www.sciencemag.org/cgi/content/full/1171599/DC1

Supporting Online Material for

Recruitment of an Area Involved in Eye Movements During Mental Arithmetic

André Knops,* Bertrand Thirion, Edward M. Hubbard, Vincent Michel, Stanislas Dehaene

*To whom correspondence should be addressed. E-mail: knops.andre@gmail.com

Published 7 May 2009 on *Science* Express
DOI: 10.1126/science.1171599

This PDF file includes:

Supporting Online Material

Materials and methods

Participants

The fMRI experiment was part of a general research program on functional neuroimaging of the human brain which was sponsored by the Atomic Energy Commission (Denis Le Bihan) and received formal approval by the relevant ethical committee (Comité de Protection des Personnes, Hopital de Bicêtre, France). 19 volunteers took part in the study after having given their written informed consent. All data analyses were based on data from 15 participants (7 female; mean age = 23.9 years; SD = 3.4 years, range 19 – 29 years). Four participants were excluded due to technical problems during acquisition, poor data quality or non-compliance with instructions.

Stimuli

Calculation Task

The arithmetic problems were created from a set of standardized problems with 50 as first operand and either a small (6) or a large numerosity (26) as the second operand. The operands were identical for addition and subtraction. Apart from the correct result, 8 deviant results were created for each arithmetic problem. These deviants were arranged as a geometric series (i.e. were linearly spaced on a logarithmic scale) and ranged from double the correct result to half of the correct result (technically, they were generated as round(c x $r^{i/4}$), where c is the correct result, r the maximal ratio between correct outcome and the lowest/highest deviant, and i an index ranging from -4 to +4). Because prior research indicated that subjects could achieve a greater precision in symbolic than in non-symbolic notation (*1*), in order to maintain an approximately similar level of difficulty the ratio r varied between notations. Deviants for Arabic problems were more closely centered on the correct outcome (r = 1.41) than were the deviants for non-symbolic problems (r = 2.5). To avoid a strategy of always selecting the response falling in the middle of the proposed range, only seven out of those nine possible results were presented. In 50% of the trials we presented the upper seven (high range), and thus the correct result was the third largest numerosity. In the other 50% of the trials the lower seven choices were shown (low range), and the correct result was therefore the fifth largest numerosity.

Since the experimental design was organized around a small number of arithmetic problems, it was important to prevent subjects from memorizing them in symbolic form. Thus, the problems and their proposed results were randomly "jittered", differently on each trial. First, the operands were jittered by a random value from 0 to ± 8 for the first operand and from 0 to ± 3 for the second operand. Second, all of the seven proposed results were jittered up or down by a random value (fixed for a given trial). This random value had a mean value of zero and was drawn from a flat distribution on a logarithmic scale, in the range ± half of the numerical interval between the correct result and the first deviant above or below it. Technically, this was achieved by drawing a random number j between -0.5 and 0.5, and defining the proposed results as round(c x $r^{(i+j)/4}$), where i again ranges from -4 to +4. For example, for the problem 53 + 6 = 59, with the closest deviants being 54 and 64, the closest proposed result could be jittered anywhere from 57 = round(59 x $1.41^{-.5/4}$) to 61 = round(59 x $1.41^{.5/4}$). We ensured that the correct outcome would never appear as a response alternative. All proposed results fell between 4 and 243.

All problems were presented both in symbolic (Arabic) notation and as non-symbolic dot patterns (Figure 1 shows an example of a symbolic trial). The notation of response alternatives in a given trial was always identical with the notation of the operands. Both notations were displayed in black within a colored circle which was presented on a black background. Each circle had a diameter of 6.1° visual angle. A single Arabic digit extended to a height of 1.1° and a width of 0.63° visual angle. Seven different colors were used for the results, while the operands' colors were identical. Color served as relevant feature only in the control task and was randomized with respect to all numerical features for the rest of the experiment. The operands were presented successively in the center of the screen. The results were presented at seven locations arranged around the screen center in an ellipsoid fashion.

To prevent the use of non-numerical cues, the sets of dots representing the non-symbolic numerosities were designed and generated using Matlab© software (www.mathworks.com) such that dot size changed, but total dot area in a given set was always fixed across stimuli. Thus, total occupied area could not serve as a cue for distinguishing between the different numerosities. As a result of this manipulation, average item size varied inversely with numerosity during the presentation of the operands (i.e., sets with smaller numerosities had larger dots). It is hard to see how such a covariation could be used to predict the outcome of an addition or a subtraction, and indeed prior research where this parameter was controlled has demonstrated that it does not affect behavioral performance in approximate arithmetic (*1, 2*). To avoid memorization effects due to repetition of a particular stimulus, on each trial the stimulus images were randomly chosen from a set of 10 precomputed images with the given numerosity.

Saccades Localizer

In order to select voxels that were systematically activated by saccades (regardless of direction and amplitude) subjects were instructed to perform one run of saccades in a block design. This saccades localizer was composed of 14 presentations of a white target cross (0.38º) to the left or right of central fixation which was indicated by a green or red target cross for 'move' or 'rest' trials (see below).

Saccades Task

In order to train the classifier, we used a slow event-related design in which subjects performed four saccades in succession either to the left or to the right. The saccades task was composed of one run of 240 trials with constant presentation of a white target cross (0.38º).

Procedure

Calculation Task

A total of 108 trials were presented in 4 runs. After each run participants were given the chance to rest. Each trial started with the presentation of a white fixation cross for 1000 ms, which was followed by an uppercase letter ('A' for addition, 'S' for subtraction or 'C' for color) for an average of 2100 ms (range: 1100 to 3100 ms) which indicated the subsequent operation to be performed. After the instruction letter the first operand was presented for 300 ms, followed by the instruction letter to ensure fixation for an average of 2100 ms (range: 1100 ms to 3100 ms). The second operand was subsequently presented for 300 ms and was likewise followed by the presentation of the instructional letter for an average of 5100 ms (range: 4100 to 6100 ms). After this delay period, during which we computed the reported contrasts in brain activation, the screen was emptied and seven proposed results appeared, one by one, every 300 ms, at one of seven possible locations and remained on screen until response but maximally for 4500 ms. The notation of the operands always was identical to that of the response alternatives in a given trial. The temporal and spatial order (and thus the numerical order of the response alternatives) in which the seven response alternatives appeared on screen was randomized for each trial. After the appearance of the last response alternative, the mouse pointer appeared in the center of the screen and the participants had to either indicate which numerosity was numerically closest to the actual result (calculation) or which response alternative was displayed against the same color as the operands (color). Responses were made by 'clicking' on the respective image with a MR-compatible joystick (Current Designs©, Philadelphia, USA). Speed was stressed over accuracy to maximize the use of approximation strategies and to avoid explicit calculation (Arabic numerals) or counting (dot patterns). During the experiment no feedback was provided to participants.

Total run duration was 8 minutes, 48 seconds (220 TRs). The paradigm was created and controlled using Python 2.4 software (http://python.org).

Saccades Localizer

Participants performed one run consisting of ten blocks of saccades, each followed by a baseline period in which identical visual stimulation was presented, but participants did not move their eyes. Each block of eye movements began and ended at fixation, and the change from the eye-movement task to the fixation task was signaled via a change in the color of the fixation cross (green = target following, red = fixation). Each block of saccades was composed of 14 presentations of a target cross (0.38º) which appeared approximately 5º (± up to 0.42º jitter in x and y) to the left or the right of fixation, or near fixation (with the same jitter) for, on average, 1000 ms (± 200ms jitter; five trials of 800 and 1200 ms, four of 1000 ms). Each block used a different, fixed order, and block order was randomized across participants. Total run duration was 4 minutes, 46 seconds (119 TRs).

Saccades Task

Subjects performed two runs of the saccades task. In each run participants were asked to constantly track with their eyes and fixate a white target cross (0.38º), which indicated both the fixation periods and the positions to which each saccade should be made. Each trial began with the subjects fixating the target cross at the center of the screen. The fixation cross would move either to the left or right, indicating the beginning of a "mini-block" of four consecutive leftward or rightward saccades. The direction of horizontal displacement was pseudo-randomized, with equal probability of leftward and rightward displacement overall. The total saccade distance for the mini-block was approximately 5° visual angle from the center position, divided into four roughly equal steps (a horizontal jitter of ± 0.42º visual angle was added on each step). In addition, a vertical jitter of ± 0.26° visual angle was added to confirm that subjects programmed precise saccades, rather than making automatic saccades. The duration of each of the steps in the mini-block was 500, 750, or 1000 ms, for a total duration of the mini-block of either 2, 3 or 4 seconds. At the end of this mini-block an inter-trial interval (ITI) of 6, 7 or 8 seconds began. To avoid confounding our saccades of interest with immediate return saccades towards the center of the screen, this ITI was divided into two phases of equal duration. The fixation cross remained at the position of the final saccade in the mini-block (extreme left or right) for the first half of the ITI (either 3, 3.5 or 4 seconds) before returning to fixation for the remainder of the ITI. This delay period of 6, 7, or 8 seconds served as the baseline for our analyses. Thus, on average, one trial lasted 10 seconds (8000 to 12000 ms). Total run duration was 6 minutes, 52 seconds (172 TRs).

Data acquisition and analysis

Functional images were acquired at Neurospin Center in Saclay, France on a 3T MR system (Siemens TrioTim Syngo) as T2*-weighted echo-planar image (EPI) volumes. Forty transverse slices covering the whole brain were obtained with a TR of 2.4 s (TE: 30 ms; flip angle, 81°; 3 x 3 x 3 mm voxels; no gap). For each participant an anatomical scan was obtained at the end of the session using a MPRAGE sequence with 160 slices covering the entire brain (TR = 2.3 s, TE = 3 ms, flip angle = 9°, voxel size: 1 x 1 x 1.1 mm, no gap).

Data were preprocessed using SPM5 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm5) implemented in Matlab© software. The first three images (7.2 seconds) in each series were discarded to allow for stable magnetization. Functional images were corrected for motion and slice-timing differences. Images were realigned to the first image in the series of the respective experiment and co-registered to the individual anatomies. For the reported random effects and classifier analyses the functional images were smoothed with a 6 mm^2 Gaussian kernel after normalization to the standard template of the Montreal Neurological Institute. Very similar results were obtained when classifier analyzes were based on unsmoothed data.

In the calculation task, for each participant and session a general linear model was created which included 3 regressors at the onset of the instructional letter ('A', 'S', or 'C'), 6 regressors at the onset of the first operand (1 for each combination of the three operations and the two notations), 12 regressors at the onset of the second operand (now crossing operation, notation and numerical size of the second operand), 1 regressor modelling the period between the onset of the response alternatives and the button press, and 6 individual motion parameters from preprocessing to capture remaining signal variations due to head motion. The canonical hemodynamic response function was used to model the BOLD response.

For the saccade localizer, each target onset was modeled with the canonical HRF to which time derivatives were added to capture variance in saccade latencies. Random effect analyses were then applied to the whole brain first-order contrast saccades > rest.

For the slow saccades task, each mini-block was modeled with the canonical HRF with time derivatives, beginning with the onset of the first saccade in the mini-block, and lasting for the duration of the entire mini-block. Random effect analyses were then applied to the whole brain first-order contrasts.

The activation level of all the voxels from the selected ROIs and for each trial was entered into a linear SVM classifier, using python bindings of the LIBSVM library. We first selected voxels that were active in our saccades localizer task, and thus no specific feature selection procedure was used, given that the size of input images (i.e. the number of voxels included in the ROIs) was typically less than 1000. The capacity parameter of the SVM classifier was set to 1.0 in all our experiments.

Cross-validation across trials was performed on a K-fold basis, where K was either 4 or 10 (this value had very little influence on the results). When the test data was different from the input data (e.g. testing the saccades classifier on brain images from the arithmetic runs), no cross-validation was used, given that the classification results are free from overfitting issues.

Results

Quantification of classification results

Overall classification accuracy provides a simpler summary measure of classification performance, but it could be a biased towards one or the other category (although in the present experiments, equal numbers of trials were always presented in each of the two contrasted categories, e.g. right versus left saccades). We therefore adopted a signal-detection theory approach to analyze the classification results. Calculating d-prime per participant provided a bias-free measure of classification results which we then tested against zero across participants. Standard signal detection paradigms address the detection of a given signal against a background of noise. This provides true positive (TP) cases when participants detect the presented signal and true negative (TN) cases when participants judge the absence of the signal when only noise is presented. In the present paradigm we defined the correct classification of both right saccades and/or addition trials as TP, and the correct classification of left saccades and/or subtraction as TN. Similarly, for the generalization from saccades to calculation we defined addition trials which were classified as right saccades as TP and subtraction trials which were classified as left saccades as TN. Adopting this definition, the accuracy rate of addition trials being classified as right saccades represents the sensitivity and the correct classification of subtraction trials as left saccades the specificity of the classifier.

Specificity of the observed classification and generalization results.

As a control, we probed classification (addition versus subtraction and left versus right saccades) and generalization (from saccades to calculation) in two control regions, the horizontal IPS (active during number processing) and the motor hand area M1. We found that addition could be distinguished from subtraction in both regions (M1: 76.7% ±2.1%, $t(14) = 12.69$, $p < .001$, mean d-prime: 1.52, ± 0.15, t(14) = 10.44, p <.001 ; hIPS: 73.7% ±1.4%, $t(14) = 17.3$, $p < .001$, mean d-prime: 1.28, ± 0.08, t(14) = 15.86, p <.001), while leftward vs. rightward saccades could be classified only from the parietal region (M1: 50.5% ±1.6%, $t(14) = .28$, $p = .77$, mean d-prime: 0.02, ± 0.08, t(14) = .3, p =.772; hIPS: 55.8% ±1.5%, $t(14) = 3.74$, $p = .002$, mean d-prime: 0.29, ± 0.08, t(14) = 3.73, p = .002). Crucially, however, neither of these regions yielded significant generalization from saccades to calculation (M1: 50.2% ±1.9%, $t(14) = 0.09$, $p = .93$, mean d-prime: 0.01, ± 0.10, t(14) = 0.16, p = 872; IPS: 48.5% ±1.6%, $t(14) = -0.94$, $p = .36$, mean d-prime: -0.09, ± 0.09, t(14) = -1.04, p = .316). Thus, the arithmetic classification in PSPL does not reflect a generic ability of the support vector machine to discover hidden information anywhere in the brain, but rather a specific form a numerical-spatial interaction in the PSPL.

Robustness of the results in PSPL

Generalization from saccades to calculation in PSPL region resulted in a positive d-prime in 12 out of 15 cases. To further consolidate our findings, we repeated the analyses with different values of the SVM regularization constant (C), which is known to have a substantial impact on the generalization of SVM classifiers. Crucially, all values of C from 10^{-1} to 10^{10} yielded identical results. Only with drastically small values of C, i.e. $C \leq 10^{-2}$ did generalization from saccades to calculation approach chance level (C = .1: mean accuracy: 55% ± 1.8%, t(14) = 2.77, p = .015, mean d-prime = 0.32, t(14) = 2.82, p = .014; C = .01: mean accuracy: 54.2% ± 1.8%, t(14) = 2.31, p = .036, mean d-prime = 0.25, t(14) = 1.88, p = .081; C = .001: mean accuracy: 52% ± 2.2%, t(14) = .94, p = .36, mean d-prime = 0.16, t(14) = 1.17, p = .26).

The machine learning literature proposes different algorithms, one of which is the support vector machine (SVM) used here as a primary tool. Unlike SVM classifiers, other algorithms like Relevance vector machines (RVM, (*3*)) and Sparse Regression Discriminant Analysis (SRDA, (*4*)) are known to be independent from parameter C. Therefore, we reanalyzed the most crucial result, i.e. the generalization from saccades to arithmetic in PSPL using these two alternative algorithms. In line with our predictions, after having been trained on left-right saccades, the classifiers again generalized to mental calculation, though with a reduced overall accuracy of 52.8% for RVM (± 1.6%, t(14) = 1.83, p = .089, mean d-prime = 0.23, t(14) = 2.12, p = .052,) and 54.7% for SRDA (± 2.1%, t(14) = 2.24, p = .042, mean d-prime = 0.29, t(14) = 2.3, p = .038). Albeit significant, the results were less robust and in the case of RVM algorithm reached significance only in a directed t-test (all reported results are two-sided).

Eye movement data

Since the interpretation of the current data set relies on the activity in posterior parietal areas that are activated by saccadic eye movements it is essential to demonstrate that the observed differences between arithmetic operations are not a mere result of differential eye movements during mental calculation. In particular, we will demonstrate that participants fixated well the center of the screen during the calculation period. To this end we calculated the median horizontal eye position during three periods of the experiment for ten participants for whom we had a good quality recording of the eye movements during scanning (for three participants no data were recorded; two data sets were not satisfying in terms of data quality due to failed calibration of the eye tracker). We analyzed the period after the onset of the instruction letter until the onset of the second operand (p1), the period after the onset of the second operand during which participants were engaging in mental calculation (p2), and, finally, the period of response selection during which eye gaze was free to move as participants actively searched the screen for the most appropriate solution (p3).

Technically, after applying a sliding median filter across 10 adjacent data samples and rejecting periods that were affected by eye blinks as well as other implausible values (e.g. negative values) from the sample we calculated the deviation from the median horizontal gaze position of each participant in each period of the experiment (p1, p2 and p3). In a next step we calculated the median horizontal deviation (and the respective standard deviation) separately for each operation (i.e. for addition and subtraction) and participant. If participants fixate the center of the screen throughout, the values should distribute symmetrically around zero, a bias to the left side of the screen would result in more negative, to the right side of the screen in more positive values.

Figure S1 of the supporting online material depicts the distribution of the median gaze positions across subjects and time. The upper row shows the distribution of the period after the onset of the first operand (over a period of 2900 ms), the middle row the horizontal gaze position during mental calculation period (i.e. after the onset of the second operand for a period of 4000 ms) and the lower row depicts the distribution during the response period (until the response but for maximally 4000 ms).

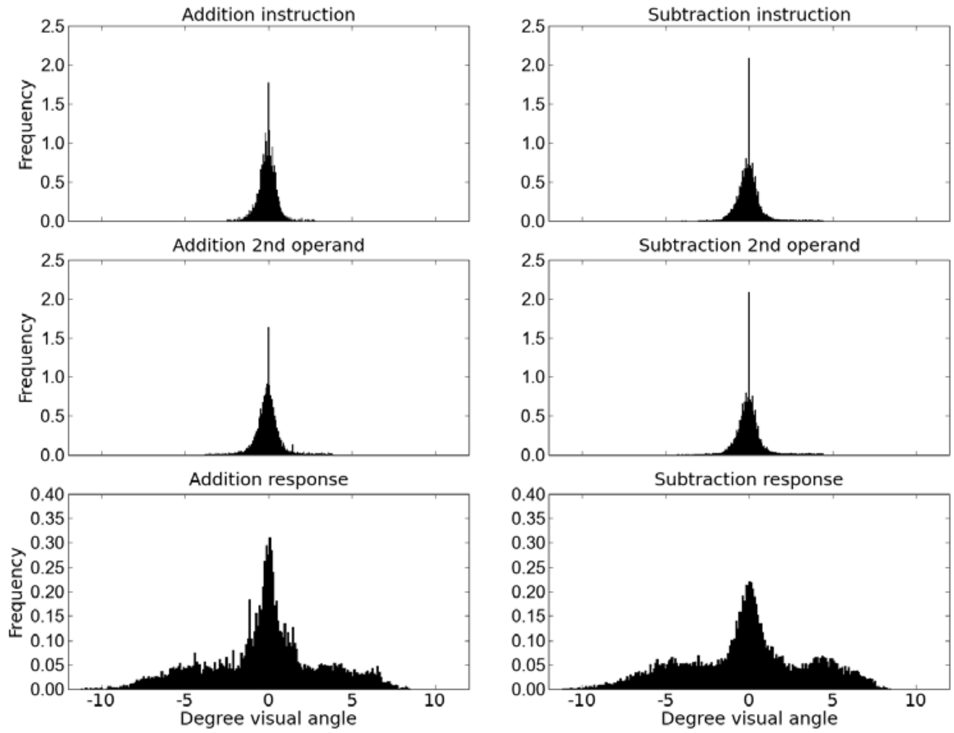

Figure S1: Histogram of median horizontal gaze position during three periods of the calculation experiment and separately for addition (left column) and subtraction (right column). Top row: Period of 2900 ms after the onset of the instruction letter. Middle row: calculation period after the onset of the second operand (note that during this period there was only the instructional letter on screen) Bottom row: histogram of median gaze position during a period of 4000ms that started with the onset of the first response alternative. During this period participants actively moved their eyes across the screen to choose a response alternative. Note difference in scaling between top/middle row and bottom row.

A repeated measures ANOVA over the variability revealed that variability increased when moving from p1 to p2 and to p3 ($F(2, 18) = 111.45$, $p < .001$). Paired tests revealed significant increases between subsequent periods of the paradigm ($p1 < p2 < p3$, all ps < .001). A significant main effect of operation implied a larger variation of subtraction as compared to addition ($F(1, 9) = 6.54$, $p = .031$). The interaction was not significant ($F < 1$).

Most crucially for the present purpose we verified whether any systematic differences in terms of gaze position between addition and subtraction might have contaminated the classifier analyzes. However, median gaze position did not change between the different periods ($F < 1$) nor did the arithmetic operation influence the gaze position ($F < 1$). No interaction between these factors was observed ($F < 1$).

Although median horizontal gaze position did not differ between addition and subtraction, the increase of variability of the gaze position across the different periods of the experiment in combination with the increased variability for subtraction as compared to addition merit a more fine grained analysis of the data. Therefore, instead of collapsing across all time points in a given period we calculated the median horizontal gaze position per participant and arithmetic operation for each time point of the period after the presentation of the second operand (p2). That is, analogous to an event-related potential we calculated the event-related gaze position, separately for addition and subtraction for each participant. In a second step we compared addition and subtraction for each time point by entering the data of all 15 participants into a t-test. This resulted in 240 t-tests and thus required a Bonferroni correction of the alpha-level. We applied a liberal correction that corresponded to calculating only 10 t-tests in a row. Figure 2 depicts the t-values for the comparison between addition and subtraction. Note that at no point in time did the mean of the median gaze positions across subjects significantly differ between the operations. If any, the differences pointed in the opposite direction, implying a slight shift of the mean gaze position to the left for addition as compared to subtraction.

In sum, the fact that 63% of the addition trials were classified as right saccades cannot be explained by brain activity accompanying systematic eye movements to the right side during this period. It should be noted that during the critical period, only the instructional letter was presented on screen, thus inducing excellent fixation. From the bottom part of figure S1 it becomes evident that systematic and frequent eye movements, during the choice period where the 7 proposed results were being scanned, result in a wide-spread distribution of horizontal eye positions that is dramatically different from the behavior observed for the periods after the instructional cue or the second operand. Despite slight differences in variability, the overall pattern of results implies that participants kept central fixation throughout the calculation period and that no systematic shift of median gaze position to the left for subtraction or to the right for addition was present that could contaminate the classification results.

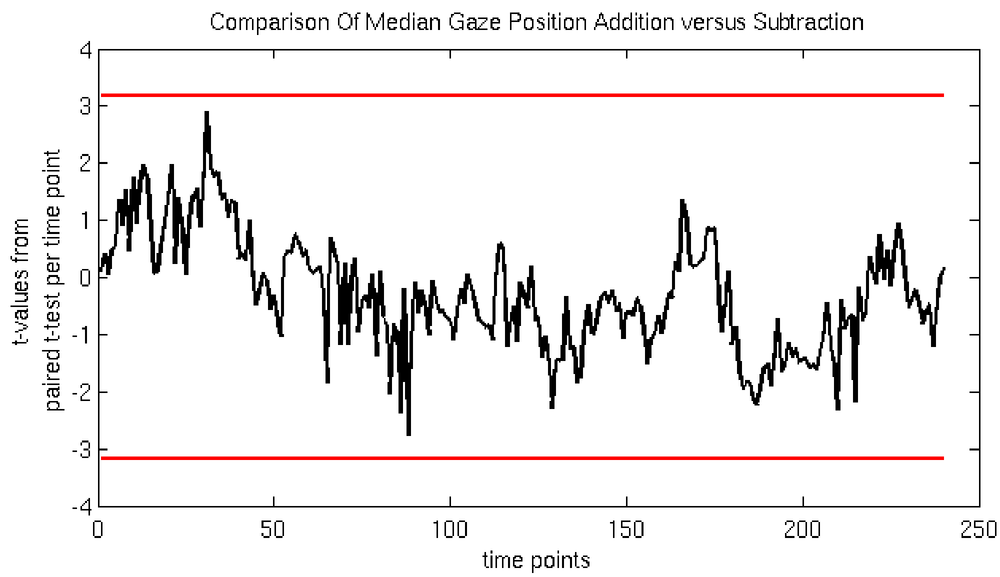

Figure S2: Evolution of the t-statistics for each time point for the comparison of horizontal gaze position between addition and subtraction. Negative t-values indicate a shift of the mean gaze position to the left. No point exceeded the significance threshold (red lines), which corresponds to a Bonferroni correction for a sequence of 10 t-tests in a row and can thus be regarded as liberal given the actual number of tests (i.e. 240).

Behavioral results

The central claim of the current study is that sensori-motor circuits in parietal cortex are recycled for high-level cognitive functions like mental arithmetic. Thus, it is crucial to demonstrate that participants indeed engaged in the assumed mental calculation process in both notations. We will demonstrate this by making use of the fact that we presented seven out of nine possible response alternatives, i.e. either the lower range or the higher range of the response alternative. This means that the closest result changes its relative position in the series of response alternatives. For the lower range the result closest to the correct outcome is in 5th position while it is in 3rd position for the higher range. If participants did engage in mental calculation, the histogram of the chosen response alternatives should be centered on these very values for the two ranges. Figure S3 depicts the distribution of the chosen response alternatives separately for each notation and operation.

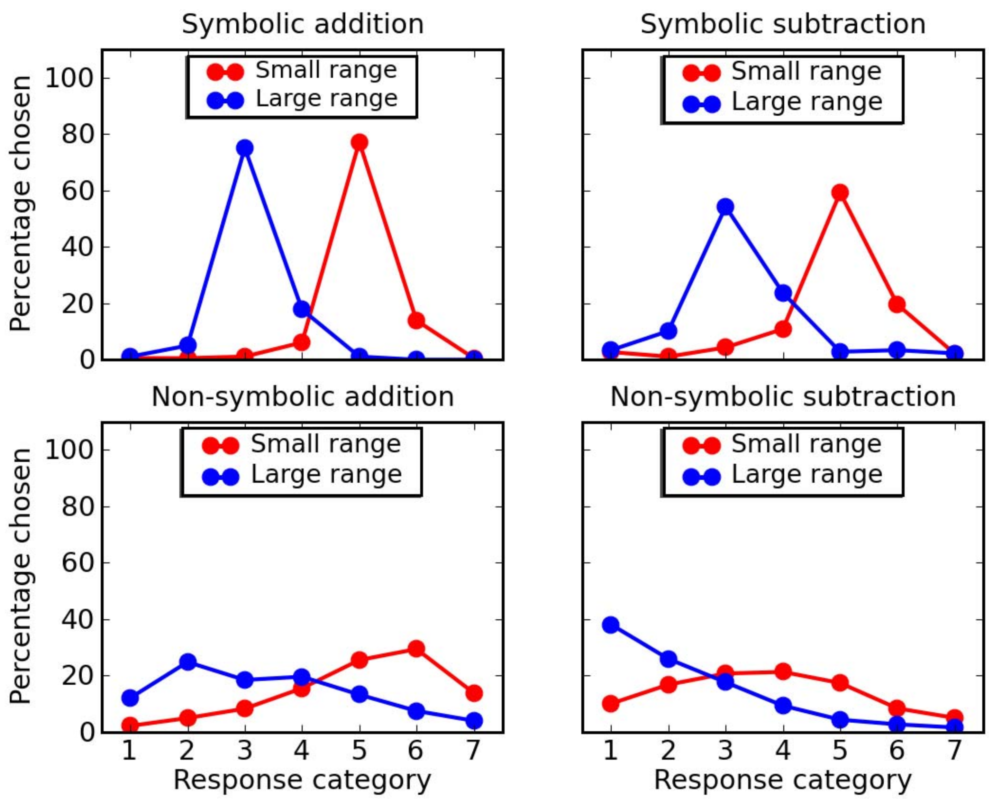

Figure S3 : Distribution of the participants' choices across the seven proposed results (averaged over all arithmetic problems, separately for each notation and operation). Participants' responses were not distributed randomly, but rather, depending on the range of response alternatives presented (high or low range), were centered on the value that was closest to the correct outcome (5th for low range and 3rd for high range). Additional influences of operation (smaller choices for subtraction than for addition) and of notation (underestimation bias for non-symbolic compared to symbolic notation) are also visible.

Figure S3 shows that the mode of the subjects' responses is affected by the proposed arithmetic problem and, for symbolic problems, peaks right at the value closest to the correct outcome. For non-symbolic calculation the distribution is more variable but still is clearly affected by the proposed arithmetic problem, with the mode falling at the values closest to the actual outcome. The distribution of choices is not flat, unlike what a random strategy would predict. Depending on the range of proposed choices, it is centered around the 3rd value when the 3rd value is the correct result and around the 5th value when the 5th value is the correct result. Since nothing in the proposed outcomes distinguishes these two types of trials, the latter finding implies that subjects take the operands into account. For subtraction, the response distribution is globally shifted towards smaller numerosities, a pattern of results replicating the results from a previous behavioral study (*1*). Most crucially, this distribution does not seem to arise from a random choice strategy, but can be explained by a combination of the overall tendency to underestimate the quantity of objects in a given set (*5*) and of the operational momentum effect (*1, 2*)

In sum, this pattern of results suggests that participants engaged in mental calculation rather than selecting any random value, both for symbolic and non-symbolic notation.

References for online supporting material

1. A. Knops, A. Viarouge, S. Dehaene, *Percept. Psychophys.* (in press).
2. K. McCrink, S. Dehaene, G. Dehaene-Lambertz, *Percept. Psychophys.* **69**, 1324 (2007).
3. M. E. Tipping, *Adv. Neural. Inf. Process. Syst.* **12**, 652 (2000).
4. D. Cai, X. He, H. J., *IEEE Trans. Knowl. Data Eng.* **20**, 1 (2008).
5. V. Izard, S. Dehaene, *Cognition* **106**, 1221 (2008).