmically scaled[2] (hypothesis 2). Hence, when a numerosity n is perceived, it generates an activation on the number line, situated on average on the point $\log(n)$. Third, the distribution of activation around $\log(n)$ is assumed to be gaussian, as proposed by Thurstone, with a constant width w (hypothesis 3). This parameter, called the *internal Weber fraction*, measures the overall level of noise inherent to the representations of numerosities. As an illustration, if $w = 0.20$, then it means that the activation evoked by a numerosity n will fall in the interval $[\log(n - 20\%), \log(n + 20\%)]$ with a probability of 0.7. Hence, for numerosity 10, the activation would fall in the interval [log(8), log(12)] in 70% of the trials. Two different evaluations of the value of the internal Weber fraction have been reported so far: the first one at 0.11 (van Oeffelen & Vos, 1982), the second one at 0.17 (Piazza et al., 2004).

[2] The fact that the internal scale is logarithmic rather than linear is debated (Brannon, Wusthoff, Gallistel, & Gibbon, 2001; Dehaene, 2003; Dehaene & Marques, 2002). Here, we do not attempt to prove that the scale is logarithmic, however, we show that based on the logarithmic scale hypothesis, it is possible to build a model accounting for several aspects of the data. Nevertheless, using the shape of the mean response functions in estimation tasks, we have developed elsewhere an argument favoring the hypothesis of a logarithmic scale over a linear scale (Izard, 2006).

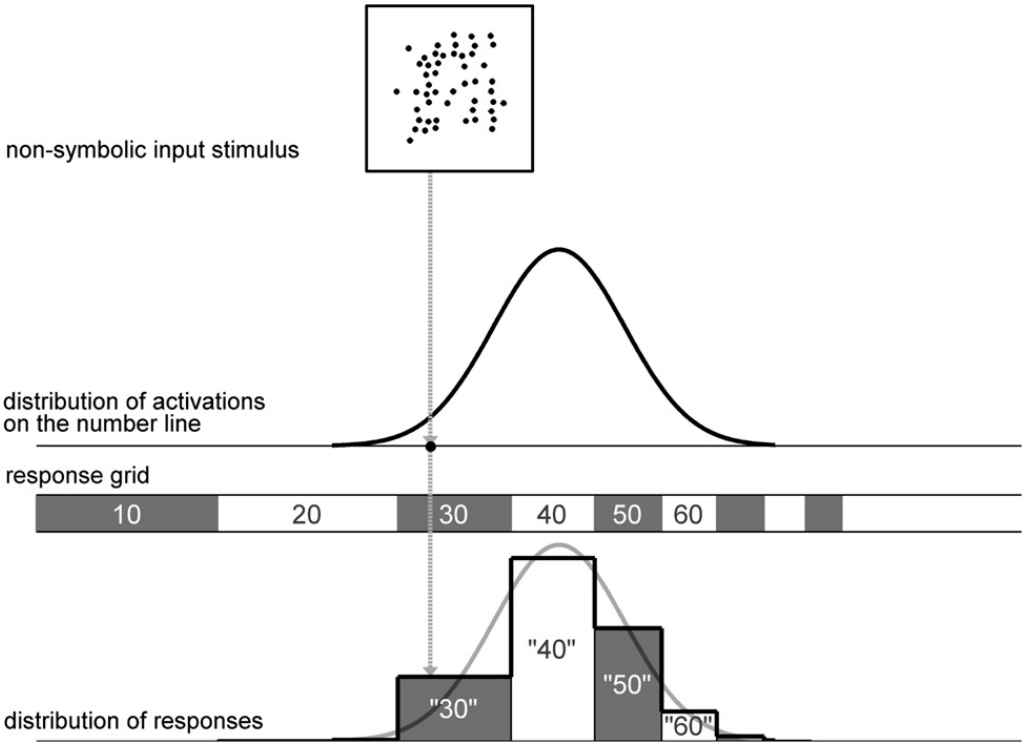

Fig. 4. General presentation of the model. Numerosities are first encoded on an internal continuum, the mental number line. Activations on the number line are then translated into verbal number words, by means of a response grid: the number line is divided in segments, each corresponding to a different verbal label. The theory section developed here precises the shape of the response grid. The bottom panel represents the distribution of responses emanating from a gaussian distribution of activations on the number line. Bar areas are proportional to the frequency of the responses, and bar widths respect the shape of the response grid. In these conditions, the curve formed by the bars (highlighted in black) follow the initial gaussian activation curve (in light gray).

To these three classical hypotheses, we have added two new ones describing how the analog representation on the number line is then transformed into a verbal numerical response. We first postulate that a list of criteria are set on the number line, defining a *response grid* (Fig. 4, hypothesis 4): the number line is divided in several segments, each of them associated with a different verbal label (Dehaene & Mehler, 1992). Finally, the grid can be stretched or compressed, and this corresponds to calibration (hypothesis 5). When no indication is given, non-calibrated estimations are generated using an idiosyncratic *spontaneous* response grid. In the second experiment where participants are calibrated, they transform this spontaneous response grid into a *calibrated* response grid to fit the indication given.

As our model has been inspired by the Signal Detection theory, correspondences can be drawn between the parameters in these two frameworks. First, our Internal Weber Fraction (parameter w) measures the precision of the internal representation, and is analogous to the sensitivity (d') in Signal Detection Theory. Although analogous, these parameters capture the sensitivity in two different ways: in the Signal

Detection Theory, the dispersion of internal representation on the internal continuum is supposed to be constant, but the distance between the two distributions (d') varies and determines the level of sensitivity. In our case, the distances between the distributions are fixed since distributions are centered on the log(n)s, but the amount of dispersion varies. Second, response bias are captured in our model by the position of the response criteria which define the response grid.

In the following theory part, our aim is to give a quantitative description of the response grid, both in the non-calibrated and in the calibrated estimation cases, and to compare the predictions of the model to several aspects of the empirical data. In our model, the response grid is not optimally adapted to the internal logarithmic scale (explaining the inaccuracy of non-calibrated estimates), but still partially adapted: the shape of the response grid is constrained. Our mathematical theory specifies the constraints defining the shape of the response grid. Then, once we have determined the shape of the response grid, we can derive the predictions of the model, and verify 1. that the predictions follow the property of scalar variability, a well-established fact about estimation data, and 2. that the response functions (mean response associated to each numerosity) are predicted to be power functions, as observed in magnitude estimation experiments. After checking for the validity of the model, we then use the distribution of responses to reconstruct the distribution of activations (or *activation curve*) on the number line, and justify *a posteriori* the choice of a gaussian distribution (hypothesis 3). Finally, we derive the quantitative value of the parameters of the model for the different groups of participants in Experiment 2 to test which parameters are responsible for calibration (sensitivity vs. bias parameters).

5.2. Shape of the response grid

We will specify the shape of the spontaneous response grid in two steps: first, we define a canonical response grid, corresponding to an optimal behavior, where a participant responds always with the closest number; and then we specify the response grids actually used by participants in relation to this canonical response grid.

In our theory, participants engaged in an estimation task represent numerosity as an activation on the internal number line; and they have to decide their verbal response on the basis of this activation only. An optimal strategy would consist in giving the number which is the closest to the activation point: hence, if the activation fell close to log(30), the response given would be '30'. This strategy defines a canonical response grid, where criteria are situated on the midpoints between the logs of the possible responses (see Fig. 5a). In the case of our task, where participants are required to respond with decade numbers only, the interval associated to the response 'R' is delimited by the criteria $c_-(R) = \frac{1}{2}(\log(R) + \log(R-10))$ and $c_+(R) = \frac{1}{2}(\log(R) + \log(R+10))$ (see Table 1 for the list of abbreviations used in the text).

If participants were using this canonical response grid, they would be quite accurate in their responses, but our empirical data show that estimations are indeed inaccurate in the absence of calibration. We propose that, instead of the canonical response grid, participants were using idiosyncratic affine transformed versions of it. Namely, the spontaneous response grid can be derived from the canonical response grid by applying

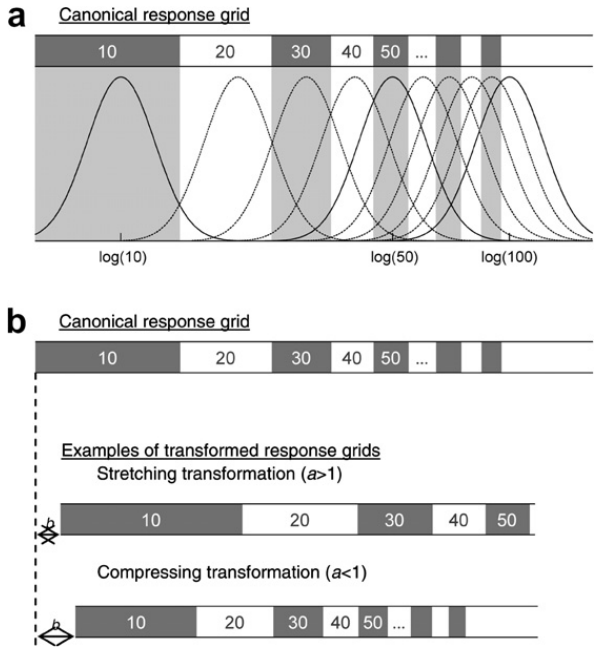

Fig. 5. Response grids. (a) In the canonical response grid, the criteria lay at the midpoint between the log(n)s. (b) The spontaneous and calibrated response grids are related to the canonical response grid is an affine transformation: the canonical response grid is either stretched or compressed (parameter a), and shifted (parameter b).

Table 1
List of abbreviations used in the text

$c_-(R), c_+(R)$	Left and right criteria delineating the segment of the mental number line associated to the verbal response 'R' for the *canonical* response grid
a, b	Parameters of the affine transformation applied to the canonical response grid. a: stretch, b: shift
a_s, b_s	Parameters defining the *spontaneous* response grid
$R_{a,b}(n)$	Mean estimate produced for stimuli of numerosity n, when the response grid used is the image of the canonical response grid by an affine transformation of parameters a, b

two successive transformations: the canonical response grid is either stretched or compressed by a first parameter a_s, then it is shifted from the origin by a second parameter b_s (see Fig. 5b). Thus, the interval associated to the response 'n' is no longer $[c_-(n), c_+(n)]$ but $[a_s c_-(n) + b_s, a_s c_+(n) + b_s]$. In other words, the participant inaccurately determines the amount of activation corresponding to the zero level (which he sets to b_s), and the value of one unit on the number line (set to a_s). Furthermore, the values of the parameters a_s and b_s vary across individuals, in line with our experimental results that spontaneous estimates are highly variable between participants.

In Experiment 2, participants were shown an inducer trial before they started to produce estimates: an array was presented, and they were told that its numerosity

should be associated with the response '30'. Experimental results show that in this case, participants calibrate their responses to the inducer, and that all the numerosities are calibrated globally. To capture the idea of a global calibration, we postulate that participants transform their *spontaneous* response grid, again by an affine transformation, in order to bring their response grid in accordance with the indication given. As the *spontaneous* response grid is an affine transformed version of the *canonical* response grid, the *calibrated* response grid is itself an affine transformed version of the *canonical* response grid, with different values of parameters a and b. For example, if during the induction period, a stimulus of numerosity n_0 is linked to response R_0, the participant will set the parameters a and b so that $ac_-(R_0) + b < \log(n_0) < ac_+(R_0) + b$. The parameters a and b therefore determine which response grid has been used, and they measure possible response bias in the participants.

Once the response grid has been defined, we can derive the predictions of the model. The probability to observe the response R following a stimulus of numerosity n is obtained by taking the integral of the activation curve for numerosity n (a gaussian centered on $\log(n)$ and of width w) on the segment of the number line associated with the response R:

$$P(R|n) = \int_{ac_-(R)+b}^{ac_+(R)+b} \mathrm{Gauss}(\log(n), w) \tag{1}$$

Fig. 6 illustrate the predictions of the model for three different conditions on the parameter a: canonical response grid ($a = 1$), stretched grid ($a > 1$) and compressed grid ($a < 1$).

5.3. Tests of the model

5.3.1. Property of scalar variability

As the property of scalar variability is a very general aspect of estimation data, models of estimation cannot be acceptable if their predictions do not follow this property. Thus, as a first step to the experimental validation of our model, we have demonstrated that it predicts scalar variability in the responses (see Supplementary Data section). This prediction is tightly related to our assumptions concerning the shape of the response grid: the scalar variability holds in the predictions only because the shape of the criteria stays always logarithmic, i.e., similar to the shape of the internal scale, even when the response grid is distorted by an affine transformation.

5.3.2. Shape of the response function

As a second requirement for a model of estimation, the model should predict that the mean response function is a power function. In magnitude estimation experiments, response functions have been repeatedly observed to be power shaped, an observation replicated on our data set.

At this point, it is important to emphasize that the shape of the response function $R_{a,b}(n)$ in our model does not necessarily coincide with the internal scale $\log(n)$. For example, if the canonical response grid is used ($a = 1$, $b = 0$), the responses are almost accurate ($R_{1,0}(n) \approx n$). However, if the response grid used is not the canonical

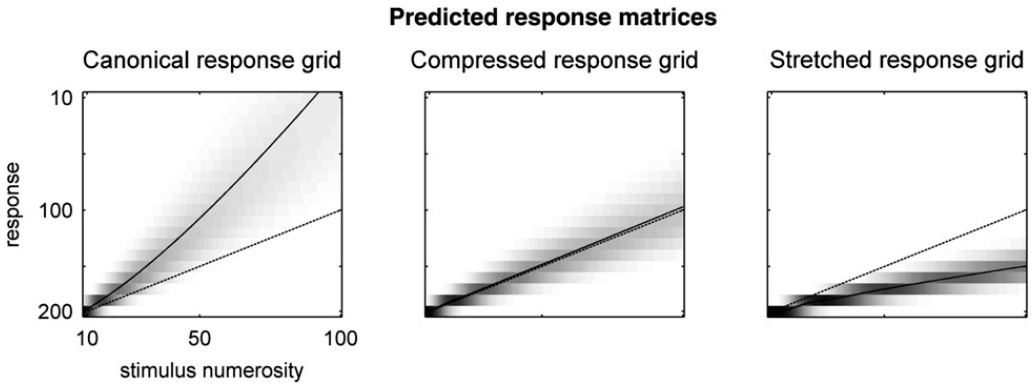

Fig. 6. Predicted response matrices. Horizontal axis: stimulus numerosity (n); Vertical axis: verbal response (R). Shades of gray indicate the probability of the response R for the stimulus n. Plain lines give the mean responses, and dotted lines indicate the correct value of the stimulus numerosity (diagonal $R = n$). Three cases are presented, corresponding to different values of the parameters of the affine transformation applied to the canonical response grid. Compressed response grid: $a < 1$; Canonical response grid: $a = 1$; Stretched response grid: $a > 1$.

response grid, the shape of $R_{a,b}(n)$ can be derived and it takes the general form of a power function (see Supplementary Data section), with exponent $\frac{1}{a}$ and coefficient $\exp(\frac{w^2}{2a^2} - \frac{b}{a})$. Again, this prediction originates from the logarithmic shape of the response grid, and from the fact that the response grid is not perfectly matched to the internal scale, but has undergone an affine transformation.

5.3.3. Activation curves

As illustrated in Fig. 4, the distribution of the responses, once brought onto the response grid, i.e., onto a logarithmic scale, is a good estimate of the distribution of activations on the number line. In our model, we postulate that the activations follow a gaussian distribution on a logarithmic scale (see Section 5.1, hypothesis 3). In this part, we release this hypothesis and challenge it by computing an estimate for the activation curve, and comparing this estimate to a gaussian fit. We also calculate the standard deviation of the estimated activation curve to obtain an estimation of the internal Weber fraction w, which represents the global amount of noise in numerosity representations.

In this section, we detail the procedure used to estimate a grand average activation curve, across all participants and all numerosities. We first evaluated separately the distributions of activations for each participant and each stimulus numerosity. To do so, distributions of responses for each participant and each numerosity (larger than 20) were placed on the response grid (i.e., on a logarithmic scale), as illustrated in Fig. 4. These distributions were then recentered and averaged across numerosities, for each participant. Furthermore, before comparing the activation curves between participants, we had to correct for the stretching/compression imposed to the canonical response grid. To obtain the value of the parameter a, giving the amount of dis-

tortion, we used the exponent of the mean response function of each participant, as our theory predicted this exponent to be equal to $\frac{1}{a}$. Once corrected, the estimated activation curve of each participant was fitted with a gaussian distribution, and we extracted the width of these gaussian fits to obtain individual values of the internal Weber fraction w. Finally, the estimated activation curves were averaged across participants, and the grand average curve was again fitted with a gaussian distribution. As can be seen in Fig. 7, the grand average curve obtained with this procedure was very close to a Gaussian.

Depending on the participant, the gaussian fit performed on the estimated activation curve explained 84–99% of the variance, and the estimated Weber fraction ranged from 0.15 to 0.41. On the average activation curve, the fit explained 99% of the variance, and the overall internal Weber fraction was estimated at 0.22 (Fig. 7A). We calculated the Z-scores associated with the part of the number line corresponding to 90% of the distribution. Plotted against the number line, the Z-scores formed a straight line going through the origin (Fig. 7b, slope $= 4.53 = 1/0.23$, $R^2 = 0.999$), indicating that the internal activation distribution is extremely close to a gaussian function.

5.4. Mechanism underlying calibration

In the preceding sections, we have given a general theory describing the estimation task in the calibrated condition as well as in the non-calibrated condition. One might

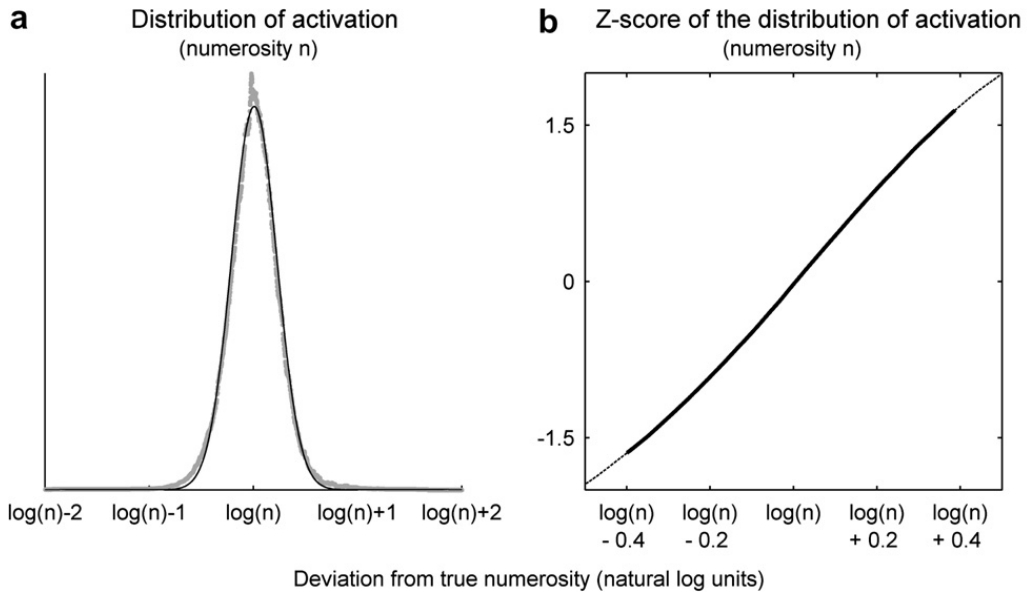

Fig. 7. Reconstruction of the activations on the number line. (a) Distribution of activations on the number line, averaged across all participants, and compared with its gaussian fit. The standard deviation of the best-fitting Gaussian curve ($R^2 = 0.99$) corresponds to the estimated internal Weber fraction: $w = 0.22$. (b) Z score of the distribution in (a), for the interval of the internal number line encompassing 90% of the data.

have argued that calibrated and non-calibrated estimation rely on fundamentally different mechanisms. By giving a unified description of these two situations, we here show that this hypothesis is not necessary.

We now use our model to precise the mechanisms underlying calibration. Does calibration occur at the stage of response encoding (sensitivity), or at the stage of response selection (response bias)? In our model, sensitivity was measured by the internal Weber fraction (w), which gives the amount of noise in the representations of numerosity. Response bias are captured by the parameters of the affine transformation imposed on the response grid (a and b). The value of these parameters was estimated for each participant and each of the two experimental sessions, and we examined whether these parameters varied in a systematic way between groups of participants.

In the preceding section, we described how w was extracted for each participant. To estimate the parameters a and b, we reversed the formulas giving the slope and intercept of the log–log regression from a, b, and w (see Section 5.3.2). Parameter a is equal to the inverse of the slope ($\frac{1}{\text{slope}}$). Parameter b corresponds to the following formula: $-\frac{\text{interc}}{\text{slope}} + \text{slope} * \frac{w^2}{2}$.

Estimated parameters for the first session were first submitted to an analysis of variance to test for a possible effect of the value of the inducer (overestimated, exact, or underestimated). For all three parameters, no significant difference between groups was observed (all ps > 0.16). As a more sensitive test, we tested how these parameters evolved as participants recalibrated their responses between the two experimental sessions. For each participant, parameters w, a and b were evaluated separately for the two experimental sessions, and then these estimated parameters were each submitted to an analysis of variance with factors group (indicating whether the value of inducer presented in the second session was larger or smaller than in the first session), and session (first or second). The sensitivity (internal Weber fraction w) stayed constant across sessions, independently of the relative value of the two inducers (all ps > 0.14). Not only did the w stay constant on average, but also within each participant, as indicated by a high level of correlation across participants between the two sessions ($R^2 = 0.61$, $p = 0.002$; see Fig. 8, left panel). On the contrary, response bias changed over the course of the experiment, since both a and b presented an effect of the session ($F(1,20) = 29$ and 23, respectively, ps < 0.001). The response grid tended to be more and more stretched over the course of the experiment, as indicated by an increase of parameter a (first session: 1.07; second session: 1.18). This evolution was independent from the calibration process, since both groups presented the same increase on a (no interaction between group and session, $F < 1$; see Fig. 8, middle panel). This progressive stretching of the response grid could be related to the natural tendency of participants to contract their response range over the course of the experiment. Parallelly, the translation parameter b (shift) tended to decrease over the course of the experiment (first session: -0.23, second session: -0.58), to compensate for the increase in stretching. Contrary to the stretch parameter, this evolution was modulated by the calibration indications, and the interaction between group and session for b approached significance ($F(1,20) = 4.3$, $p = 0.051$; see Fig. 8, right panel). According to these analyses, calibration does not modify the sensitivity of the participants, but occurs at the level of

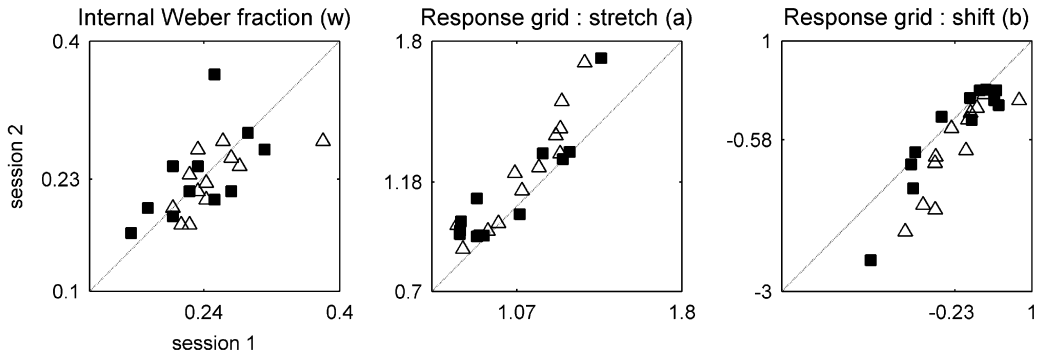

Fig. 8. Evolution of the internal Weber fraction (w, left panel), and the two parameters of the affine transformation applied to the response grid (stretch on the middle panel and shift on the right panel) between the two experimental sessions in Experiment 2. X-axes represent the value of these parameters for the first session, and Y-axes the values for the second session. Doted lines indicate the equality diagonal (first session = second session). Each dot represents one participant: participants presented with a larger inducer in the second session are marked as filled squares, and participants presented with a smaller inducer in the second session are marked as unfilled triangles. Values indicated on the axes correspond to the means for the first and second sessions. The graphs illustrate how each subject possesses an idiosyncratic value of parameters w, a and b, which is reproducible across the two sessions, and also demonstrate that only the parameters a and b are affected by calibration.

response selection, by changing the response bias. Finally, even if all participants tend to modify a and b in the same way, assigning a value to these parameters remains largely idiosyncratic, as indicated by the absence of effect between groups in the first session, and by a high level of correlation between the values of a and b for the two sessions across participants (a: $R^2 = 0.92$, $p < 0.0001$; b: $R^2 = 0.89$, $p < 0.0001$).

5.5. Theory: discussion

The model of the number line has previously proved its efficiency in various numerical tasks involving two alternatives forced choice responses: comparison, same/different judgment (Piazza et al., 2004), identification of a numerosity amongst two different possibilities (van Oeffelen & Vos, 1982), mental addition and subtraction of numerosities (Barth et al., 2006; Pica et al., 2004; see Dehaene, 2007 for a review and theoretical outlook). Here, we extended the model to another situation: a numerosity estimation task, where response alternatives are a priori infinite. Our model relies on three classical hypotheses: (1) numerosities are encoded on a number line; (2) the scale of this internal number line is logarithmic; (3) for a given numerosity, the dispersion of activations on the number line follows a gaussian distribution. Furthermore, two new hypotheses have been added to account for the need to give numerical responses in estimation tasks: (4) number words are associated to segments of the number line, thus defining a response grid; (5) responses can be calibrated by applying a transformation to this response grid. More precisely, when the response grid needs to be calibrated, the participant applies an affine transforma-

tion to its spontaneous response grid. Because it is a global transformation applied to the whole response grid, the idea of an affine transformation captures our experimental result that all numerosities are calibrated at once.

The model is a good predictor of several aspects of the data: first, the shape of the mean response function (predicted to be a power function), and second, the dispersion of the responses around the mean, which follows scalar variability. Moreover, we were able to reconstruct the distribution of activations on the number line from the responses to our numerosity estimation task, and prove that the activation curves are gaussian, as it is postulated in the model (hypothesis 3).

5.5.1. On the process of calibration

The model incorporates several parameters which could serve to underly the calibration process. On one hand, the internal Weber fraction measures the sensitivity of the participants. On the other hand, the process of calibration is modeled by an affine transformation applied to the response grid. All these parameters are idiosyncratic and vary between participants. Nevertheless, when the participants are presented with a new value of inducer and need to recalibrate their responses, the intrinsic variability of the internal continuum (sensitivity) remains constant, while the stimulus-response mapping is shifted dynamically to fit the new inducer value.

5.5.2. Precision of the representation of numerosity

The internal Weber fraction w, a measure of the amount of dispersion of the activations on the number line, was estimated to be 0.22 from our data. This value is slightly higher than previous estimations (Piazza et al., 2004; van Oeffelen & Vos, 1982). However, note that in our design, contrary to previous estimations of the Weber's fraction, the participants did not receive any feedback. Receiving feedback may help participants identify the influence of low-level continuous parameters biasing their estimation. Secondly, in previous attempts to evaluate the internal Weber fraction, the authors did not always control for non-numerical confounds, and the participants may have used these confounds to improve their performance. The first estimation of the internal Weber fraction was reported by van Oeffelen and Vos (1982) at 0.11. Actually, in their experiment, several non-numerical parameters such as total luminance and density of the pattern were confounded with numerosity. These parameters are known to be very influential in estimation of dots arrays (Allik & Tuulmets, 1991; Frith & Frith, 1972), and this might explain why they found a value smaller than ours. More recently, Piazza et al. (2004) controlled for non-numerical variables and found an internal Weber fraction of 0.17. The difference with our 0.22 estimation may result from the fact that feedback was given on each trial in their experiments.

6. Conclusion

We close by noting that one goal of psychology is to establish precise laws of behavior. The present numerosity estimation task provides highly regular data following predictable quantitative laws. We have developed a comprehensive theory

of the estimation task, modeling the mapping between numerical symbols and quantities, and capturing the regularities observed in the data at several levels. The present theory and experimental results provide important tools which, in the future, could be used to explore the interface between verbal and non-verbal numerical cognition, for instance in brain lesioned dyscalculic patients or in young children.

Acknowledgements

Supported by INSERM, and a McDonnell Foundation centennial fellowship (S. D.). The manuscript was partially written while V. I. was hosted by Elizabeth Spelke at the Department of Psychology of Harvard University, and sponsored by a grant of the Fyssen Foundation. We thank Elizabeth Spelke, Manuela Piazza, Claire Sergent, Miles Shuman, Anna Wilson, Susannah Revkin, Xavier Seron, Christophe Pallier, and Matthieu Le Corre for their valuable comments on earlier versions of this manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cognition.2007.06.004.

References

Allik, J., & Tuulmets, T. (1991). Occupancy model of perceived numerosity. *Perception and Psychophysics, 49*(4), 303–314.

Barth, H., Kanwisher, N., & Spelke, E. S. (2003). The construction of large number representations in adults. *Cognition, 86*, 201–221.

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E. S. (2006). Nonsymbolic arithmetic in adults and young children. *Cognition, 98*, 199–222.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 41*(6), 189–201.

Brannon, E. M., & Terrace, H. S. (2000). Representation of the numerosities 1–9 by rhesus macaques (macaca mulatta). *Journal of Experimental Psychology: Animal Behavior Processes, 26*(1), 31–49.

Brannon, E. M., Wusthoff, C. J., Gallistel, C. R., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science, 12*(3), 238–243.

Cantlon, J. F., & Brannon, E. M. (2006). Shared system for ordering small and large numbers in monkeys and humans. *Psychological Science, 17*(5), 401–406.

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

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

Dantzig, T. (1967). *Number: The language of science*. New York: The Free Press.

Dehaene, S. (1997). *La bosse des maths*. Paris: Odile Jacob Science.

Dehaene, S. (2003). The neural basis of the weber-fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences, 7*(4), 145–147.

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard & Y. Rossetti (Eds.), *Attention and Performance XXII. Sensori-motor foundations of higher cognition*. Cambridge, Mass: Harvard University Press.

Dehaene, S., & Marques, J. F. (2002). Cognitive euroscience: Scalar variability in price estimation and the cognitive consequences of switching to the euro. *The Quaterly Journal of Experimental Psychology, 55*(3), 705–731.

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

Flombaum, J., Junge, & J. Hauser, M. D. (2005). Rhesus monkeys (macaca mulatta) spontaneously compute addition operations over large numbers.

Frith, C. D., & Frith, U. (1972). The solitaire illusion: An illusion of numerosity. *Perception and Psychophysics, 11*(6), 409–410.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences, 4*, 59–65.

Gallistel, C. R., & Gelman, R. (2005). Mathematical cognition. In K. Holyoak & R. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 559–588). Cambridge University Press.

Ginsburg, N. (1978). Perceived numerosity, item arrangement, and expectancy. *American Journal of Psychology, 91*(2), 267–273.

Hauser, M. D., Tsao, F., Garcia, P., & Spelke, E. S. (2003). Evolutionary foundations of number: Spontaneous representation of numerical magnitudes by cotton-top tamarins. *Proceedings of the Royal Society London Series B Biological Science, 270*(1523), 1441–1446.

Hollingsworth, W. H., Simmons, J. P., Coates, T., & Cross, H. A. (1991). Perceived numerosity as a function of array number, speed of array development, and density of array items. *Bulletin of the Psychonomic Society, 29*(5), 448–450.

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

Izard, V. (2006). Interaction entre les représentations numériques verbales et non-verbales: étude théorique et expérimental. Ph.D dissertation, Université Paris VI, France.

Krueger, L. E. (1972). Perceived numerosity. *Perception and Psychophysics, 11*(1), 5–9.

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

Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception and Psychophysics, 35*(6), 536–542.

Krueger, L. E. (1989). Reconciling fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences, 12*, 251–320.

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

Logie, R. H., & Baddeley, A. D. (1987). Cognitive processes in counting. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13*(2), 310–326.

Masin, S. C. (1983). Null effect of intramodal stimulus-range variation on the exponent for numer-ousness. *Perceptual and Motor Skills, 56*, 851–855.

McCrink, K., & Wynn, K. (2004). large number addition and subtraction by 9-month-old infants. *Psychological Science, 15*(11), 776–781.

Minturn, A. L., & Reese, T. W. (1951). The effect of differential reinforcement on the discrimination of visual number. *Journal of Psychology, 31*, 201–231.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215*, 1519–1520.

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron, 44*(3), 547–555.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an amazonian indigene group. *Science, 306*, 499–503.

Spelke, E. S., & Tsivkin, S. (2001). Language and the brain: A bilingual training study. *Cognition, 78*, 45–88.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*(3), 153–181.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286.
van Oeffelen, M. P., & Vos, P. G. (1982). A probabilistic model for the discrimination of visual number. *Perception and Psychophysics, 32*, 163–170.
Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Non-verbal counting in humans: The psychophysics of number representation. *Psychological Science, 10*(2), 130–137.
Wickens, Thomas D. (2002). *Elementary Signal Detection Theory*. New-York: Oxford University Press.
Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 358*, 749–750.
Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74*, Bl–Bll.