Nonverbal arithmetic in humans:
Light from noise

Sara Cordes
Duke University, Durham, North Carolina

C. R. Gallistel and Rochel Gelman
Rutgers University, New Brunswick, New Jersey

and

Peter Latham
University College London, London, England

Animal and human data suggest the existence of a cross-species system of analog number representation (e.g., Cordes, Gelman, Gallistel, & Whalen, 2001; Meck & Church, 1983), which may mediate the computation of statistical regularities in the environment (Gallistel, Gelman, & Cordes, 2006). However, evidence of arithmetic manipulation of these nonverbal magnitude representations is sparse and lacking in depth. This study uses the analysis of variability as a tool for understanding properties of these combinatorial processes. Human subjects participated in tasks requiring responses dependent upon the addition, subtraction, or reproduction of nonverbal counts. Variance analyses revealed that the magnitude of both inputs and answer contributed to the variability in the arithmetic responses, with operand variability dominating. Other contributing factors to the observed variability and implications for logarithmic versus scalar models of magnitude representation are discussed in light of these results.

Humans and other animals appear to compute descriptive statistics in a variety of domains—from language (e.g., Aslin, Saffran, & Newport, 1999), to foraging (Gallistel, 1990), to vision (e.g., Ariely, 2001) and motor skills (Trommershäuser, Maloney, & Landy, 2003). These statistics may derive from mental magnitudes representing elementary abstractions like number, duration, and distance (Gallistel, Gelman, & Cordes, 2006). These magnitude representations are approximate (i.e., fuzzy estimates) and nonverbal—animals, preverbal human infants, and adult humans engaged in conflicting verbal tasks use them, suggesting a common cross-species mechanism for representing quantity (e.g., Cordes, Gelman, Gallistel, & Whalen, 2001; Meck & Church, 1983; Xu & Spelke, 2000).

There is a growing psychophysical literature on the nonverbal representation of these basic quantities (Brannon & Roitman, 2003; Dehaene, 1997; Gallistel & Gelman, 2005). A well-established finding is that performance obeys Weber's law; that is, the ease with which two subjective quantities (e.g., numbers) can be ordered is proportional to their ratio (Brannon & Terrace, 2000, 2002; Buckley & Gilman, 1974; Dehaene, Dupoux, & Mehler, 1990; Moyer & Landauer, 1973; Parkman, 1971). This Weber-characteristic is also evidenced when subjects must repeatedly state the number of rapid arrhythmic flashes they have seen or rapidly press a button a given

number of times without verbal counting; the standard deviation of the resulting distribution is proportional to its mean (Cordes et al., 2001; Whalen, Gallistel, & Gelman, 1999). This scalar characteristic of the variability has led to the widely accepted hypothesis that discrete quantity (number) is represented nonverbally by continuous, noisy symbols (or signals) called *mental magnitudes*.

It has been shown that both animal and human subjects can do arithmetic with these nonverbal numerical symbols (Barth et al., 2006; Barth, La Mont, Lipton, & Spelke, 2005; Brannon, Wusthoff, Gallistel, & Gibbon, 2001; Cordes, King, & Gallistel, 2007; Gallistel & Gelman, 2005; Gibbon & Church, 1981; McCrink & Wynn, 2004; Pica, Lemer, Izard, & Dehaene, 2004). For example, one particular line of work with bilingual adult humans comparing "exact" (language-dependent) versus "approximate" (thought to employ nonverbal magnitudes) arithmetic finds that when subjects must choose the exact answer to arithmetic problems (e.g., 16 + 18 = 41 or 34), the reaction time is longer when the problem is posed in the subject's second (nonnative) language (Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999; Spelke & Tsivkin, 2001). By contrast, when subjects must only approximate the answer (give the response closest to the correct answer—e.g., 16 + 18 = 41 or 36), they choose equally rapidly regardless of the language of presenta-

S. Cordes, scordes@duke.edu

tion. These studies and others suggest that while exact arithmetic engages arithmetic skills taught in school and dependent on language, approximate arithmetic taps into primitive nonverbal calculation abilities. These nonverbal arithmetic abilities, presumably employing nonverbal quantity representations, are language-independent and inexact—mean responses approximate the correct answer, yet responses are quite variable.

Despite the growing literature on nonverbal arithmetic capabilities, little is known about the sources of the noise in nonverbal representations of number, nor about the properties of the computations that take these noisy mental magnitudes as inputs. For example, neural network models of arithmetic have varied greatly in where noise is implemented, simply because this information is currently not known (Verguts & Fias, 2004; Zorzi, Stoianov, & Umiltà, 2005). One approach to the behavioral analysis of these processes is by way of a noise analysis (see Church, Meck, & Gibbon, 1994; Gibbon & Church, 1992): How does the variability and resulting uncertainty in the representation of the inputs propagate through the computation to determine the variability in the result? In this article, we analyze the propagation of noise through the nonverbal computation of numerical sums and differences.

We compare the variability in the performance of two nonverbal numerical tapping tasks by the same subjects. The first task requires the subjects only to tap out the number of rapid, arrhythmic flashes they have just seen. The second requires them to compute the sum or difference of two sequences of flashes and tap out the (signed) result. The reasoning behind this experimental comparison is explained by reference to the flow diagrams in Figure 1.
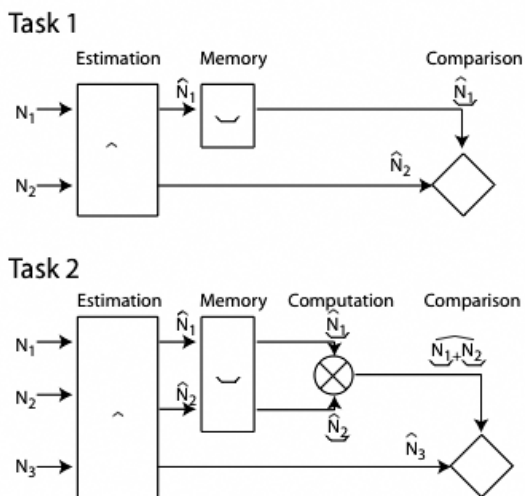

Figure 1. Flow diagrams decomposing the two tapping tasks into stages at which noise might enter: estimation, memory, computation, and comparison. The diamond represents the comparison of two estimates (one remembered and one online); the circle with an X represents unweighted linear combination (either the addition or the subtraction) of two remembered numbers. N_i, objective number (of flashes); $\hat{N}_i$ ("N_i-hat"), the online subjective estimate of N_i; $\underset{\smile}{\hat{N}_i}$ ("N_i-hat-basket"), the remembered estimate.

The first task (our baseline task) requires the subject to estimate the number of flashes seen ($\hat{N}_1$), remember the estimate, and compare that remembered number to their current (online) estimate of the accumulated number of taps they have so far made ($\hat{N}_2$). The subject stops when this current estimate (number of taps) equals the remembered number of flashes. Noise (variability) may enter at any or all of the stages. The nonverbal enumeration process that generates an online estimate of the number of events (flashes or taps) may generate different estimates from one occasion to the next (estimation noise, that is, variable subjective numbers for repetitions of the same objective number). An estimate made on one trial may be forgotten or confused with an earlier estimate or otherwise degraded between the time when it is read into memory and the time when it is retrieved from memory for comparison with the accumulating number of taps (memory noise). And the comparison process may itself be noisy, as it is in diffusion models of the process (Buckley & Gilman, 1974; cf. Ratcliff & Smith, 2004).

In the second task (an arithmetic task), the subject must estimate and remember the number of flashes in the first sequence ($\hat{N}_1$), estimate and remember the number of flashes in the second sequence ($\hat{N}_2$), add or subtract the two remembered numbers and compare the resulting subjective sum or difference ($\hat{N}_1 \pm \hat{N}_2$) to an online estimate of the accumulated number of taps ($\hat{N}_3$). Relative to the first (baseline) task, the second also has the potential for noise generated from the computational process (as postulated by Zorzi et al., 2006) and additional memory noise (since more values are stored in memory).

The expected results of a comparison of the response variability in the second (arithmetic) task to that of the first (baseline) task depend on which sources of noise one assumes predominate. If, for example, memory is the dominant source of noise (that is, much greater than the other sources), then one would expect the variance in responding for a given sum or difference in the arithmetic task to be the sum of the variances observed for the two operands in the baseline task, because the variance of the sum or difference of two independent random variables is the sum of their variances ($\sigma^2_{x\pm y} = \sigma^2_x \pm \sigma^2_y$). It does not matter whether the variables are added or subtracted; the variance in the result depends only on the magnitude of the operands, not on the magnitude of the result: The expected variance for $15 + 15 = 30$ is the same as for $15 - 15 = 0$.

If, on the other hand, all of the noise is in the comparison process, then the variability in the second task should depend only on the magnitude of the result, not on the magnitude of the operands. Moreover, it should be the same in the second task as in the first, because in both cases there is only one comparison stage. Thus, the expected variance for $15 + 15 = 30$ is the same as for straightforward estimations of 30 in the baseline task (no computation).

Our analysis also bears on the question of the quantitative relation between objective and subjective numerical magnitude, at the point where the mental magnitudes are added and subtracted. One common hypothesis explaining the Weber-characteristic of numerical representations is that subjective number is proportional to the logarithm of

objective number and that the noise in a subjective magnitude is independent of the magnitude (i.e., $\sigma = k$, a constant, where σ is the standard deviation in a distribution of mental magnitudes; Buckley & Gilman, 1974; Dehaene, 2001, 2002; Moyer & Landauer, 1967). An alternative, less popular hypothesis, called the *scalar variability hypothesis*, is that subjective number is proportional to objective number and so is the noise (i.e., $\sigma = kN$, where N represents objective number; Gallistel & Gelman, 1992; Gibbon, 1977; Meck & Church, 1983). Although both hypotheses account equally well for the behavioral data from numerical estimation and discrimination tasks, there is debate as to whether this is also the case for data from tasks requiring nonverbal computations (see Dehaene, 2001, and Gallistel, Brannon, Gibbon, & Wusthoff, 2001). One of the insights from the theory of measurement is that mental (brain) quantities cannot be measured by behavioral methods except at points of combination (Stevens, 1951). Thus, the question as to how well each theory accounts for magnitude arithmetic is a crucial one in this debate.

In the case of nonverbal arithmetic, the scalar variability hypothesis predicts that subjective sums and differences scale proportionally with the objective sums and differences—that is, arithmetic is a straightforward process. The logarithmic hypothesis, on the other hand, requires the further assumption that the sums and differences of subjective numbers are computed by means of look-up tables (Dehaene, 2001), because straightforward addition and subtraction simply do not exist in the logarithmic domain $[\log(N_1 \pm N_2) \neq \log(N_1) \pm \log(N_2)]$. The only mathematically defined way (the only closed procedure) for obtaining the log of the sum or difference of two objective quantities from their logarithms (the putative corresponding subjective quantities) is by taking antilogs and then adding or subtracting those antilogs. When the antilogs of the subjective quantities have been taken, the resulting subjective quantities, which are the ones that enter into the subjective combinatorial operation (subjective addition or subjective subtraction), are proportional to the corresponding objective quantities—in accord with the scalar variability hypothesis. Models that avoid the intermediate step of taking antilogs do so by resort to look-up tables (Dehaene, 2001).

If the logarithmic hypothesis is augmented to include look-up tables for implicit arithmetic, then hard-to-predict things should happen to variability when the subjective difference of two values approaches 0 and becomes negative, since the logarithms of these values do not exist. To illustrate, given a subtraction table (much like the multiplication tables taught in school) with rows corresponding to the subjective representation of the minuend and columns corresponding to that of the subtrahend, then one can mentally locate the row corresponding to log 8 (subjective representation of 8) and the column corresponding to log 5 (subjective 5), and discover that the value at their intersection, representing their difference, is log 3, and respond accordingly without problem. However, when one must subtract (e.g., 5 − 5 or 5 − 8), what value is found at the corresponding intersection? The logarithm of 0 is minus infinity, and negative quantities do not have logarithms. Whereas the scalar variability hypothesis has

clear predictions of systematic responding to computations involving a zero or negative difference, the logarithmic account does not. Thus, it is particularly interesting to compare the variability in subtraction answers around 0 to the variability well away from 0.

The goals of the present study are twofold. First, by comparing response variability obtained in arithmetic tasks to that of a baseline reproduction task, we provide an analysis of both the quantity and quality of the various sources of noise in the resultants of nonverbal arithmetic. Second, by comparing variability in the subtraction task when the difference is near 0 with variability for differences well away from 0, we ask whether there is evidence for a singularity at 0, the existence or absence of which distinguishes between the logarithmic versus nonlogarithmic hypotheses.

METHOD

Subjects

Four undergraduate students (2 male, 2 female) at Rutgers University, New Brunswick, participated in this study. Subjects were paid to participate in a total of five 1-h sessions.[1]

Materials and Stimuli

PsyScope 1.01 was used to present stimuli on a 19-in. color Macintosh computer screen. A Power Macintosh G4 was used to record responses. Subjects indicated their responses using both a lab-constructed response box with two large buttons (connected to a PsyScope button box) and the computer keyboard.

Stimuli construction. Stimuli were sequences of rapid, arrhythmic dot flashes. Dots were formed by printing the character "l" in Wingdings black font, size 120, on a white background. Both dot flash durations and interflash interval durations were randomly chosen from a geometric (i.e., discrete exponential) distribution with an expectation of 100 msec (minimum of 50 msec). This prevented the minimum flash duration from being less than 50 msec, and made the expected flash + interflash duration 200 msec. These parameters made the dot flash sequences both arrhythmic and extremely rapid—making verbal counting and/or timing both highly inefficient and inaccurate. In the baseline condition, the flash sequence was presented in the center of the screen. In the subtraction and addition conditions, flash sequences were presented sequentially on the left and right sides of the screen (25% and 75% of the horizontal dimension, centered vertically). The order of side presentation (left vs. right) was randomly determined on each trial. The number of dot flashes presented on any given trial was also randomly chosen from a geometric distribution with an expectation of 15 (minimum 1, maximum 100).

Procedure

The experiment took place in a dark room with the experimenter seated nearby to ensure subject compliance with instructions and to answer questions. Subjects were told that their responses were not expected to be accurate, but to perform to the best of their ability. It was stressed that subjects were not to count verbally at any time during the session, but just to get a "feel" for the numbers, and to make all button responses as rapidly as possible.

All subjects participated in the baseline condition first (one session), while the order of participation in the subtraction (two sessions) and addition (two sessions) tasks were counterbalanced across subjects. Each session had two scheduled 5-min breaks. Subjects were encouraged to take additional breaks when necessary. Practice trials were presented at the beginning of each new session as well as following both of the 5-min breaks during the session. All sessions were audio-recorded.

Baseline task. All trials began with a *Ready?* prompt in the center of the screen. When ready, the subject pressed a button and immediately saw a dot flash sequence in the center of the screen. Following the termination of the sequence, subjects were prompted with *How many?* in the center of the screen. At this point, subjects rapidly pressed a button on the button box as many times as they felt they had seen the dot flash. To prevent the counting of buttonpresses, subjects were asked to say "the" with every press. Thus, they said "the" at the moment when they would ordinarily say a count word if verbally counting (vocally or subvocally), and, because subjects pressed as rapidly as they could, the stream of "the"s formed a continuous and crowded articulatory sequence, precluding verbal counting (this method has been employed as an effective measure of articulatory suppression of count words by both Logie & Baddeley, 1987, and Cordes et al., 2001). Subjects then pressed the computer space bar to indicate completion of the trial. Subjects participated in a total of 150 trials in the 1-h baseline session.

Subtraction task. All trials began with a *Ready?* prompt on either the right or left side of the computer screen (position of the first prompt was randomized across trials). After a buttonpress, subjects were presented with a dot flash sequence on that same side. Again, subjects were asked to refrain from counting the flashes verbally, just to get a "feel" for how many flashes they saw. Subjects then saw another *Ready?* prompt on the opposite side of the screen. After another buttonpress, another dot flash sequence was presented on that side of the screen. Following the presentation of both sequences, subjects saw the *How many?* prompt in the center of the screen. At this point, subjects rapidly pressed the button (on the button box) corresponding to the side with fewer flashes as many times as they felt was the numerical difference between the two sequences of flashes without counting, again while saying "the" with every press. For example, if a subject saw 9 flashes on the right side and 15 flashes on the left side, she would rapidly press the right button approximately six times. Subjects then pressed the computer space bar to indicate completion of the trial. Occasionally, the numbers of flashes on each side were the same. Subjects were instructed that in this case, to indicate a zero response, just to press the space bar (without pressing the button box). The subtraction task consisted of a total of 300 trials, 150 trials per session.

Addition task. Trials in the addition task were exactly the same as the subtraction task except that in addition trials, subjects were instructed to press the button corresponding to the side with fewer flashes as many times as they felt was the *sum* of the two sequences of flashes. So, given stimuli as in the above example (9 on right, 15 on left), the subject would rapidly press the right button approximately 24 times. All other trial attributes were identical.

Data Analyses

Those trials requiring the greatest number of presses were excluded from data analyses; that is the baseline, addition, and subtraction data were sorted by objective number, sum, and absolute difference (of flashes), respectively, and the last 10% of trials were excluded (amounting to a total of 15 trials in baseline, and 30 trials in each of the arithmetic conditions per subject). The decision to employ this data-limiting strategy was made prior to analyses for a number of reasons. First, the number of flashes in each flash sequence was randomly chosen from a geometric distribution. Thus, the data points obtained in this larger flash range were relatively sparse, and may have provided less reliable estimates of the variability in this range. Second, since these trials were those requiring subjects to make the most number of presses, there were concerns regarding possible subject fatigue adversely affecting the data in this range. Finally, results from previous pilot work using fixed target values (as opposed to values chosen from a distribution) showed signs of anchoring effects, such that subjects tended to grossly underestimate the larger flash sequences. This result initially influenced our decision to choose flash sequence values from a distribution (as opposed to from a list of predetermined targets), and later provided a similar rationale for excluding the final 10% of the data.

Infrequently, the button box recorded two buttonpresses when the subject made only one press. In these cases, the recorded interresponse time was either negative or only a few milliseconds—a physically impossible feat. Therefore, buttonpresses with a recorded interpress interval of less than 5 msec were excluded from data analyses. Also, interresponse intervals greater than three standard deviations away from the mean were excluded from analysis, so as to avoid an inaccurate reflection of the average responding.

For purposes of data scoring and analysis, data from the subtraction task were treated as follows: Flashes on the left side of the screen were attributed negative values and flashes on the right side were attributed positive values. Because the subject's presses were compensatory (opposite in sign to the perceived difference in the number of flashes), they were given the opposite sign: presses on the left button were scored as positive and presses on the right button as negative. Thus, for example, in a trial in which there were 6 flashes on the left and 8 on the right and the subject pressed the right key three times, the scored answer would be -3 (a response indicating a perceived difference of 3 in favor of the left side when the true difference was 2 in favor of the right side). The signed error would be -3 (response) $- (+2)$ (true difference) $= -5$ (a left-directed error of magnitude 5). On a trial with 11 flashes on the left and 2 on the right (a true difference of -9), in which the subject pressed the right button seven times (perceived difference of -7), the scored answer would be -7, and the signed error $-7 - (-9) = +2$ (a right-directed error of magnitude 2).

RESULTS AND DISCUSSION

Data from the baseline condition (Figure 2) were comparable to nonverbal counting data obtained in previous reproduction studies (Cordes et al., 2001; Whalen et al., 1999). Plots of the data revealed the number of presses (z) increased as an approximately linear function of the number of flashes presented (x), as did variability (i.e., the standard deviation of the distribution of presses); that is, $z = sx$ and $\sigma_z = wsx$, where s is the slope of the regression line relating objective to subjective number and w is the the Weber fraction (coefficient of variation).[2] This pattern of scalar variability is ubiquitous in both the human and animal nonverbal counting data.

Data from the addition and subtraction conditions are plotted in Figures 3 and 4, respectively. The number of presses made increased in proportion to the sum (addition) or difference (subtraction) of the number of flashes, indicating that subjects could do the requested combinatorial operation and base their responding upon the resulting mental magnitude. Results from the both the baseline and arithmetic tasks demonstrate that humans represent numerosity nonverbally while providing some of the first direct evidence that these representations are subject to arithmetic manipulation. The finding that nonverbal arithmetic was a feasible, nontrivial task for our subjects is significant in that it strengthens the claim that computations involving these magnitude representations may serve as a basis for learning.

Variance Analyses

The variability in the addition data appears to increase as a function of the objective sum of the operands but not the variability in the subtraction data. Why is this the case? If we assume that the variability in the baseline condition is dominated by variability in the underlying mental
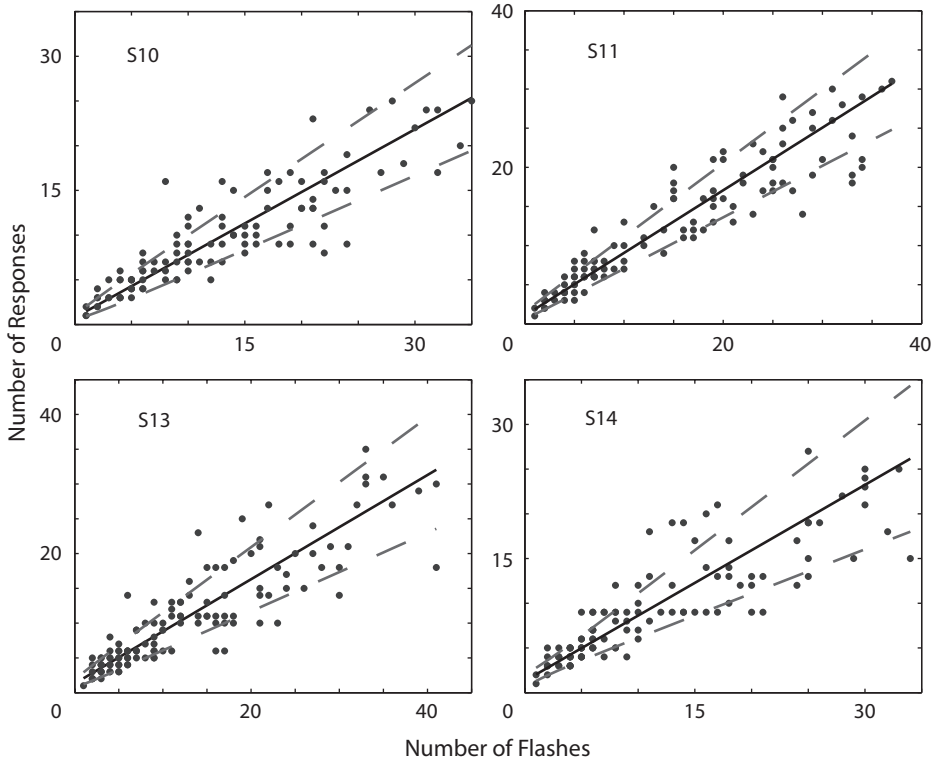
Figure 2. Number of responses in the baseline condition from all subjects plotted as a function of the number of flashes presented. The solid lines through the data are the regression lines and the dashed lines represent ±1 standard deviation. The regression parameters and the estimates of the standard deviation (as a function of number of flashes) were determined by fitting the data with the Killeen variance equation (Equation 4).

magnitudes (as opposed to the comparison process) and that there is no source of covariance in the two estimates made on each combinatorial trial, then the input variances will add. That is, the variance obtained in our addition and subtraction tasks should equal the sum of the variances associated with the individual operands (as determined by our baseline data), because, for random normally distributed variables, x and y,

$$\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y + 2\text{cov}(x,y) \qquad (1)$$

and

$$\sigma^2_{x-y} = \sigma^2_x + \sigma^2_y - 2\text{cov}(x,y), \qquad (2)$$

where σ^2_{x+y} and σ^2_{x-y} are the variances of the sums and differences, and σ^2_x and σ^2_y are the variances of the operands.

Note that with independent operands (hence, zero covariance), the variances associated with the sum and the difference of any pair of operands are the same; they are both equal to the sum of the operand variances. When operand variability scales with operand magnitude

$$\sigma_{x+y} \text{ or } \sigma_{x-y} = \sqrt{\sigma^2_x + \sigma^2_y}$$
$$= \sqrt{(wsx)^2 + (wsy)^2} = ws\sqrt{x^2 + y^2}. \quad (3)$$

That is, the variability in the results will depend on $\sqrt{x^2 + y^2}$, which we call the *magnitude* of an operand pair, by analogy with the magnitude of a complex number. On this analysis,

answer variability should covary with answer magnitude in the case of addition, because the magnitude of the answer covaries with the magnitude of the operand pair (thus the clear pattern of scalar variability in Figure 3). In the case of subtraction, however, it does not; pairs of arbitrarily small or large operands can yield the same answer—for instance, $3 - 2 = 1$ and $9,999 - 9,998 = 1$ (thus, when the data are plotted as a function of the answer as in Figure 4, variability patterns appear unsystematic). The variability patterns in Figures 3 and 4 therefore imply that answer variability is dominated by operand variability, variability in the inputs to the combinatorial operations. The question then becomes, to what extent is the variability in the answers predictable from the variability in the operands? That is, is arithmetic variability purely a function of noise in the representation of the operands, or do other sources play a role?

Within-Task Variance Characterization

Before predicting the variability in the addition and subtraction conditions from the variability in the baseline condition, we must accurately characterize variability under baseline conditions. Is it in fact purely scalar?

Baseline condition. To answer this question, we did a weighted linear regression, minimizing the sum of the squared deviations weighted by the inverse of the square of x. Deviations were weighted in this way to offset the approximately scalar heteroscedasticity evident in the
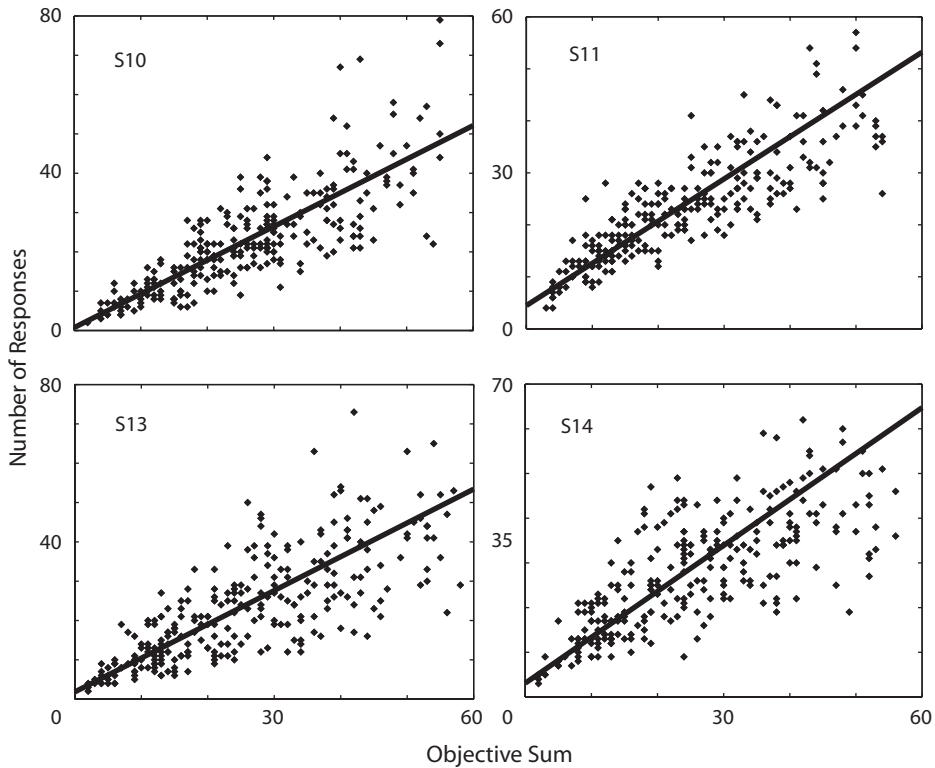
Figure 3. Number of responses in the addition condition plotted as a function of the correct sum of the number of flashes presented. Lines through the data were obtained through simple linear regressions of the data (not used in data analyses, but drawn here to illustrate the linear nature of the data). Note the clear exhibition of scalar variability in the data.

plots, the tendency of the variance (σ_z^2) to increase as the square of the predicted value. If the variability (i.e., the standard deviation, σ_z) about the resulting regression line, $z = ax + b$, increases strictly in proportion to the value of x (the independent variable)—that is, if it is scalar—then there will be no trends in the residuals when they have been normalized by the predicted variability. That is,

$$\frac{|z - z'|}{\sigma_z}$$

should be independent of x. Alternatively, trends in the normalized residuals would suggest the influence of non-scalar noise sources.

The normalized residuals are plotted in Figure 5. They show a clear trend: Large deviations from predictions are more common for small test values. This implies nonscalar components of the variability in these data, and in particular components that make a relatively greater contribution to overall variability at small values than at large.

Killeen and colleagues (Fetterman & Killeen, 1995; Killeen & Weiss, 1987) have characterized the variance associated with interval timing and counting as the composition of scalar, binomial, and additive variances:

$$\sigma_x^2 = Ax^2 + Bx + C, \qquad (4)$$

where the coefficient C corresponds to additive sources of variability, B to binomial sources, and A to scalar sources of noise (where $\sqrt{A} = ws$). What leads to the influence of

these additional, nonscalar sources? Additive sources of variability are brought about by factors such as a failure to immediately stop pressing the button once the target number of presses is attained, and a failure to stop the nonverbal count process immediately after the last flash on the screen (resulting in counting more flashes than seen); that is, trial-to-trial variations whose magnitude is independent of the number being estimated. Counting errors—errors from skipped items or double counts—yield binomial variability. Binomial variability arises whenever there is an approximately constant probability of deviation for each item or event in an array or sequence.

As Equation 4 makes clear, the contributions of both additive and binomial variability relative to the scalar component must diminish as the number of flashes gets larger, because additive sources are independent of the number of flashes (i.e., constant) and binomial variability grows in proportion only to the square root of the number of flashes. Thus these additional sources of variance must be responsible for the pattern obtained in Figure 5.

We fit the Killeen variance equation to our data by an iterative search for the maximum likelihood values for A, B, and C. The parameters of the regression line are an analytic function of the variance parameters (see Appendix B), so, in determining the values of A, B, and C, one determines the value of the slope and intercept of the regression line. The maximum likelihood values for these parameters are given in Table 1. Generalized likelihood
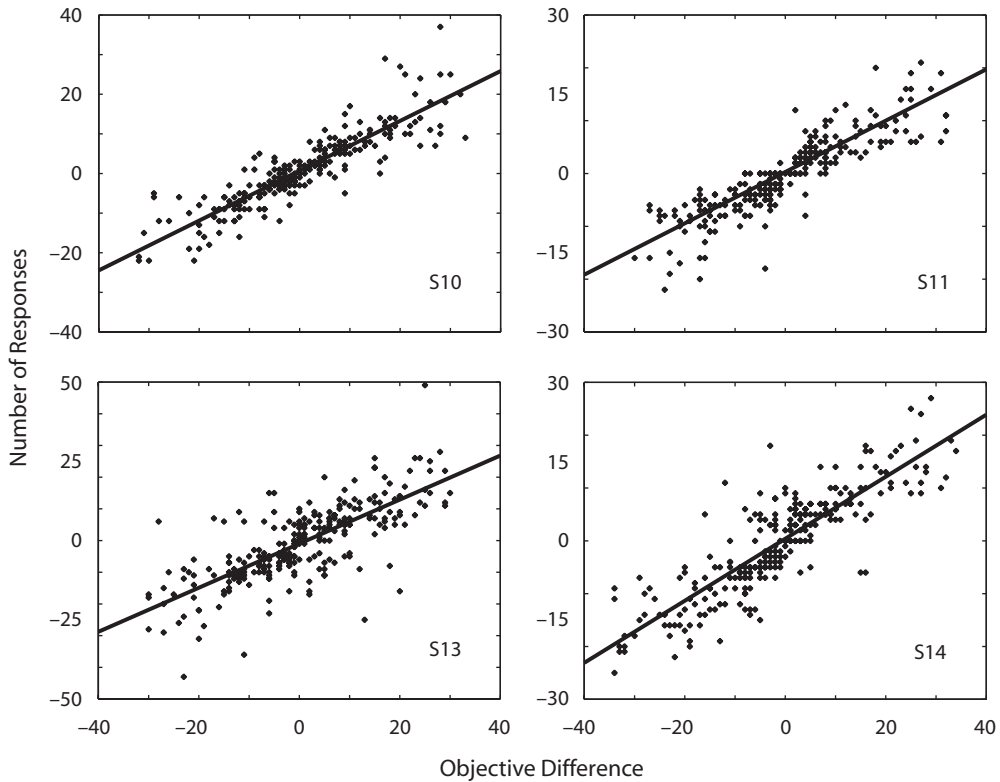
Figure 4. Number of responses in the subtraction condition (right responses given negative sign, left responses positive sign) plotted as a function of the correct difference in the number of flashes presented in the two flash sequences (number of flashes on right minus number on left). Gross linear fits of the data, obtained through simple linear regressions of the data, are plotted as well to illustrate the linear nature of the data. Notice the lack of a clear pattern of response variability.

ratio tests (Mood & Graybill, 1963) confirmed that the fits of the data using the Killeen variance equation (Equation 4) were significantly better ($p < .001$) than the purely scalar fits of the data in all 4 subjects [in Subject 10 (S10), $\chi^2(2, N = 135) = 15.83$; in S11, $\chi^2(2, N = 135) = 33.13$; in S13, $\chi_v^2(2, N = 135) = 25.25$; in S14, $\chi_v^2(2, N = 135) = 25.68$]. This is the first time these alternative (nonscalar) variability sources have been estimated in human nonverbal counting data.

Plots of the baseline data with the new regression lines and lines of ± 1 standard deviation show that this model provides a fairly accurate fit of the spread around the regression line (Figure 2). As can be seen in Table 1, however, there are noticeable differences in exactly how much contribution each noise source made to each individual's

Table 1
Variance and Regression Parameters Obtained Through Fits of the Baseline Data Using the Killeen Variance Equation (Equation 4): $\sigma_x^2 = Ax^2 + Bx + C$

	Regression	Additive	Binomial	Scalar
S10	$0.70x + 0.78$	0	0.27	0.021
S11	$0.80x + 1.05$	0.20	0.19	0.020
S13	$0.75x + 1.32$	0.31	0.39	0.033
S14	$0.73x + 1.31$	0.47	0	0.057

Note—S10, Subject 10; S11, Subject 11; S13, Subject 13; S14, Subject 14.

data. For example, the likelihood fits of S10's data did not reveal any additive variability but revealed a contribution of binomial variability, whereas S14's data revealed the opposite pattern. What can account for these individual differences? There are two likely contributing factors. First, it is highly probable that there are both individual and day-to-day differences in factors related to performing in this task. For example, differences in attention may have lead to differences in the number of items not counted or double-counted and/or in the amount of time it took to stop the nonverbal count. Our ability to specify the sources of these individual differences is limited by our lack of certainty about the true values of the variance coefficients for a given subject. That is, though we can be fairly certain that nonscalar influences affected our data, the precision in our maximum likelihood estimates for the scalar and binomial variance components is low. Contour plots of the loss surfaces showed strong negative covariation between these estimates; very different combinations of the binomial and scalar variance coefficients gave equally likely fits (Figure 6). A similar trade-off is not found with the additive component, suggesting that estimates for this parameter are reasonably independent of the estimates for the other variance parameters. Although discussions of the total amount of response variability from each subject can be conducted with confidence, the relative amounts of sca-
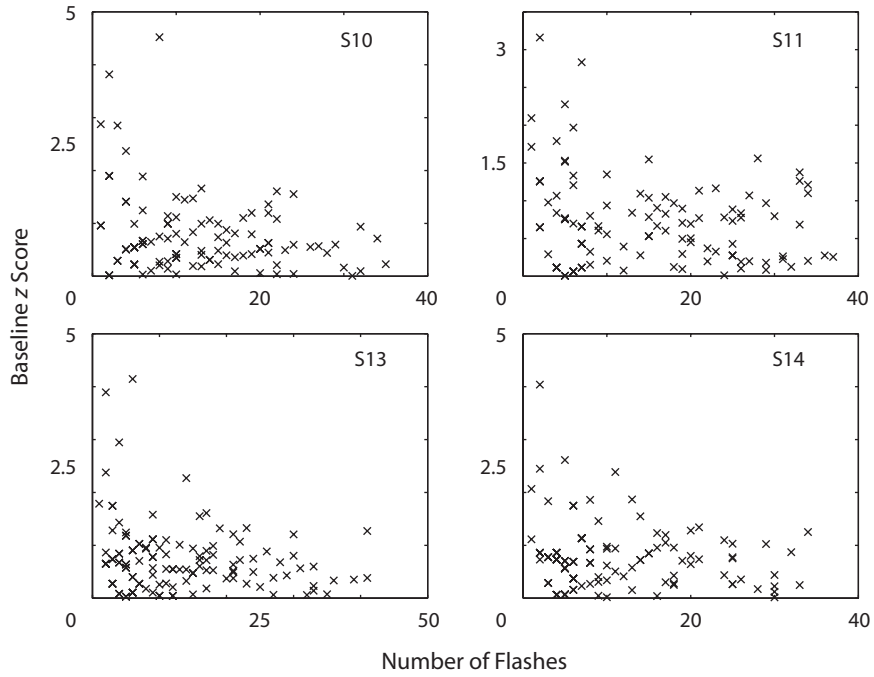
Figure 5. The *z* scores (absolute value of the difference between the observed and weighted regression-predicted values divided by the standard deviation at that point) from the baseline condition plotted as a function of the number of flashes presented.
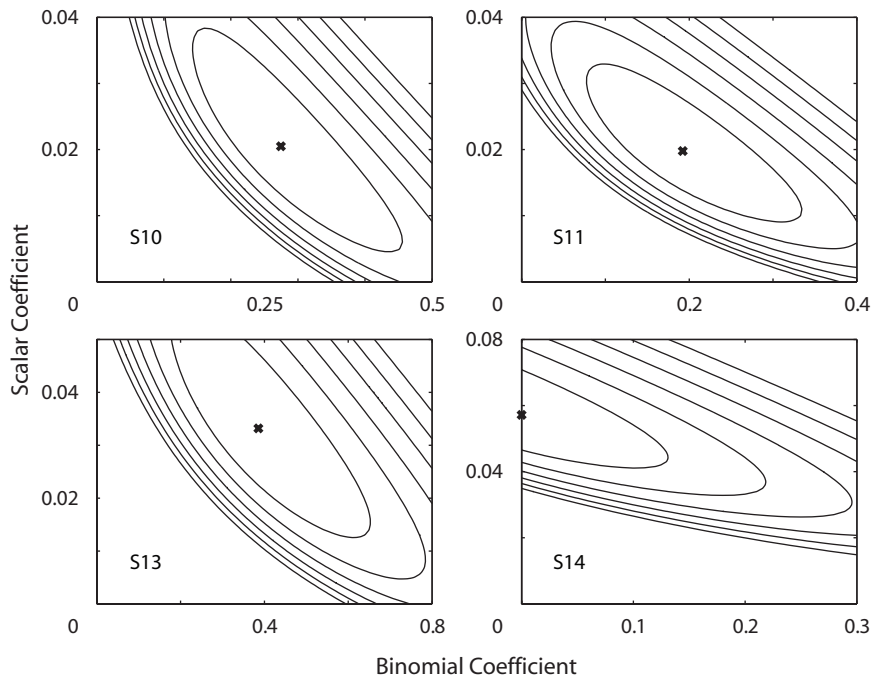

Figure 6. Equal-likelihood contour plots of the amount of scalar variability versus binomial variability attributed to each of the fits of the data for the baseline condition. The contour lines are at common log unit intervals; each more remote line represents a reduction by a factor of 10 in likelihood. *X* denotes the coefficient values returned by the parameter search.
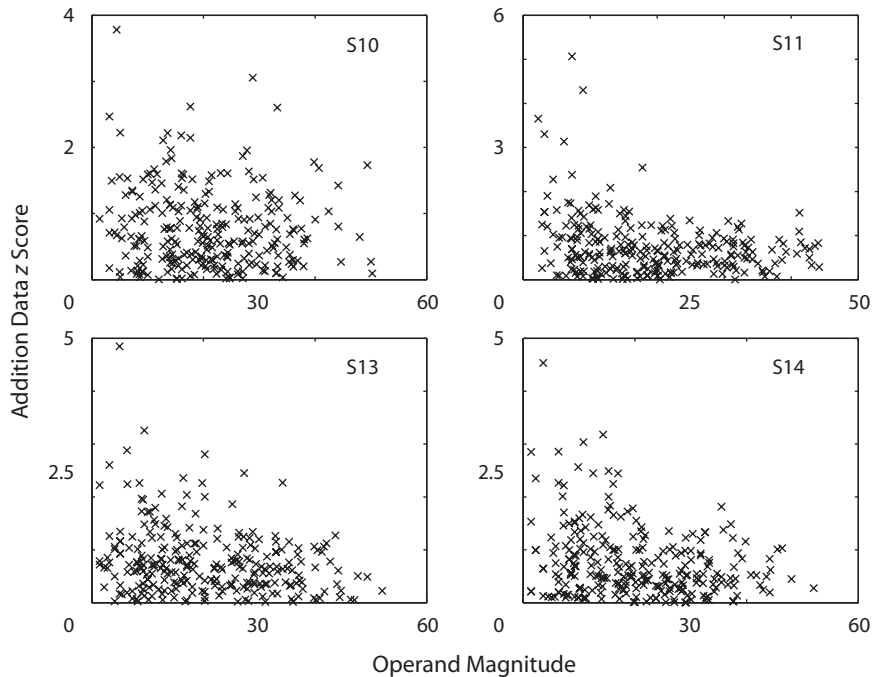
Figure 7. The z scores (absolute value of the difference between the observed and regression-predicted values divided by the standard deviation at that point) from the addition condition plotted as a function of the magnitude of the number of flashes presented.

lar and binomial variability in the data from a given subject cannot be estimated with any confidence.

Arithmetic data. Given the presence of both additive and binomial variabilities in the baseline data, it was likely these sources also played a part in the variance obtained in the addition and subtraction cases. Preliminary analyses confirmed this. Again, if response variability were exclusively a function of noise in the representations of the operands (i.e., variability was scalar with respect to the operands only), then there would not be a relationship between the distribution of z scores and the magnitude of the operands; that is, variability in the plots would be flat. Plots of the z scores obtained from multiple weighted regressions of the arithmetic data (with weights again inversely proportional to the square of the magnitude of the operands $\sigma_i^2 \sim x_i^2 + y_i^2)^3$ as a function of the magnitude of the operands revealed trends (see Figures 7 and 8). The z scores followed a pattern similar to that seen in the baseline data. Both sets of z scores showed relatively more variability when subjects were presented small numbers of flashes, again suggesting additive and/or binomial components.

To determine exactly how much noise was contributed by these nonscalar sources, the variance of the arithmetic data was also fit by iterative search for the maximum likelihood parameters of the bivariate form of the Killeen variance equation.

$$\sigma_{x\pm y}'^2 = A'\left(x^2 + y^2\right) + B'(|x| + |y|) + 2C'. \quad (5)$$

We first examined the simplest-case scenario—the case in which the impact of variables other than inherent noise in the representations is minimal. Thus, calculations were

based on the following two assumptions: (1) representations of the two operands were independent [i.e., $\text{cov}(x, y) = 0$] and (2) the implementation of an arithmetic operation did not create any additional noise in the representations. As in the baseline case, the regression parameters were analytically derived from the variance parameters returned by the search. (Derivations of the loss equation and the regression parameters are described in Appendix B.)

The parameters obtained from this new search appeared to describe the data well. The new z scores obtained from the Killeen fit of the data were plotted as a function of operand magnitude, answer magnitude, and predicted standard deviation. None of these plots revealed significant trends, suggesting that the Killeen parameter search led to an appropriate representation of the scatter in the arithmetic data. This was confirmed by generalized likelihood ratio tests comparing the new fits with the purely scalar model. The Killeen variance fits were a significantly better characterizations ($p < .0001$) of the data than the purely scalar model in both the addition and subtraction data for all subjects, except for S10's subtraction data [addition data for S10, $\chi^2(2, N = 271) = 126.38$; for S11, $\chi^2(2, N = 271) = 68.86$; for S13, $\chi^2(2, N = 270) = 20.07$; for S14, $\chi^2(2, N = 271) = 51.12$; subtraction data for S10, $\chi^2(2, N = 271) = 3.96, p = .14$; for S11, $\chi^2(2, N = 272) = 81.29$; for S13, $\chi^2(2, N = 269) = 77.34$; for S14, $\chi^2(2, N = 273) = 89.05$].

Across-Task Variability

Given the uncertainty in the estimates of the variance components (the values for the variance parameters A, B,
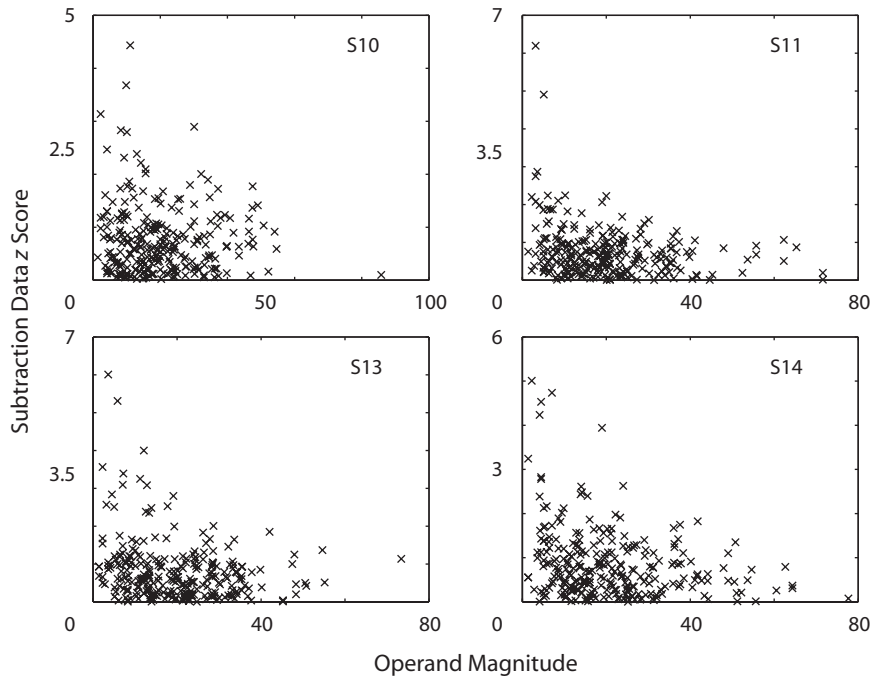
Figure 8. The *z* scores (absolute value of the difference between the observed and regression-predicted values divided by the standard deviation at that point) from the subtraction condition plotted as a function of the magnitude of the number of flashes presented.

and *C*), predicting each component in the arithmetic data from the estimate of the corresponding component in the baseline data would not be justified. Instead, we asked to what extent the total variance estimates for the two operands obtained from the baseline condition predicted the variance estimates obtained by fitting the data from the arithmetic conditions. The model of the baseline data yields an estimate of the variance associated with each individual operand. This estimate can be compared on an operand-pair-by-operand-pair basis with the estimates for each pair obtained from the multiple-regression models for the arithmetic data. Specifically, for each operand, we obtained from the maximum-likelihood model of the baseline data the estimate of the variance to be expected. The predicted standard deviation for the sum or difference of the pair of operands was the square root of the sum of the two variance estimates (one for each operand) obtained from the baseline condition. These predictions were plotted against the observed standard deviation for that pair; that is, the standard deviation obtained from the maximum-likelihood model for the arithmetic data. In other words, we took each pair of operands presented to the subject in the arithmetic task (e.g., 18 + 7) and computed the standard deviation predicted by baseline—that is,

$$\sigma_{\text{pred}(18+7)} = \sqrt{\sigma^2_{\text{base}(18)} + \sigma^2_{\text{base}(7)}}$$

—and plotted it against the standard deviation estimated in the arithmetic task for that particular pair of operands (σ_{18+7}). Figure 9 plots the observed estimates against predicted estimates. If the variability in the responses could be predicted from the estimates of the variability in

the operands, the points in Figure 9 would cluster around a line with slope = 1 (solid line).

There are three points to notice from these plots. First, the plots from the addition task in all four subjects fall above the solid comparison line (slope = 1). This means that the standard deviations obtained from the Killeen fits of the addition data were greater than predicted. While the obtained variability is greater in all subjects, the differing shapes of the curves reflect possible discrepancies in estimates of the individual variance parameters. For example, the exaggerated curvature seen in S14's data is most likely an artifact of the lack of binomial contributions estimated in her baseline data. Regardless, it is not certain that any one source of variability can account for the discrepancy between baseline predictions and the obtained variability. For the reasons listed above (i.e., individual differences and large confidence limits on estimates of the parameters), these differences in curvature were deemed orthogonal to the question of overall variability patterns.

Second, the subtraction data do not seem to differ significantly from a slope of one in any common systematic fashion across subjects. That is, 2 subjects (S11 and S14) appear to have produced less variable data in the subtraction task than predicted by their baseline estimates, whereas 1 subject (S13) produced more variable data in this task, and the last subject's data (S10) matched predictions fairly well. So, although the addition data appear to be more variable than predictions of the simple model of additive variance, the subtraction data match predictions fairly well.

Finally, for a given pair of operands, the addition condition consistently yielded more variable data than the subtraction task. This is most evident when the standard de-
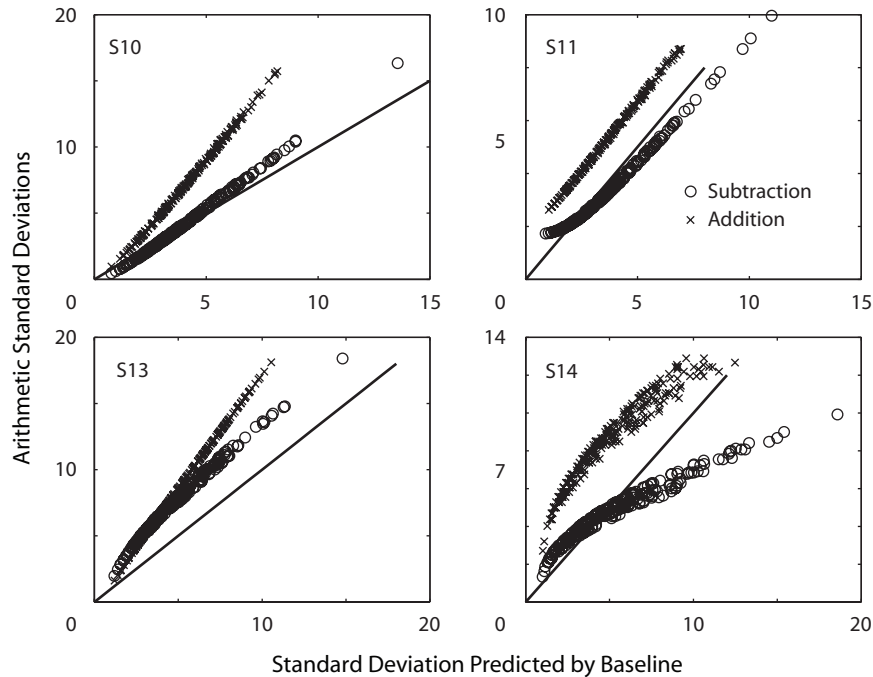
Figure 9. Standard deviation of the baseline predictions (without answer magnitude) versus the standard deviation of the Killeen fit of the arithmetic data (using Equation 5). The solid lines represent a slope of 1.

viation of the addition data is plotted against the standard deviation of the subtraction data in each subject (Figure 10). In this case, all four plots are above a slope of 1. (Recall that if variability in the answers could be predicted simply from summing the variance estimates associated with the operands, then the variance in the sum of two operands would be the same as the variance in their difference.)

The plots in Figure 9 indicate that the variability in the sums and differences of two nonverbal numerical estimates cannot be adequately predicted from the sum of the variance estimates associated with the operands via the baseline task alone. This is not surprising, because there are likely to be at least two sources of variability in the arithmetic tasks—variability in the nonverbal estimates of the number of flashes (operands) and in the representation and tapping out of the computed sum or difference (answer)—and these would be expected to obey different combination rules. When two nonverbal estimates are mentally summed or subtracted, the variance in the resulting subjective sum or difference should be the sum of the variances associated with the operands, that is to say, of the variability in those estimates. Once the sum or difference is computed, however, that estimate is then stored in memory where it is subject to additional scalar noise and then buttonpresses are counted out to this value, providing additional additive and binomial noise (counting error variability). This additional noise contributed by the representation and counting out of the computed answer should thus be scalar and binomial with respect to the magnitude of the answer (i.e., $|x + y|$).

The importance of this distinction may be appreciated by considering the extreme case. If there were no additional variability provided by the answer representation,

then (as outlined in the introduction) the variability observed in both addition and subtraction answers would depend only on the variability estimates for the operands. The variance in the addition and subtraction answers would be the same, and it would be equal to the sum of the variance estimates for the two operands obtained from the baseline data. In this case, the variances obtained on trials where subjects were required to compute the sum of say, 30 and 29, should be identical to the variance associated with the responses on those trials where the difference between these two values was computed; that is, responses as high as 59 buttonpresses should be just as variable as those around 1. This model, which is rather counterintuitive, is the one we have so far tested, and it clearly fails.

The consideration of the extreme case shows how something like the pattern of results we observe would arise. A conclusion from this consideration is that we ought to include in our regression terms that depend on the sums (and differences) of the operands. Therefore, the baseline predictions as well as fits of the arithmetic data were modified to include potential answer noise. That is, for the case of the arithmetic data, Equation 5 was modified to include variance associated with the magnitude of the answer as follows:

$$\sigma''^2_{x\pm y} = A''\left(x^2 + y^2\right) + B''(|x| + |y|) + 3C''$$
$$+ A''(x + y)^2 + B''(|x + y|). \qquad (6)$$

Notice that we have added three terms to the variance equation (Equation 5) intended to allow for a dependence on the square of the sum of the operands (scalar answer-dependent variability) and on the simple sum (binomial answer-dependent variability), and for additional additive noise. To
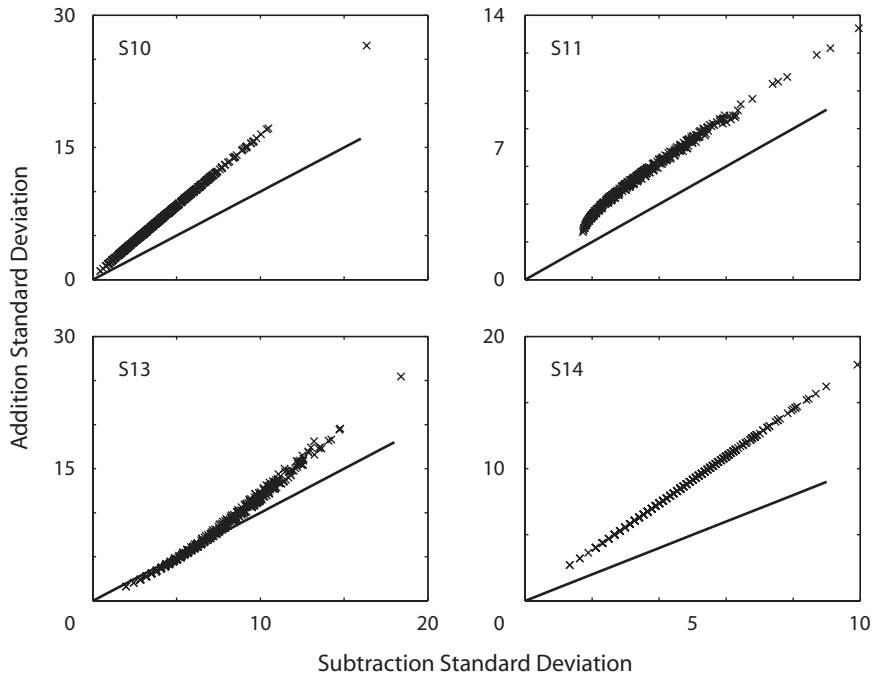
Figure 10. Standard deviation (as determined by fits using Equation 5) of the addition data as a function of the standard deviation of the subtraction data. The solid lines represent a slope of 1.

make the iterative search for maximum likelihood values tractable, we have forced the coefficients of scalar and binomial variability to be the same for these two terms as for the original two terms. In other words, we have assumed that the mixture of scalar and binomial variability is the same for the estimates of flash number (operand representations) and the answer estimates.[4] Our rationale for proceeding was simply to see whether allowing for a dependence on answer magnitude as well as operand pair magnitude improved our ability to predict the addition and subtraction data from the baseline data. These parameters obtained from this new fit of the arithmetic data are listed in Table 2.

The standard deviations obtained from these new baseline predictions were re-plotted against the standard deviations from the new fits of the arithmetic data including answer noise (Figure 11). Comparisons of the plots in Figure 11 with those in Figure 9 (the original plots without variability due to the answer magnitude) reveal that, while it is not clear whether the plots of the subtraction data benefit substantially from this new fit, the addition data from all four subjects are significantly closer to a line with slope one, indicating it is likely that the magnitude of the answer to the arithmetic manipulation also contributed to the response variability obtained.

Despite the improvement in these new fits, it is clear that the variability obtained in the arithmetic conditions does not entirely match up with that predicted by each subject's baseline performance. In addition, in 3 out of 4 subjects, the addition data remain more variable than the subtraction data. It is entirely possible that a fit that allowed the coefficients of scalar and binomial variability to be different for the estimation and the tapping-out components of the

baseline variability would do better. It is also possible that these final discrepancies could be attributed to a lack of independence in the representations of the two operands. That is, $\text{cov}(x, y) \neq 0$, such that the probability of the subject overestimating both operands or underestimating both operands on any given trial was greater than chance.

How could covariation in the operands—that is, in the subjective estimates of the two flash numerosities—arise, given that the objective numerosities were independent? Subjects might have attempted to compensate for a (incorrectly) perceived under or overestimation of the first operand by purposely over or underestimating the second. Subject interviews do not provide any evidence to suggest this is the case. A second, and perhaps more likely, expla-

Table 2
Variance and Regression Parameters Obtained Through Fitting the Arithmetic Data Using the Bivariate Form of the Killeen Variance Equation, Including Terms for Answer Noise (Equation 6): $\sigma^2_{x \pm y} = A''(x^2 + y^2) + B''(|x| + |y|) + 3C'' + A''(x + y)^2 + B''(|x + y|)$

	Regression	Additive	Binomial	Scalar
Subtraction Data				
S10	$0.69(x + y) + 0.21$	0	0.064	0.024
S11	$0.55(x + y) - 0.14$	0.767	0.048	0.012
S13	$0.74(x + y) - 0.82$	0	1.580	0.012
S14	$0.62(x + y) + 0.49$	0	0.663	0
Addition Data				
S10	$0.84(x + y) + 0.98$	0	0.161	0.036
S11	$0.71(x + y) + 6.31$	2.280	0.138	0.013
S13	$0.82(x + y) + 2.37$	0	0.519	0.043
S14	$0.86(x + y) + 5.91$	0.487	1.470	0

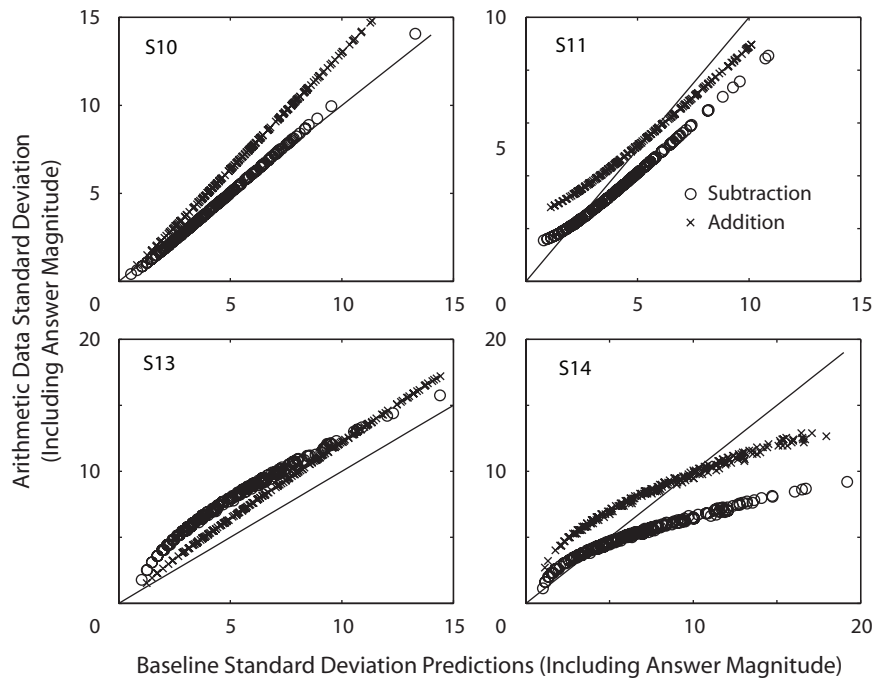Note—S10, Subject 10; S11, Subject 11; S13, Subject 13; S14, Subject 14.

Figure 11. Standard deviation of the modified baseline predictions (including answer noise) as a function of the standard deviation of the arithmetic data. The predictions and fits of the data were obtained using Equation 6 as a variance model.

nation for a nonzero operand covariation is that there is $1/f$ noise in the process of nonverbally estimating numerosities. This ubiquitous source of noise in psychophysical tasks is a fluctuation in time in psychophysical judgments, in which the power density is inversely proportional to the frequency, that is, the low-frequency, long-period fluctuations are bigger than the high-frequency, short-period fluctuations (Gilden, Thornton, & Mallon, 1995). Period is typically measured in number of trials. Thus, a low frequency fluctuation is a change in the tendency to over or underestimate numerosity that is cyclical, with cycle duration extending over many trials. Such a fluctuation would induce covariation in the subjective numerosities of the two flash sequences presented on each trial.

The arithmetic data were fit using a model including covariance. The inclusion of this component, however, did not appear to significantly improve the fit of the data, with the possible exception of S10's subtraction data. This model was not pursued any further, due to the difficulty in distinguishing the contributions of the various noise sources from the covariance, and thus a difficulty in comparing these new fits to baseline predictions.

DISCUSSION

The Loci of Variation

In summary, the variability obtained in our addition and subtraction data was primarily a function of variability in the representation of the operands and of the answer. Our results imply that the noise (variability) observed in the results from the arithmetic composition of nonverbal mental magnitudes derives primarily from noise in the represen-

tations themselves, rather than noise in the comparison process (i.e., the comparison between the representation of the sum and the number of buttonpresses made thus far). If the variability arose primarily in the comparison process, then it would depend primarily on the magnitude of the resultant, rather than on the magnitude of the operands. That is, the variability associated with $2 + 3$ would be equivalent to that associated with $58 - 53$. We find the opposite; it depends primarily on the magnitude of an operand pair.

A significant portion of the behavioral noise in the task, however, does appear to arise from the comparison process, because the variability for an operand pair with a given magnitude is substantially greater when the resultant is their sum than when the resultant is their difference. If the variability in the behavioral consequences of arithmetic composition depended only on the magnitude of the operand pair, then, for a given pair, the variability in sums and differences would be the same. There is, however, a second plausible source for the discrepancy in the variability of sums and differences. Factors that tended to covary over- and underestimates for the numerosities of the two flash sequences presented on compositional trials would also produce greater variability in the sums than in the differences, because the covariance term is additive for sums and subtractive for differences. The ubiquitous phenomenon of $1/f$ noise in psychophysical judgments makes the assumption of appreciable covariance plausible.

The two possibilities—variability in the comparison process and covariance in operand noise—are not mutually exclusive. If the discrepancy between variabilities in the sums and differences derived entirely from covaria-

tion between operands, then the two variabilities would straddle the variability predicted on the assumption of simple additivity: The sums would have greater than predicted variability and the differences less than predicted. If the discrepancy derived entirely from the comparison process, then both variabilities would be greater than predicted by the additivity of input variances, with the variability in the sums being more excessive than the variability in the differences. The actual patterns differ somewhat between subjects, but the variability in the differences tends to approximate the additive prediction, whereas the variability in the sums clearly exceeds it. This pattern is consistent with the plausible assumption that there is both some operand covariance and variability in the comparison process (variability dependent on answer magnitude).

Variability in the translation of mental magnitudes into numbers of presses is the most plausible source for the additive and binomial components of the variance in our baseline task, which involved simply duplicating in the number of one's presses the number of flashes seen. Our task requires subjects to produce behavior—a number of presses—that varies in the same way as the inputs (numbers of flashes). It may be possible to get a less contaminated, albeit more indirect, view of the variability in the underlying number-representing signals and their composition using two-alternative forced choice procedures (see Barth et al., 2006).

The Form of the Mapping From Objective to Subjective Magnitude

The Weber-law characteristic in judgments of numerical (and other) order is susceptible to at least two contrasting explanations: *Model 1*. The subjective magnitude (brain signal) is proportional to the logarithm of the objective magnitude and the noise in that signal is constant, that is, independent of signal magnitude (Dehaene et al., 1990). *Model 2*. The subjective magnitude is proportional to the objective magnitude and so is its variability (Gibbon, 1977). In either model, the overlap between two signal distributions—hence, the discriminability of the objective variables that the two signals represent—is a function of the ratio of their means (Weber's law).

Studies revealing arithmetic competence (specifically, in the case of subtraction) in nonhuman animals (Brannon et al., 2001; Gibbon & Church, 1981) support the second kind of model (scalar mapping and scalar variability) put forth by Gibbon (1977). Dehaene (2001) has suggested that mental magnitudes are proportional to the logarithms of objective magnitudes and that, to obtain from them the mental magnitude corresponding to an objective difference, the brain uses a neuronal look-up table. A look-up table uses the subjective representations of two inputs as row and column header values that together determine the location (cell) within the table that contains the resultant. The table overcomes the fact that logarithmic transformation is a nonlinear operator—that is, $\log(x_1 + x_2) \neq \log(x_1) + \log(x_2)$. In this model, the value obtained from adding (or subtracting) the two input signals [$\log(x_1) \pm$

$\log(x_2)$] is not the same as the value (resultant) that the relevant table delivers [$\log(x_1 \pm x_2)$]. What it delivers is the logarithm of the sum or difference of the objective values represented by those input signals. The table is able to do this because there is no necessary connection between the values used to access a location and the value stored at that location. The retrieval of the value stored at a given address in a table is a nonarithmetic data-processing operation.

Any version of a table look-up model for combining mental magnitudes that are a logarithmic function of the objective magnitudes they represent must find a way to deal with the singularity in the logarithmic mapping at 0. The logarithm of 0 is minus infinity, which is to say that it is a physically unrealizable answer and must be treated as a special case. Zero was the most common objectively correct answer in our subtraction condition. Most importantly, the variability in our subjects' responses in the vicinity of 0 was of a piece with the variability in answers comfortably away from 0. From the perspective of a noise analysis, there was nothing special about subjective differences at or near 0.

A table look-up model with logarithmically compressed mental magnitudes as representations of both entrant and answer magnitudes will have to have special machinery for dealing with the case where the subjective difference is 0— that is, where the numerosities of the flashes on the two sides are subjectively equal. It will also have to deal with the related problem that when subtraction is implemented in this way, it is not a closed operation. Negative quantities have no logarithms, so a mental magnitude can never be subtracted from a mental magnitude smaller than it. The direction (sign) of a difference must be determined before the table is entered. Given that the order of presentation of the large and small operands in our subtraction task was randomized, a look-up table would require the additional task of determining which input value corresponded to the row value and which to the column. This is not required in the case of a scalar mapping. How to handle differences that include zero and negative values (signed differences) in such a way that the propagation of variability from operands to answers behaves just as it does well away from the singularity at 0 presents an interesting challenge to such a model. If this challenge cannot be surmounted, then our data argue in favor of an approximately scalar mapping from objective magnitudes to subjective magnitudes, at least at the point where subtraction is implemented.

Alternative Estimation Strategies

We assume that our subjects based their responses on estimates of the numbers of flashes in the flash sequence and that they did not make those estimates by means of subvocal verbal counting. Here, we review the thinking behind those assumptions.

Did subjects verbally count flashes? The speed of the flash sequences made it unlikely that subjects used verbal counting to keep track of the number of flashes. Subjects reported that they felt the flash sequences were presented much too quickly to be counted. They expressed surprise at the rapidity of the flash sequences, and doubted, before-

hand, their ability to extract the numerosity of the display. The experimenter in the testing room could also verify that subjects did not vocalize count words. The expected durations of both the dot flashes and interflash intervals were very short—100 msec each (resulting in a flash + interflash interval duration expectation of 200 msec). Previous work in our lab (Cordes et al., 2001) revealed that subjects are able to verbally count events at a maximum rate of about 5.7 items/sec (which corresponds to 175 msec per item). This is longer than the *median* flash–flash interval, which was about 160 msec. Thus, the majority of flashes (approximately 60%) occurred less than 175 msec apart.

Although it is clear that subjects were unable to verbally count the flashes presented to them on the screen, it is possible that once they obtained a nonverbal estimate of the number of flashes in each sequence in the arithmetic tasks, this nonverbal estimate was converted to a verbal/symbolic representation on which the arithmetic operation was performed. That is, subjects may have seen a sequence of flashes and thought "that was about 6," the second sequence and thought "that was about 15," and computed the difference of these verbal symbols, "15 − 6 = 9," and then estimated nine buttonpresses accordingly. Although this is a possibility, none of the subjects reported engaging in such a strategy, and the conflicting articulatory suppression task made this an unlikely approach. Unfortunately, however, this possibility cannot be ruled out in the present experiments or in any other study addressing nonverbal arithmetic in numerically fluent individuals.

Did subjects verbally count their presses? Subject reports also suggested that subjects did not verbally count their presses. All subjects claimed they did not, and that they had no idea how accurate their responses were. This corroborates previous research using similar articulatory suppression techniques. While tapping their responses, subjects verbally repeated "the" with each tap. This technique, first employed by Logie and Baddeley (1987), has since been used in nonverbal counting tasks, and evidence suggests it is successful in preventing articulation of the verbal count words (Cordes et al., 2001).

Analysis of interresponse intervals [the amount of time between the nth and $(n+1)$th presses] also suggests that subjects did not verbally count their responses. Interresponse intervals were reasonably flat and unremarkable, which would not be expected if "chunking" strategies were employed when responding. The interresponse intervals were also very short, generally shorter than the previously mentioned 175-msec verbal counting criterion, again suggesting the unlikelihood of verbal counting.[5]

Were subjects timing rather than counting? In an attempt to partially deconfound time and number, the flashes in this study were presented arrhythmically. Thus, it would be an inefficient strategy for subjects to base responding on flash sequence durations, as the duration of a five-flash sequence occasionally lasted the same amount of time as a ten-flash sequence.

Sequence number and duration were, nonetheless, highly correlated. Our subjects could have timed the duration of the flash sequences to estimate the number of flashes in

them and they could have attempted to produce a given number of presses by timing the duration of their tapping. Although we think this unlikely, it is not material to our purpose, so we did not pursue it further. The current literature indicates that the nonverbal representations of time and of number have identical psychophysical properties (Meck & Church, 1983; Meck, Church, & Gibbon, 1985; Roberts, 1995). Our purpose was to use noise analyses to determine the locus of the variability in tasks in which behavior depends on the nonverbal representation and combination of abstract quantities like number and duration, and to constrain models of the relation between the objective and subjective magnitudes of these variables. If our subjects did use time as their basis for responding, as opposed to number, then the conclusions we draw apply to the representation of duration by means of mental magnitudes.

CONCLUSIONS

Our results imply that the primary source of variability in the behavioral consequences of mental arithmetic with nonverbal analog representations of number is variability (noise or uncertainty) in the subjective magnitudes entering into the arithmetic composition—not in the comparison process. However, some of the variability depends on the magnitude of the resultant. This is not surprising, given that subjects indicated the magnitude of the resultant by making a proportionate number of buttonpresses. The departures from additive operand variability in our arithmetic data can be explained by a combination of two things: (1) Additional noise was accumulated (scalar with respect to answer magnitude) once the answers were computed, in the process of responding, and (2) there was some covariance in the magnitudes representing the operands. The covariance increased variability in the addition condition and decreased variability it in the subtraction condition.

In addition, our study is the first to examine the influences of nonscalar noise sources on nonverbal counting data. Although, as previously assumed, scalar variability appears to be the most prominent characteristic of the data, our results suggest that the contribution of these binomial and additive sources are robust and should not be ignored (as evidenced in Figures 5, 7, and 8).

Importantly, the variability in the results of mental subtraction remains roughly constant as the difference passes through 0 and its sign reverses. This presents a challenge to models in which the mental magnitudes representing number are assumed to be proportional to the logarithm of the number they represent, which assumption requires that addition and subtraction be implemented by table look-up. The problem arises from the fact that the logarithms of 0 and negative quantities do not exist.

Our approach provides a general strategy for the behavioral analysis of the nonverbal mental arithmetic that mediates the extraction of descriptive statistics from quantitative experience. Analyses of the variability in responding sheds light on basic combinatorial mental capacities—a feat that mean response data cannot begin to tackle. In the present study, these variability analyses have illuminated not only

the quality of nonverbal magnitude representations but also on the processes involved in nonverbal arithmetic.

REFERENCES

ARIELY, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, **12**, 157-162.

ASLIN, R., SAFFRAN, J., & NEWPORT, E. (1999). Statistical learning in linguistic and nonlinguistic domains. In B. MacWhinney (Ed.), *The emergence of language* (pp. 359-380). Mahwah, NJ: Erlbaum.

BARTH, H., LA MONT, K., LIPTON, J., DEHAENE, S., KANWISHER, N., & SPELKE, E. (2006). Non-symbolic arithmetic in adults and young children. *Cognition*, **98**, 199-222.

BARTH, H., LA MONT, K., LIPTON, J., & SPELKE, E. S. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences*, **102**, 14116-14121.

BRANNON, E. M., & ROITMAN, J. D. (2003). Nonverbal representations of time and number in animals and human infants. In W. H. Meck (Ed.), *Functional and neural mechanisms of interval timing* (pp. 143-182). Boca Raton, FL: CRC Press.

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**, 31-49.

BRANNON, E. M., & TERRACE, H. S. (2002). The evolution and ontogeny of ordinal numerical ability. In M. Bekoff, C. Allen, & G. M. Burghardt (Eds.), *The cognitive animal: Empirical and theoretical perspectives on animal cognition* (pp. 197-204). Cambridge, MA: MIT Press.

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**, 238-243.

BUCKLEY, P. B., & GILMAN, C. B. (1974). Comparison of digits and dot patterns. *Journal of Experimental Psychology*, **103**, 1131-1136.

CHURCH, R. M., MECK, W. H., & GIBBON, J. (1994). Application of scalar timing theory to individual trials. *Journal of Experimental Psychology: Animal Behavior Processes*, **20**, 135-155.

CORDES, S., GELMAN, R., GALLISTEL, C. R., & WHALEN, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both small and large numbers. *Psychonomic Bulletin & Review*, **8**, 698-707.

CORDES, S., KING, A. P., & GALLISTEL, C. R. (2007). Time left in the mouse. *Behavioural Processes*, **74**, 142-151.

DEHAENE, S. (1997). *The number sense.* Oxford: Oxford University Press.

DEHAENE, S. (2001). Subtracting pigeons: Logarithmic or linear? *Psychological Science*, **12**, 244-246.

DEHAENE, S. (2002). Single neuron arithmetic. *Science*, **297**, 1652-1653.

DEHAENE, S., DUPOUX, E., & MEHLER, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception & Performance*, **16**, 626-641.

DEHAENE, S., SPELKE, E., PINEL, P., STANESCU, R., & TSIVKIN, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, **284**, 970-974.

FETTERMAN, J., & KILLEEN, P. (1995). Categorical scaling of time: Implications for clock-counter models. *Journal of Experimental Psychology: Animal Behavior Processes*, **21**, 43-63.

GALLISTEL, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

GALLISTEL, C. R., BRANNON, E. M., GIBBON, J., & WUSTHOFF, C. J. (2001). "Subtracting pigeons: Logarithmic or linear?": Reply. *Psychological Science*, **12**, 247.

GALLISTEL, C. R., & GELMAN, R. (1992). Preverbal and verbal counting and computation. *Cognition*, **44**, 43-74.

GALLISTEL, C. R., & GELMAN, R. (2005). Mathematical cognition. In K. Holyoak & R. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 559-588). New York: Cambridge University Press.

GALLISTEL, C. R., GELMAN, R., & CORDES, S. (2006). The cultural and evolutionary history of the real numbers. In S. C. Levinson & J. Pierre (Eds.), *Evolution and culture: A Fyssen Foundation symposium* (pp. 247-274). Cambridge, MA: MIT Press.

GIBBON, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, **84**, 279-335.

GIBBON, J., & CHURCH, R. M. (1981). Time left: Linear versus logarithmic subjective time. *Journal of the Experimental Analysis of Behavior*, **7**, 87-107.

GIBBON, J., & CHURCH, R. M. (1992). Comparison of variance and covariance patterns in parallel and serial theories of timing. *Journal of the Experimental Analysis of Behavior*, **57**, 393-406.

GILDEN, D., THORNTON, T., & MALLON, M. (1995). 1/f noise in human cognition. *Science*, **267**, 1837-1839.

KILLEEN, P., & WEISS, N. (1987). Optimal timing and the Weber function. *Psychological Review*, **94**, 455-468.

LOGIE, R., & BADDELEY, A. (1987). Cognitive processes in counting. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 310-326.

MCCRINK, K., & WYNN, K. (2004). Large number addition and subtraction by 9-month-old infants. *Psychological Science*, **15**, 776-781.

MECK, W. H., & CHURCH, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, **9**, 320-334.

MECK, W. H., CHURCH, R. M., & GIBBON, J. (1985). Temporal integration in duration and number discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, **11**, 591-597.

MOOD, A. M., & GRAYBILL, F. A. (1963). *Introduction to the theory of statistics* (pp. 297-301). New York: McGraw-Hill.

MOYER, R. S., & LANDAUER, T. K. (1967). Time required for judgments of numerical inequality. *Nature*, **215**, 1519-1520.

MOYER, R. S., & LANDAUER, T. K. (1973). Determinants of reaction time for digit inequality judgments. *Bulletin of the Psychonomic Society*, **1**, 167-168.

PARKMAN, J. M. (1971). Temporal aspects of digit and letter inequality judgments. *Journal of Experimental Psychology*, **91**, 191-205.

PICA, P., LEMER, C., IZARD, V., & DEHAENE, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, **306**, 499-503.

RATCLIFF, R., & SMITH, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, **111**, 333-367.

ROBERTS, W. A. (1995). Simultaneous numerical and temporal processing in the pigeon. *Current Directions in Psychological Science*, **4**, 47-51.

SPELKE, E. S., & TSIVKIN, S. (2001). Language and number: A bilingual training study. *Cognition*, **78**, 45-88.

STEVENS, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: Wiley.

TROMMERSHÄUSER, J., MALONEY, L. T., & LANDY, M. S. (2003). Statistical decision theory and trade-offs in the control of motor response. *Spatial Vision*, **16**, 255-275.

VERGUTS, T., & FIAS, W. (2004). Representation of number in animals and humans: A neural model. *Journal of Cognitive Neuroscience*, **16**, 1493-1504.

WHALEN, J., GALLISTEL, C. R., & GELMAN, R. (1999). Non-verbal counting in humans: The psychophysics of number representation. *Psychological Science*, **10**, 130-137.

XU, F., & SPELKE, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, **74**, B1-B11.

ZORZI, M., STOIANOV, I., & UMILTÀ, C. (2005). Computational modeling of numerical cognition. In J. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 67-83). New York: Psychology Press.

NOTES

1. One subject (Subject 11) was asked to participate in the baseline condition a second time. Therefore, this subject participated in a total of six sessions.

2. This was initially not the case for Subject 11's baseline data. It was statistically confirmed that this subject's data were significantly different from the other subjects' baseline data. Given that the anomalous nature of this data set was attributed to task variables, as opposed to uncharacteristic magnitude representations, this subject participated in the baseline condition a second time. All results reported here are from Subject 11's second run in the baseline condition. (See Appendix A for the initial baseline data obtained from this subject.)

3. These weights were chosen because, under purely scalar conditions, the variance in the arithmetic data was expected to be a function of the magnitude of the operands.

4. It was assumed that the binomial coefficients obtained in the parameter searches of the baseline data described the binomial variability due to error in the counting of the flashes and in the counting of the button presses. Thus, this coefficient was double the variability of a single count. So, for the computation of the baseline predictions, the binomial term (B'' in Equation 6) was divided in half.

5. Toward the end of some of the longer responses, there did appear to be a tendency either to slow down the rate of responding (e.g., S11 addition) or to produce buttonpresses at a more variable rate (e.g., S10 addition). Both of these trends were most likely an artifact of fatigue and do not implicate verbal counting.

APPENDIX A
Subject 11's Initial Baseline Data

The initial baseline data from Subject 11 appeared somewhat curvilinear, as though the subject became fatigued when pressing the button more than around 30 times, or anchoring effects influenced responding. Due to the asymptotic behavior of this data set, we were unable to obtain a satisfactory linear fit of the data using both weighted and unweighted linear regressions, as evidenced in Figure A1 (panel A).

Plots of the z scores of this data as a function of the number of flashes (Figure A1, panel B) revealed an inordinate amount of variability throughout (as compared with the other subjects), and especially when presented large flash sequences—both reflecting an imperfect fit of the linear regression. Further analyses using a series of two sample Kolmogorov–Smirnov tests of the z scores revealed that the baseline data for this subject reflected significantly more variability than baseline data for the other 3 subjects ($p < .001$).

Due to the anomalous nature of Subject 11's initial baseline data, this subject participated a second time in the baseline condition. Data obtained from this second run were not significantly different from the other subjects' baseline data, and thus were considered a more accurate measure of this subject's internal representation of number. The data reported in the body of this article are from Subject 11's second baseline condition.
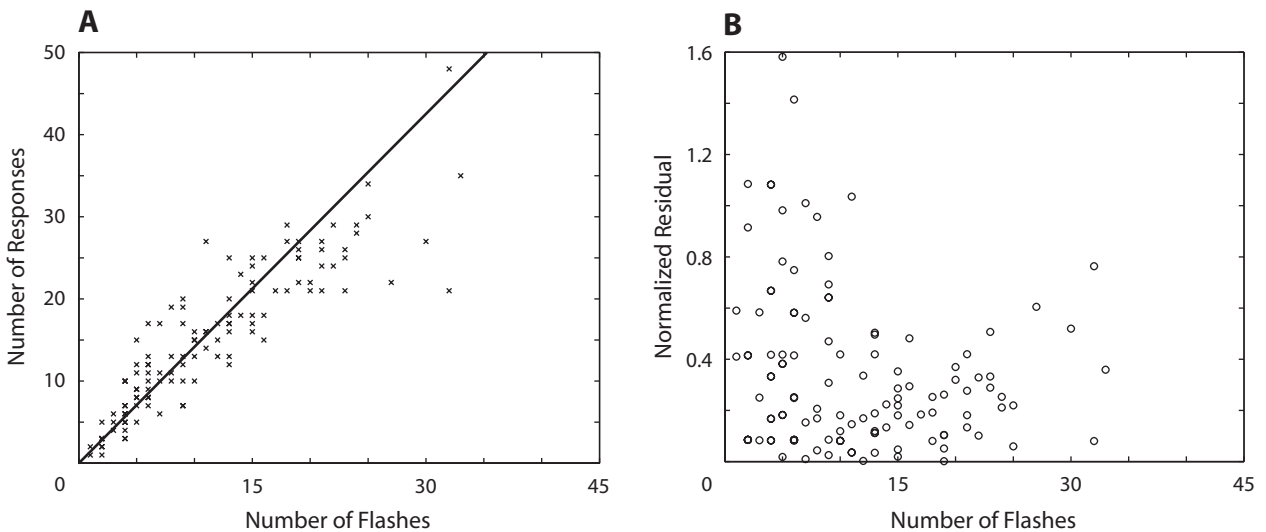

Figure A1. (A) Subject 11's original baseline data and regression line plotted as a function of the number of flashes presented. Note the curvilinear nature of the data toward the larger flash range. (B) The z scores from Subject 11's original baseline data plotted as a function of the number of flashes presented. Compared with the baseline z scores from other subjects (Figure 2), these data are notably more variable, especially in the larger flash range.

APPENDIX B

Model 1: Baseline Condition

To find the parameters of the weighted regression, we assume that the mean of the dependent variable (the number of counts) is a linear function of the true number of counts, and that the noise around that mean is Gaussian. Formally we write this as

$$z_i = ax_i + b + \eta_i, \tag{B1}$$

where z_i is the number of counts for trial i, x_i is the true number of counts, and η_i is a zero-mean Gaussian random variable. In this section, we consider the baseline condition, for which the variance of the noise is given by Equation 4 of the main text,

$$\mathrm{Var}(\eta_i) = Ax_i^2 + Bx_i + C. \tag{B2}$$

What Equations B1 and B2 tell us is that the probability of observing a particular z_i given x_i *and* the parameters of the model is

$$P(z_i \mid x_i, a, b, A, B, C) = \frac{\exp\left[-(1/2)(z_i - ax_i - b)^2 \big/ \left(Ax_i^2 + Bx_i + C \right) \right]}{\left[2\pi \left(Ax_i^2 + Bx_i + C \right) \right]^{1/2}}.$$

Since the samples are drawn independently, the probability of observing a whole set of zs given a set of xs is

$$P(\mathbf{z} \mid \mathbf{x}, a, b, A, B, C) = \prod_{i=1}^{N} \frac{\exp\left[-(1/2)(z_i - ax_i - b)^2 \big/ \left(Ax_i^2 + Bx_i + C \right) \right]}{\left[2\pi \left(Ax_i^2 + Bx_i + C \right) \right]^{1/2}}, \tag{B3}$$

where N is the number of observed (x_i, z_i) pairs, $\mathbf{x} \equiv (x_1, x_2, \ldots, x_N)$ denotes the set of xs, and $\mathbf{z} \equiv (z_1, z_2, \ldots, z_N)$ denotes the set of zs.

We now come to the key step, which is the use of Bayes's theorem to turn Equation B3 into a probability distribution over the parameters. Applying Bayes's theorem and performing a few lines of algebra, we find that the probability distribution of the parameters given $\mathbf{x}$ and $\mathbf{z}$ (that is, given all the observations) denoted $P(a, b, A, B, C \mid \mathbf{x}, \mathbf{z})$ is given by

$$P(a, b, A, B, C \mid \mathbf{x}, \mathbf{z}) \propto P(\mathbf{z} \mid \mathbf{x}, a, b, A, B, C) P(a, b, A, B, C), \tag{B4}$$

where $P(a, b, A, B, C)$ is the prior probability distribution over the parameters and we have ignored the constant of proportionality, as it is independent of the parameters.

Assuming uniform priors (except that A, B, and C must be nonnegative) and using Equation B3 for the first term on the right-hand side of Equation B4, we have

$$P(a, b, A, B, C \mid \mathbf{x}, \mathbf{z}) \propto \prod_{i=1}^{N} \frac{\exp\left[-(1/2)(z_i - ax_i - b)^2 \big/ \left(Ax_i^2 + Bx_i + C \right) \right]}{\left[2\pi \left(Ax_i^2 + Bx_i + C \right) \right]^{1/2}}.$$

It is convenient to write this as

$$P(a, b, A, B, C \mid \mathbf{x}, \mathbf{z}) \propto e^{-L(a, b, A, B, C)}, \tag{B5}$$

where $L(a, b, A, B, C)$, the loss function, is given by

$$L(a, b, A, B, C) = \frac{1}{2} \sum_{i=1}^{N} \frac{(z_i - ax_i - b)^2}{Ax_i^2 + Bx_i + C} + \log\left(Ax_i^2 + Bx_i + C \right). \tag{B6}$$

The first term in the loss function is the familiar one from linear least-squares regression. The second one penalizes large variance (without it, the loss function could be made arbitrarily small by making the variance arbitrarily large).

Given Equations B5 and B6, there are two natural ways to proceed. The more intuitive is to choose parameters that maximize the loss function; this is the maximum-likelihood approach (so called because we used uniform priors, so e^{-L} is the likelihood of the data). This is the approach we took here. An alternative is to compute the mean value of the parameters. In the limit of a large amount of data, the two approaches yield the same answer.

Models 2 and 3: Addition and Subtraction

Models 2 and 3 differ from Model 1 in two ways: There is a slightly more complicated dependence on the true counts (since there are two true counts instead of one) and the variance is different. However, the formalism is very similar. In particular, the regression equations for the two models are written

$$z_i^{\pm} = a'\left(x_i \pm y_i \right) + b' + \eta_i' \tag{B7a}$$

and

$$z_i^{\pm} = a''\left(x_i \pm y_i \right) + b'' + \eta_i'', \tag{B7b}$$

APPENDIX B (Continued)

where Equations B7a and B7b refer to Models 2 and 3, respectively; z_i^+ is the dependent variable for addition; and z_i^- is the dependent variable for subtraction. The variances of η_i' and η_i'', denoted σ'^2 and σ''^2, respectively, are given by

$$\sigma'^2 = A'\left(x_i^2 + y_i^2\right) + B'\left(|x_i| + |y_i|\right) + 2C'$$

and

$$\sigma''^2 = A''\left(x_i^2 + y_i^2\right) + B''\left(|x_i| + |y_i|\right) + 3C'' + A''\left(x_i + y_i\right)^2 + B''|x_i + y_i|.$$

(See Equation 5 for variance of Model 2 and Equation 6 for variance of Model 3.)

The derivation of the loss function for Models 2 and 3 proceeds exactly as it did for Model 1: Write down an expression of $P(\mathbf{z}\,|\,\mathbf{x}, \mathbf{y}, \text{parameters})$, use Bayes's theorem to derive an expression for $P(\text{parameters}\,|\,\mathbf{x}, \mathbf{y}, \mathbf{z})$, then assume uniform priors. Carrying out these steps, we find that

$$L^{\pm'}(a', b', A', B', C') = \frac{1}{2}\sum_{i=1}^{N}\frac{\left[z_i - a'\left(x_i \pm y_i\right) - b'\right]^2}{\sigma_i'^2} + \log\sigma_i'^2$$

and

$$L^{\pm''}(a'', b'', A'', B'', C'') = \frac{1}{2}\sum_{i=1}^{N}\frac{\left[z_i - a''\left(x_i \pm y_i\right) - b''\right]^2}{\sigma_i''^2} + \log\sigma_i''^2.$$

As in Model 1, the probability distribution for the parameters is proportional by $e^{-L^{\pm'}}$ for Model 2 and $e^{-L^{\pm''}}$ for Model 3. The regression parameters were taken to be the most likely ones; that is, the ones that maximized $e^{-L^{\pm'}}$ and $e^{-L^{\pm''}}$.