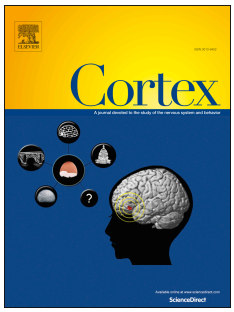
Accepted Manuscript



Decoding the processing stages of mental arithmetic with magnetoencephalography

Pedro Pinheiro-Chagas, Manuela Piazza, Stanislas Dehaene

PII: S0010-9452(18)30235-1

DOI: 10.1016/j.cortex.2018.07.018

Reference: CORTEX 2368

To appear in: *Cortex*

Received Date: 7 November 2017

Revised Date: 25 May 2018

Accepted Date: 16 July 2018

Please cite this article as: Pinheiro-Chagas P, Piazza M, Dehaene S, Decoding the processing stages of mental arithmetic with magnetoencephalography, *CORTEX* (2018), doi: 10.1016/j.cortex.2018.07.018.

Decoding the processing stages of mental arithmetic with magnetoencephalography

Pedro Pinheiro-Chagas[1] Manuela Piazza[2] and Stanislas Dehaene[1,3]

1.Cognitive Neuroimaging Unit, CEA DRF/I2BM, INSERM, Université Paris-Sud, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France
2. Center for Mind/Brain Sciences, University of Trento, 38068 Rovereto, Italy
3. Collège de France, 11 Place Marcelin Berthelot, 75005 Paris, France

Corresponding author:
Pedro Pinheiro-Chagas
ppinheirochagas@gmail.com
Cognitive Neuroimaging Unit
DRF/JOLIOT/NEUROSPIN/UNICOG
Bât. 145 - Point Courrier 156
F-91191 GIF SUR YVETTE Cedex FRANCE

Abstract

Elementary arithmetic is highly prevalent in our daily lives. However, despite decades of research, we are only beginning to understand how the brain solves simple calculations. Here, we applied machine learning techniques to magnetoencephalography (MEG) signals in an effort to decompose the successive processing stages and mental transformations underlying elementary arithmetic. Adults subjects verified single-digit addition and subtraction problems such as 3+2=9 in which each successive symbol was presented sequentially. MEG signals revealed a cascade of partially overlapping brain states. While the first operand could be transiently decoded above chance level, primarily based on its visual properties, the decoding of the second operand was more accurate and lasted longer. Representational similarity analyses suggested that this decoding rested on both visual and magnitude codes. We were also able to decode the operation type (additions vs. subtraction) during practically the entire trial after the presentation of the operation sign. At the decision stage, MEG indicated a fast and highly overlapping temporal dynamics for (1) identifying the proposed result, (2) judging whether it was correct or incorrect, and (3) pressing the response button. Surprisingly, however, the internally computed result could not be decoded. Our results provide a first comprehensive picture of the unfolding processing stages underlying arithmetic calculations at a single-trial level, and suggest that externally and internally generated neural codes may have different neural substrates.

Highlights

We used time-resolved MVPA to characterize the processing stages of mental arithmetic

Results revealed the unfolding of a cascade of partially overlapping brain states

Each brain state was highly dynamic and corresponded to different calculation features

We were able to decode the identity of the operands, operation type and correctness

Externally and internally generated codes seem to have different neural substrates

Keywords

mental arithmetic, magnetoencephalography, decoding, representational similarity analysis

1. Introduction

The ability to understand the world through mathematics is a uniquely human competence. Understanding how mathematics is implemented in the brain is therefore fundamental to our comprehension of the mechanisms of high-level symbolic cognition. Brain-imaging evidence suggests that even professional-level competence in mathematics is grounded in an evolutionary ancient set of areas that, in young children and non-human primates, is involved in simple number processing (Amalric & Dehaene, 2016, 2017). Elementary arithmetic thus appears as one of the fundamental building blocks of higher mathematics. Here, our goal is to begin to decompose the successive processing stages and mental transformations underlying elementary arithmetic, using the recently emerging technique of MEG decoding (Grootswagers, Wardle, & Carlson, 2016; King & Dehaene, 2014).

Traditionally, research in cognitive arithmetic has relied on behavioral methods and used mental chronometry to infer the covert processing stages of mental calculations. Behavioral research discovered that response time (RT) during calculation increases with the size of the operands, a finding which has been called the “problem-size” effect and which led to the proposal of several models of mental arithmetic (Ashcraft & Battaglia, 1978; Butterworth, Zorzi, Girelli, & Jonckheere, 2001; Campbell, 1994; Groen & Parkman, 1972; Uittenhove, Thevenot, & Barrouillet, 2016; Zbrodoff & Logan, 2005).

However, since Because RT is only a summary measure of the entire processing chain, it can only provide indirect information on the nature and relative timing of the various stages. Recently, more direct behavioral methods, such as continuous measures of finger pointing, have helped characterized the covert processing stages of arithmetic processing (Dotan & Dehaene, 2013, 2015; Pinheiro-Chagas, Dotan, Piazza, & Dehaene, 2017). Pinheiro-Chagas et al. (2017) monitored the finger trajectory of adult subjects, while they were asked to point to the result of single-digit calculations on a number line. Results revealed that additions and subtractions are computed by a stepwise displacement on the mental number line, starting with the larger operand (*max*), irrespectively of its position in the problem, and incrementally adding or subtracting the smaller operand (*min*). They also found a transient effect of the operator

sign (a plus sign attracted the finger to the right [larger results] and a minus sign to the left [smaller results]) around the time that subjects were processing the second operand. However, while such behavioral methods can be considerably informative about the duration and serial organization of cognitive computations, they remain limited in capturing processes that may happen simultaneously.

To supplement behavioral research, several studies have tried to decipher the neural code for numbers. Initial electrophysiology findings revealed the existence of single neurons tuned to specific numerosities in the monkey ventral intraparietal (VIP) and lateral prefrontal cortices (lPFC) (Nieder, 2016). These results were corroborated by human fMRI studies that demonstrated tuning curves for numbers in the intraparietal sulcus (IPS) (Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Piazza, Pinel, Le Bihan, & Dehaene, 2007) and a topographical organization of numerosites in the lateral parietal cortex (Harvey, Klein, Petridou, & Dumoulin, 2013). Machine learning was also used to successfully decode the identity of numbers from fMRI activity in parietal cortex (Eger, Pinel, Dehaene, & Kleinschmidt, 2015; Eger et al., 2009) However, these studies only investigated simple magnitude perception and comparison tasks. At present, due to the difficulty of training monkeys in arithmetic tasks, electrophysiological studies have not yet obtained direct information about the neural transformations underlying mental calculation, and fMRI measurements in humans are probably too slow to characterize them.

Only a few studies have tried to decompose the brain states during arithmetic processing, using a combination of mental chronometry and time-resolved brain imaging. Dehaene (1996) combined event relate potentials (ERPs) and the additive-factors method (Sternberg, 1969) to parse the processing stages involved in a number comparison (between two visually presented stimuli). By manipulating orthogonal features of the stimuli and the task, the author showed that the ERPs were first modulated by notation (Arabic numerals vs. number words, at ~110 - 170 ms), followed by the numerical distance (close vs. far, at ~190-300 ms) and finally by the lateralization of motor response (left vs. right, at ~250 – 400 ms). More recently, using a modified version of the arithmetic verification with ERPs, Avancini, Soltész, & Szucs (2015) identified a series of overlapping cognitive processes during calculation, such as the

identification of the stimuli properties, magnitude comparison and judgment of correctness.

Progress in understanding the spatial-temporal dynamics of mental calculations have recently increased with a series of novel electrocorticography (ECoG) findings. Using a sequentially presented addition task, a recent study revealed a series of successive brain activations: starting around ~90 ms, the number form area (NFA, lateral ventral temporal cortex) responds to digits, irrespective of whether or not they are presented in a calculation context (Shum et al., 2013); slightly later at ~100 ms, adjacent sites in the posterior inferior temporal gyrus (pITG) respond to numbers only when they are manipulated in the context of a calculation. Furthermore, activity at those ventral calculation-selective populations showed high correlations with activity in the vicinity of the intraparietal sulcus (IPS), which is traditionally considered the main number processing hub in the brain (Dehaene, Piazza, Pinel, & Cohen, 2003). Pinheiro-Chagas, Daitch, Parvizi, and Dehaene, (2017) further determined that both regions were affected by problem-size, though in different ways: pITG shows a fast peak which was inversely proportional to problem size, while IPS shows a more progressive activity whose integral is proportional to problem size. Thus, both regions seem to be involved in magnitude processing, but these findings do not resolve the nature of the underlying neural codes for the operands, nor do they provide a comprehensive picture of the series of unfolding computations.

In the present study, we aimed to evaluate whether magnetoencephalography (MEG) could resolve this issue. We combined MEG recordings with time-resolved multivariate pattern analysis (MVPA), specifically decoding (King & Dehaene, 2014) and representational similarity analysis (Kriegeskorte & Kievit, 2013), in order to characterize the series of processing stages and mental transformations underlying elementary arithmetic. Time-resolved decoding and MVPA have been successfully applied to characterize several cognitive functions such as working memory (King et al., 2016; Trübutschek et al., 2017; Wolff, Jochim, Akyürek, & Stokes, 2017) and object recognition and categorization (Carlson, Hogendoorn, Kanai, Mesik, & Turret, 2011; Carlson, Tovar, Alink, & Kriegeskorte, 2013; Cichy, Pantazis, & Oliva, 2014; Isik, Meyers, Leibo, & Poggio, 2014). Furthermore, MVPA can then shed light on the nature

of underlying codes (Diedrichsen & Kriegeskorte, 2017), exceeding the capacity of traditional ERP univariate-level analysis to capture fine-grained representations (Pantazis et al., 2017).

In the present task, subjects were asked to verify single-digit addition and subtraction problems, such as 3+2=5. Each of the symbols was presented sequentially for 400 ms, separated by 385 ms, so that we could analyze brain activity at each step. Specifically, we aimed at answering the following questions. First, can we decode the identity of operands? If so, can we distinguish neural codes for digit symbols and for the corresponding quantities? What is their temporal dynamics? Is this information sustained or transient? When and for how long can we decode the operation type? Can we then track the emergence of the internally computed result? Can we dissect the comparison and decision processes by which subjects classify the proposed result as correct or incorrect? Are these processes completely serial or do they partially overlap in a form of a cascade of computations that can be simultaneously decoded? Finally, are the neural codes independent of each other, or do they overlap? We were interested in the possibility that the neural codes for addition versus subtraction (active just after the presentation of the operation sign) would overlap with those for large versus small numbers, as such an overlap would readily explain the psychological observation that additions induce a bias towards larger numbers and subtraction towards smaller numbers (Knops, Viarouge, Dehaene, et al., 2009; Knops, Thirion, Hubbard, Michel, & Dehaene, 2009; McCrink, Dehaene, & Dehaene-Lambertz, 2007; Pinhas & Fischer, 2008; Pinheiro-Chagas, Dotan, et al., 2017).

2. Methods

2.1. Protocol and experimental design

Twenty healthy adults were scanned with MEG (23 ± 2 years old, 10 females, all right handed). Subjects had normal vision. The experiment lasted ~45 minutes, for which subjects were financially compensated. The study was approved by the local Ethics Committee and all subjects provided written informed consent before participation.

Subjects were asked to verify the accuracy of sequentially presented single-digit additions and subtractions problems in the form of A ± B = C (see Figure 1A). Each stimulus appeared for 400 ms, with an inter-stimuli interval of 385 ms. Subjects were instructed to generate an internal estimate of the result in advance of its appearance, and to further incite them, on half of the trials the appearance of C was delayed for an additional 385 ms. Inter-trial interval was 2,000 ms. Stimuli were white with a 1.5° visual angle, presented on a black background and projected on a screen with a refreshing rate of 60 Hz, placed 100 cm away from subject's head. The experiment was programmed in Python, mostly using the PsychoPy package (Peirce, 2007).

Subjects were asked to respond as fast and as accurate as possible if C was correct or incorrect, by pressing a button with their left of right thumb. In half of the blocks, left/right were associated with correct/incorrect and then switched. The association order was randomized across subjects. Stimuli were composed of the 16 addition and 16 subtraction problems consisting of all combinations of the operands: A = [3, 4, 5, 6] and B = [0, 1, 2, 3]. The correct results C ranged from 0 to 9, in the following proportions: 0 : 3.12 %, 1 : 6.25 %, 2 : 9.38 %, 3 : 15.62 %, 4 : 15.62 %, 5 : 15.62 %, 6 : 15.62 %, 7 : 9.38 %, 8 : 6.25 %, 9 : 3.12 %. On half the trials, C was correct. On the other half, C was ± [1, 2, 3, 4] distant from the correct result. A list of incorrect C's was generated for each subject with the single goal of maximizing the homogeneity of their distribution across trials.

Each experimental block took ~4.5 min and consisted of 32 calculation trials and 8 non-calculation trials, of the form "A = = = C", which are not analyzed in the current publication. Subjects completed 10 experimental blocks, comprising a total of 320 calculation trials.

2.2. Preprocessing

MEG signals were recorded with an ElektaNeuromag® MEG system (Helsinki, Finland), comprising 306 sensors (102 triples of 2 orthogonal planar gradiometers and 1 magnetometer) in a helmet-shaped array. Subjects' head position relative to the MEG sensors was estimated with four head position coils (HPI) placed on the frontal and pre-auricular areas, digitized with a 3-dimensional Fastrak system (Polhemus, USA), and

triangulated before each block of trials. Three pairs of electrodes recorded electrocardiograms (EMG) as well as the horizontal and vertical electro-oculograms (EOG). All signals were sampled at 1 kHz. MEG signals were hardware band-pass filtered between 0.1 Hz and 330 Hz, and active compensated for external noise with Maxshield (ElektaNeuromag). After visual inspection of bad channels, raw MEG signals were cleaned with the signal space separation method (Taulu & Simola, 2006) provided by MaxFilter (ElektaNeuromag) to 1. suppress magnetic interferences and 2. interpolate bad sensors. All further preprocessing steps were done with the Matlab Fieldtrip Toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011).

The MEG raw signals were epoched between -500 ms and +4,500 ms with respect to the onset of the first operand (A, see Figure 1A) and downsampled to 250 Hz. Trials contaminated by muscular or other artifact, were identified and rejected in a semi-automated procedure that used the variance across the MEG sensors. Next, we applied independent component analysis (ICA) to identify and remove artifacts caused by eye blinks and heartbeats. We then visually inspected the topographies of the first 30 components and subtracted the contaminated components from the data.

Further preprocessing was dependent on the nature of the analysis. For decoding and representational similarity analysis (RSA), epochs were low-pass filtered at 30 Hz and downsampled to 125 Hz. For time-frequency analysis, the spectral power of the non-low-pass-filtered epochs were estimated with parameters adapted to low and high frequency ranges. For the low-frequency range (2 – 34 Hz, steps: 1 Hz), data segments extracted from a sliding time window (length: 500 ms, steps: 40 ms) between 2 and 10 Hz and with a length of 5 oscillation cycles per frequency between 10 and 34 Hz was tapered with a single Hanning window. For the high-frequency range (34 – 100 Hz, steps: 2 Hz), data segments extracted from a sliding time window (length: 200 ms, steps: 40 ms) were multitapered and the frequency smoothing was set to 20% of each frequency value. Finally, epochs were cropped in three time windows: time-locked to A (-200 ms to 3,200 ms), time-locked to C (-200 ms to +800 ms) and time-locked to the response (-800 ms to +200 ms).

2.3. Decoding

All decoding analysis were performed using the Python scikit-learn (Pedregosa et al., 2011) and MNE (Gramfort et al., 2013) packages. The multivariate estimators aimed at predicting a vector of labels (y) from a matrix of features composed by single-trial MEG amplitude signals (X, shape = n_{trials} x ($n_{sensors}$ x $1_{time\ sample}$). Decoding analyses systematically consisted of the following steps: (1) fitting a linear estimator to a training subset of X (X_{train}); (2) predicting an estimate of y on a separate test set ($\hat{y}_{test}$); (3) assessing the decoding score of these predictions as compared to the true value of y. This procedure was repeated for each time sample separately. First, we used a standard transformation to z-scores in each channel at each time point across trials, in order to concomitantly include all 306 MEG sensors, pooling over magnetometers and gradiometers. Next, we fitted the data with a linear model to find the hyperplane that maximally predicts y from X while minimizing the loss function. Three main estimators were used: linear support vector machine (SVM) classifier, logistic regression classifier and Ridge regression, using the default parameters of the scikit-learn (e.g. lambda = 1). For multiclass problems using SVM, a 'one-versus-one' decision function was used. All decoding analyses were performed within subject and across trials, with an 8-fold stratified folding cross-validation scheme to maximize the homogeneity of distribution across training and testing sets. Decoding scores ($y,\hat{y}$) were quantified using the average classification accuracy for SVM and the Kendall's *tau* for Ridge regression. Statistical analyses were based on second-level tests across subjects. More specifically, we tested whether the classification scores were higher than theoretical chance value or 0, for classification accuracy and Kendall's *tau*, respectively, using one-sample t-test with random-effect Monte-Carlo cluster statistics for multiple comparison correction (Maris and Oostenveld, 2007), using the default parameters of the MNE *spatio_temporal_cluster_1samp_test* function.

As a note, we acknowledge that for the classification analysis, the most recommended approach is to use empirical chance level, estimated by running several iterations with labels shuffling. However, in the present case, this procedure would be prohibitively expensive in terms of computation time, given that in our time-resolved decoding approach, an independent classifier is trained at each time point. And our epochs have 4.8 s, that is, 600 time points when downsampled to 125 Hz. Crucially, we

have no reason to expect that empirical chance would greatly deviate from theoretical chance in our dataset, because all classes were fully balanced in all decoding analysis. This was done by randomly selecting a subsample of trials from the labels with more trials to match the one with fewer trials. We only needed to exclude a maximum of 4 trials per subject in each decoding analysis. Furthermore, we did not perform any baseline correction in the data, therefore the baseline period is meaningful. As can be seen in Figure 1, there was no single time point at which the decoding accuracy was significantly higher than theoretical chance level during the baseline period or during the period in which the information to be decoded was unavailable to the subjects. Therefore, we are confident that our results do not originate from false positives.

2.4. Temporal generalization

We also tested if each estimator fitted across trials at time *t* could accurately predict the $\hat{y}$ value at time *t'*, therefore probing whether the coding pattern is similar between *t* and *t'*. We applied this systematically across all pairs of time samples, resulting in a temporal generalization matrix (King & Dehaene, 2014).

2.5. Riemannian geometry

We also applied an estimator based on Riemannian geometry, using a covariance matrix estimation that integrates the temporal information, using the open source tools developed by Jean-Rémi King and Alexandre Barachant (https://github.com/kingjr/jr-tools, https://github.com/Team-BK/Biomag2016). More specifically, the model relies on the tangent space mapping of the covariance matrix described in (Barachant, Bonnet, Congedo, & Jutten, 2013). We started by decomposing the low-pass filtered data with a Principal Component Analysis (PCA) and taking the first 70 components for dimensionality reduction. Next, we used the ERPCov model (Barachant & Congedo, 2014), which is useful to capture both evoked and task induced responses, since it embeds the temporal information of the signal by concatenating, along the sensor axis, the averaged ERF (across trial) of each class before estimating the spatial covariance matrix. Finally, we mapped the covariance matrix to the tangent space and fitted a SVM or logistic regression classifier with our

standard cross-validation scheme. The use of Riemannian geometry has been shown to increase performances in sensorimotor rhythm (SMR)-based brain-computer interface (BCI) and more recently in MEG decoding of cognitive features (Biomag 2016 Decoding Competition).

2.6. Representational Similarity Analysis (RSA)

Several RSA models were constructed to test specific relationships between different dimensions of the stimuli and the MEG signals (Cichy et al., 2014; Diedrichsen & Kriegeskorte, 2017; Kriegeskorte & Kievit, 2013). RSA analyses systematically consisted of (1) averaging conditions across trials; (2) pair-wise correlating the conditions across the MEG sensors at each time point; (3) creating a symmetric dissimilarity matrix, equal to 1 - Spearman's rank correlation coefficient; (4) correlating the observed matrix with the theoretical similarity matrices predicted by different types of neural codes for the stimuli (see below). This procedure was repeated for each time sample separately. From the z-scored data, 32x32 representational dissimilarity matrices (RDM) were constructed using the 32 additions and subtraction problems, sorted by first operand, then by operation (additions first) and finally by second operand (3+0, 3+1, 3+2, 3+3, 3-0, 3-1, 3-2, 3-3, etc.). Seven theoretical RDM were constructed with the same structure and based on the magnitude dissimilarity (numerical distance) or visual dissimilarity (see method below) of operand 1, operand 2 and correct result and based on category for addition vs. subtractions (see Figure 4). Visual dissimilarity matrices were calculated using the Gabor Filterbank method, as implemented in the Matlab Image Similarity Toolbox. (https://github.com/daseibert/image_similarity_toolbox).

This method projects the image onto a Gabor wavelet pyramid as a model for primary visual cortex, simplified from (Kay, Naselaris, Prenger, & Gallant, 2008). The filters span eight orientations (multiples of .125π), four sizes (with the central edge covering 100%, 33%, 11%, and 3.7% of the image), and *X, Y* positions across the image (such that filters tile the space for each filter size). The resulting vector of filter responses are then compared between images, using the Euclidean distance. The method replicates the dissimilarity matrix of neural responses of the inferior temporal

cortex (IT) in both humans and monkeys (Kriegeskorte et al., 2008) (Readme File of the toolbox).

We then used Spearman's rank correlation test to evaluate the relationship between the observed and theoretical matrices. All RSA analyses were first computed within each subject, then statistical analyses were based on second-level tests across subjects, using the same method as in the decoding analysis, to test if the correlation coefficient was higher than 0.

3. Results

Twenty healthy adults were asked to verify the accuracy of successively presented single-digit additions and subtractions problems with matched operands in the form of A ± B = C, where in half of the trials C was incorrect (see Figure 1A and Methods). Accuracy was very high (average = 98.8%). Reaction time was faster for correct as compared to incorrect proposed results ($mean_{correct}$ = 519 ms, $SD_{correct}$ = 117 ms, $mean_{incorrect}$ = 622 ms, $SD_{incorrect}$ = 134 ms, $F(1, 19)$ = 68.796, p = 0.013; η^2 = 0.149). Within the trials with an incorrect result, no distance effect was found across the four absolute distances between the proposed and the correct results ($mean_1$ = 630 ms, SD_1 = 145 ms, $mean_2$ = 616 ms, SD_2 = 130 ms, $mean_3$ = 628 ms, SD_3 = 136 ms, $mean_4$ = 615 ms, SD_4 = 141 ms, $F(3, 57)$ = 0.781, p = 0.508; η^2 = 0.002). And no significant difference was observed when combining the trials in which parity was violated (distance 1 or 3) and those in which it was preserved (distance 2 or 4) ($t(19)$ = 0.437, p = 0.662, *Cohen's d* = 0.097). Finally, we also did not observe a problem-size effect, considering both operand 1 (*max*) (b = 4.919 ms; p = 0.69) and operand 2 (*min*) (b = -3.249 ms; p = 0.797). This is expected, since calculation was probably performed between the onset of operand 2 and the equal sign, therefore subjects most likely already had the correct result in mind when the proposed result was presented.

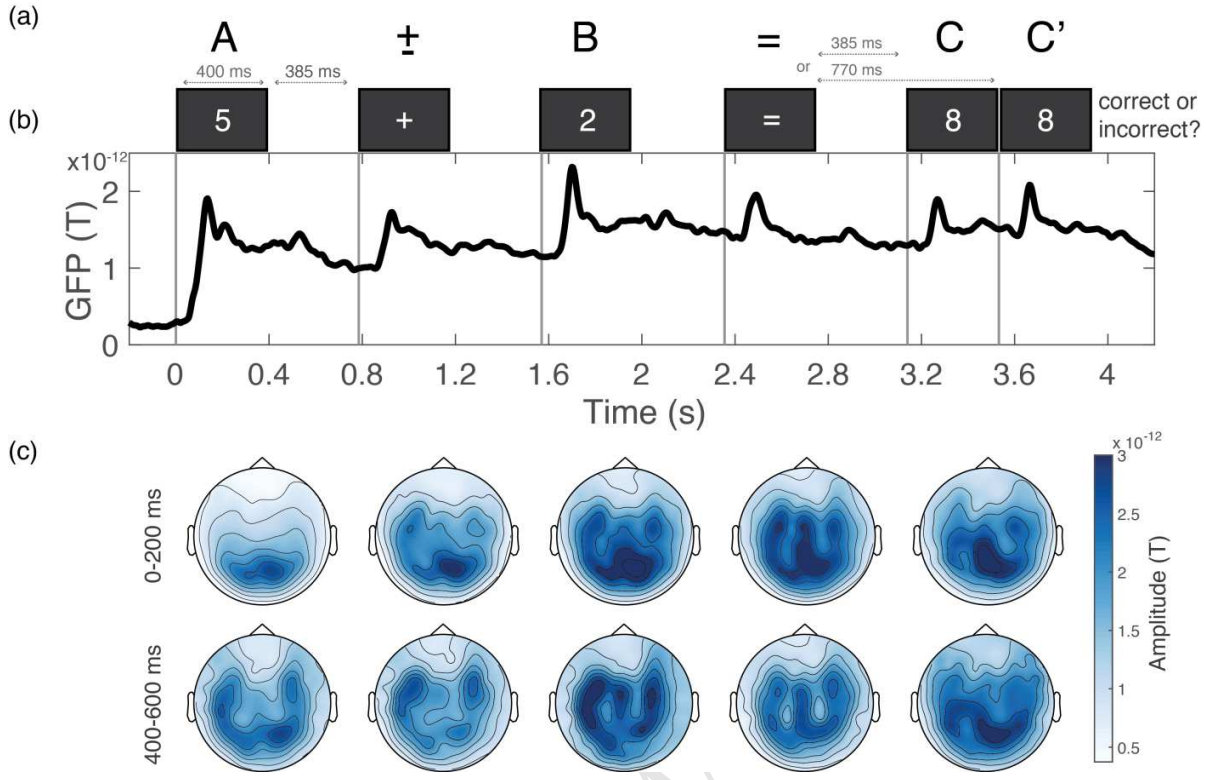


Figure 1. Sustained activity and signal propagation from posterior to anterior sensors

(A) Experimental design. Subjects were asked to verify the accuracy of sequentially presented single-digit additions and subtractions problems in the form of A ± B = C, with an 785 ms asynchrony. On half the trials, the presentation of C was delayed by an additional 385 ms. (B) Global Field Power (GFP), estimated using the MEG gradiometers and baseline corrected. After the onset of each stimulus event, GFP sharply peaked and remained above baseline for the entire trial. (C) Averaged MEG gradiometers topographies calculated between 0 – 200 ms and 400 – 600 ms after each stimulus. The signal propagates from posterior to anterior sensors after the onset of each stimulus and overall across the entire trial.

3.1. Sustained activity across the entire trial

In order to investigate weather overall activity was transient or sustained across the entire trial, we calculated the Global Field Power (GFP) (Lehmann & Skrandies, 1980), for the MEG gradiometers sensors and then normalized with the reference of a baseline period of -200 ms from the onset of operand 1.. As can be seen from Figure 1B, GFP increases right after (~100 ms) the presentation of each event and then slowly decreased until the presentation of the next event, but without returning to baseline, thus confirming that the overall activity was sustained across the entire trial. The evoked

brain activity evolved across time from more anterior sensors in the first 200 ms after the stimuli onset to more posterior sensors in the following period of 400 – 600 ms. Qualitative exploration showed that the second operand produced a higher and wider occipital-parietal-frontal activation as compared to the first operand, in both early and later time windows (Figure 1C). Therefore, this sustained activity allows us to investigate in more detail the mental transformations occurring during the entire trial.

3.2. Decoding the processing stages of mental arithmetic

We next investigated whether we could decode the series of processing stages underlying mental calculation, from the perception and representation of the operands to the operation type and response selection. For this purpose, we cropped the epochs in three different time windows: time-locked to operand 1 (-200 ms to 3,200 ms), time-locked to C (-200 ms to +800 ms) and time-locked to the RT (-800 ms to +200 ms). For each time window, we used seven different classifiers (SVM, see Methods) to decode operand 1 [values: 3, 4, 5, 6], operation [additions, subtractions], operand 2 [0, 1, 2, 3], correct result [3, 4, 5, 6; chance = 0.25], proposed result [3, 4, 5, 6], correctness of the operation as judged by the subject [correct or incorrect, including only the accurate responses], and response side [left vs. right button press, including only the accurate responses]. Note that for the correct and proposed result we only included the trials in which their values were [3, 4, 5, 6], since those were homogeneously distributed (15.62 % of trials each, see Methods).

3.3. Operand 1

The classification accuracy for operand 1 became significantly above chance starting at 112 ms after its onset, with a peak at 152 ms, and lasted until 640 ms ($p < 0.05$, corrected for multiple comparisons).

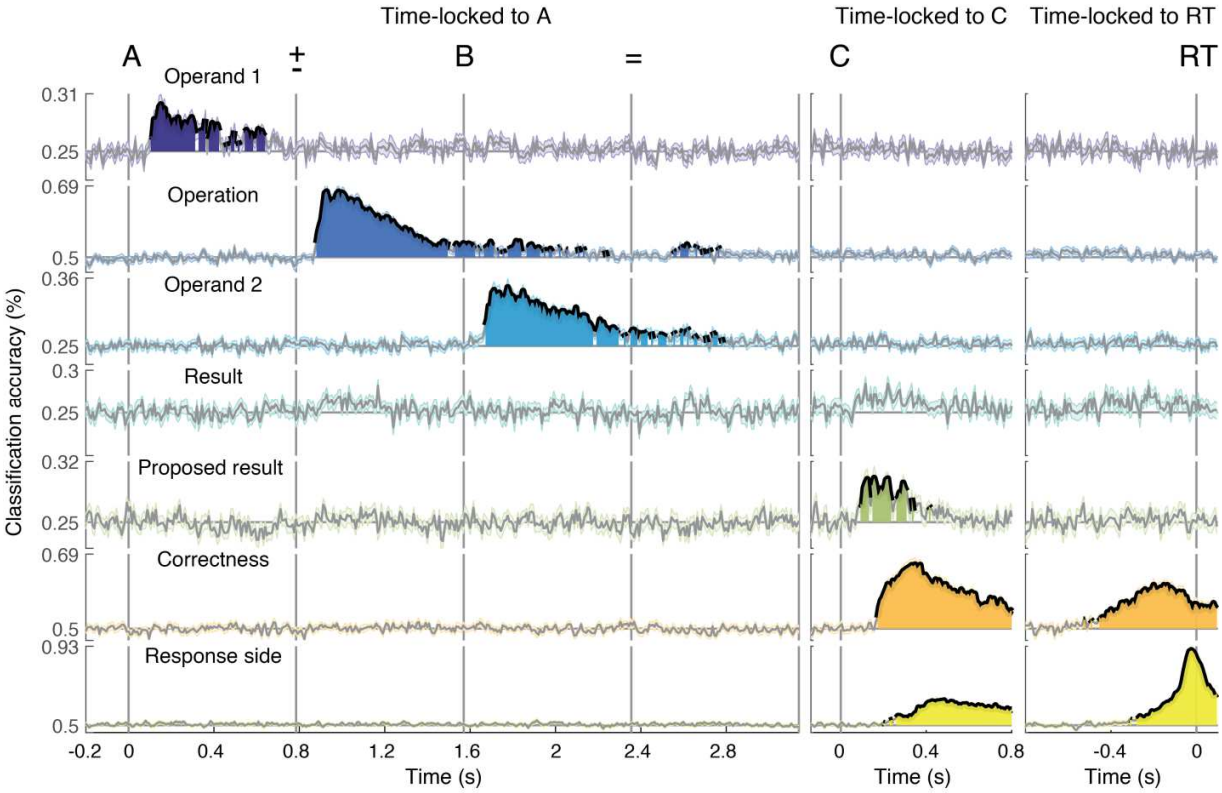


Figure 2. Decoding the time course of the processing stages underlying calculation

A series of SVM estimators were applied to classify the different features at each time sample, using the signal amplitude of all MEG sensors. Trials were time-locked to three windows of interest: after the onset of the operand 1 (A), proposed result (C) and RT. Gray horizontal lines indicate theoretical chance level. Operand 1, operand 2, result and proposed result involved 4 classes each (theoretical chance level = 25 %) and operation, correctness and response side involved binary classifiers (theoretical chance level = 50 %.). Thick lines and filled areas represent time periods in which the second-level statistical tests across subjects revealed a classification accuracy significantly above chance (cluster corrected for multiple comparisons, $p < 0.05$).

3.4. Operation type

The decoding scores for the operation became significantly higher than chance at 880 ms (i.e. 95 ms after the onset of the operation sign at 785 ms) with the peak at 928 ms, then dropping after the offset of sign, but remaining above chance almost all the way though the onset of the equal sign, and then transiently recovering above-chance performance level after the onset of the equal sign ($p < 0.05$, corrected).

3.5. Cross generalization from operation type to operand 2

The high initial classification score of the operation is most likely due to the visual difference between the plus and minus signs, but could also reflect task-specific preparation, such as operator priming (Fayol & Thevenot, 2012) as well as visual-spatial mechanisms or spatial-numerical associations (Hartmann, Mast, & Fischer, 2015; Masson & Pesenti, 2014; Mathieu, Epinat-duclos, Léone, Fayol, & Thevenot, 2017; Mathieu, Gourjon, Couderc, Thevenot, & Prado, 2016). Indeed, behavioral studies have shown that addition leads to a bias towards large numbers, and subtraction a bias towards small numbers (Knops, Viarouge, Dehaene, et al., 2009; Knops, Thirion, et al., 2009; McCrink et al., 2007; Pinhas & Fischer, 2008), which could suggest that the neural codes for add/subtract and for larger/smaller numbers overlap. To test this hypothesis, we trained a logistic regression classifier to decode subtractions vs. additions and tested if it could cross-generalize to small vs. larger numbers for both operand 1 (in which 3 & 4 received the same label as subtraction and 5 & 6 as addition) and operand 2 (in which 0 & 1 received the small label as subtraction and 2 & 3 as addition).

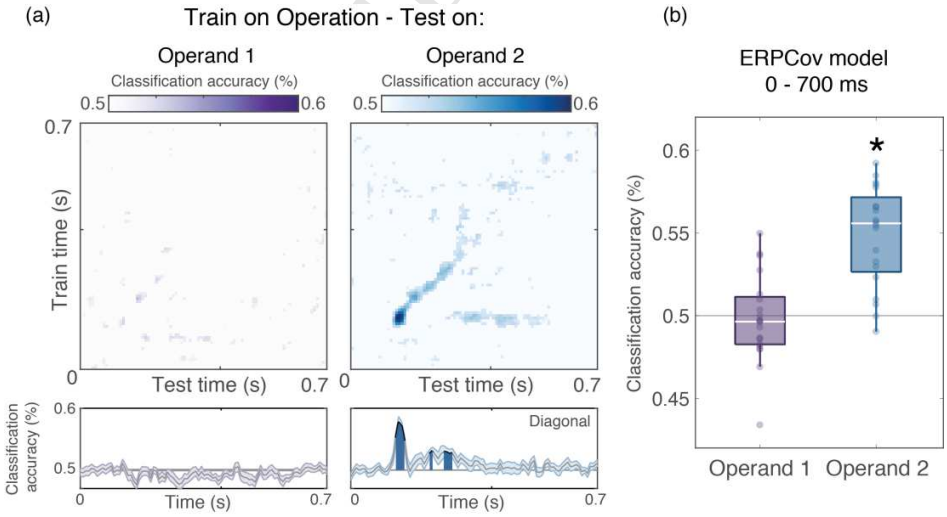


Figure 3. Cross-decoding from operation to operands
A logistic regression classifier was used to decode subtractions vs. additions and then tested if it could generalize to respectively smaller vs. larger numbers for both operand 1 (in which 3 & 4 received the same label as subtraction and 5 & 6 as addition) and operand 2 (in which 0 & 1 – subtraction and 2 & 3 addition). The time window used was

between 0 – 700 ms, locked to each stimulus. (A) Top squared plots show the generalization across time matrices, with only classification accuracies significantly above chance ($p < 0.05$, uncorrected). Bottom plots show the diagonal of the upper matrices, where train and test times were the same. Gray horizontal lines indicate theoretical chance level (0.5). Thick lines and filled areas represent time periods with classification accuracy significantly above chance (cluster corrected for multiple comparisons, $p < 0.05$). (B) Boxplots represent classification scores across subjects (individual dots) for the ERPCov model, which integrates the information over 0 – 700 ms (* = $p < 0.01$, second-level 1-sampled t-test).

As can be seen in Figure 3, cross-generalization from operation was only significant at the time of operand 2, but not for operand 1 (even when using a more robust Riemannian geometry based model which integrates the temporal information). Those results therefore suggest the existence of a transient (~128 – 288 ms) common code between additions and subtractions and larger and smaller operands 2, respectively.

3.6. Operand 2

The decoding scores for operand 2 started to be significantly above chance at 1,672 ms (i.e. 102 ms after its onset at 1,570 ms) with a peak at 1,776 ms, then dropping after the its offset, but remaining above chance until 2,770 ms ($p < 0.05$, corrected). Therefore, between the onset of operand2 and the offset of the equal sign, both the operation (addition vs. subtractions) and the operand 2 could be decoded simultaneously from the same MEG data. Importantly, comparisons showed that operand 2 was decoded with higher classification accuracy than operand 1 (between 0 – 400 ms: mean operand 2 = 0.31, SD = 0.029; mean operand 1 = 0.27, SD = 0.016, $F(1, 19) = 67.706$, $p < 0.001$; $\eta^2 = 0.414$ and between 400 - 800 ms: mean operand 2 = 0.284, SD = 0.018; mean operand 1 = 0.26, SD = 0.012, $F(1, 19) = 41.776$, $p < 0.001$; $\eta^2 = 0.382$), and for a longer time period (see also Figure S). This observation suggests that more intense brain activity occurred after operand 2 than after operand 1, in agreement with the fact that, at this time, subjects were able to start their calculation, a process whose length depends on the size of the *min* operand (or the smallest operand) (Groen & Parkman, 1972; Pinheiro-Chagas, Dotan, et al., 2017; Uittenhove et al., 2016), which in the present experiment is always operand 2. A potential confound that could

explain the higher decoding accuracy observed in operand 2 is the presence of 0, since it has been proposed that problems with 0 might engage a non-calculation rule-based strategy (Ashcraft & Battaglia, 1978), therefore facilitating their classification. We tested and refuted this possibility, by excluding the 0s. Even with a smaller data set, the classifier for operand 2 significantly outperformed the one for operand 1 (0 – 400 ms: $F(1,19) = 17.643$, $p < 0.001$, $\eta^2 = 0.146$ and 400 – 800 ms: $F(1,19) = 18.410$, $p < 0.001$, $\eta^2 = 0.188$).

3.7. Proposed result

As expected, the proposed result was not decodable before its appearance on screen. Similarly to operand 1, it was transiently decoded starting from 92 ms, with a peak at 166 ms and remained above chance only until around its offset ($p < 0.05$, corrected).

3.8. Correctness

The correctness of the trial judged by the subject was significantly classified above chance from 172 ms after the onset of the proposed result with a peak at 248 ms and remained significant all the way until the end of the epoch ($p < 0.05$, corrected). We did not observe any significant decoding score for the absolute distance between proposed and correct result (1 - 4), in line with the absence of a distance effect in RT. Relative to the onset of the proposed result, the response side started to be significantly classified above chance at 196 ms, with a peak at 484 ms, and this effect also lasted until the end of the epoch ($p < 0.05$, corrected). Note that the classifiers for response side and response correctness were orthogonal, since the response buttons were switched in the middle of the experiment (see Methods). A better way to look at the relationship between the judgment of the correctness and the response side, is to time-lock the epochs to the key press. This analysis clearly showed a slow ramping of the classification score for the correctness starting at -428 ms with a peak at -100 ms followed by a drop just before the response ($p < 0.05$, corrected). On the other hand, the fast ramping of the classification score for the response side started at -212 ms and

sharply increased to almost perfect classification at around 24 ms before the button press.

3.9. Generalization across time

To investigate the dynamics of calculation, we conducted a generalization across time decoding analyses (King & Dehaene, 2014), which revealed that the features of operand 1, operation sign, and operand 2 were decodable when train and test times were approximately the same ('*diagonal decoding*' Figure A.1). This analysis therefore suggests that each of these items launched a series of internal processes whose underlying codes dynamically changed along the trial. Nevertheless, the generalization-across-time matrix was broader for operation sign and for operand 2, transiently turning into a square pattern characteristic of sustained activity (Figure A.1). Furthermore, while the operand 1 and the proposed result were only transiently decoded during the time window that the stimuli was visually present, the operation, operand 2, correctness and response side had classification scores above chance that lasted for a longer time window.

3.10. Representational similarity

The decoding analysis does not directly reveal the precise stimulus dimensions that allowed the classifier to perform above chance level. In particular, we wanted to further investigate the representational geometries underlying the responses evoked by the operands and the result. For that, we turned to representational similarity analyses (RSA).

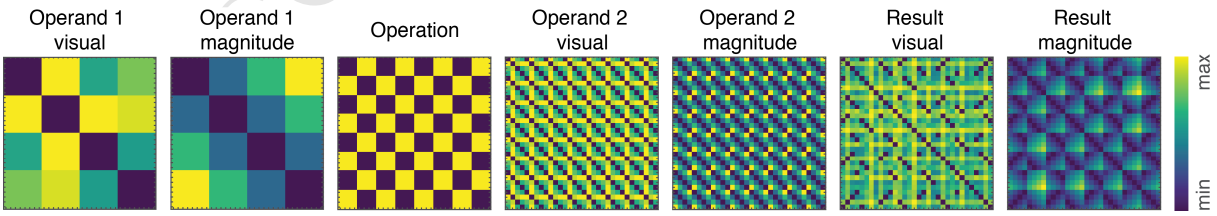


Figure 4. Theoretical predictors of dissimilarity matrices
Dissimilarity matrices were calculated using all 32 additions and subtraction problems, sorted by operand 1, then by operation (additions first) and finally by operand 2: (3+0, 3+1, 3+2, 3+3, 3-0, 3-1, 3-2, 3-3, etc., see Figure A.2). Visual models were calculated

using a method that rates the similarity of the digits based on their putative responses in inferior temporal cortex. Magnitude models used the numerical distance between numbers. For the operation, the matrix was composed by 0s (same operation) and 1s (different operations).

Several RSA models were constructed to test specific relationships between different dimensions of the stimuli and the MEG signals. The theoretical representational dissimilarity matrices (RDM) were constructed using the 32 additions and subtraction problems, which we sorted by operand 1, then by operation (additions first) and finally by operand 2: (3+0, 3+1, 3+2, 3+3, 3-0, 3-1, 3-2, 3-3, etc., see Figure A.2). Seven theoretical RDM matrices were constructed, either based on the magnitude dissimilarity (numerical distance) or visual dissimilarity (using a method that captures the hypothetical responses of inferior temporal cortex), separately for operand 1, operand 2, and the correct result, plus a matrix for category-based similarity for addition vs. subtractions (see Figure 4). Those theoretical matrices were used as regressors on the observed matrices derived from the MEG data, i.e., the dissimilarities between the 32 averaged event-related MEG topographies. Such regressions were conducted at each time step, thus allowing us to visualize the time course of the corresponding neural codes.

We first tested whether and when the visual and magnitude dimensions of the operands could be recovered from MEG signals. As can be seen from Figure 5, both the visual and magnitude models of the operands had significant correlations with the observed RDM following operand onset. Specifically, the visual model of operand 1 showed a significant effect at 128 ms after visual appearance of operand 1, with a peak at 168 ms and lasting up to 544 ms ($p < 0.05$, corrected for multiple comparisons). Around the same time, the magnitude model for operand 1 had a smaller, but significant effect, starting at 112 ms with a peak at 328 ms and lasting until 328 ms ($p < 0.05$, corrected). For operand 2 the pattern was somehow inverted. The magnitude model had a stronger effect which started at 1,664 ms (94 ms after the onset of operand 2, which occurred at 1,570 ms). This effect peaked at 1,704 ms and lasted until 2,608 ms ($p < 0.05$, corrected), i.e. longer than the visual model (start = 1,672 ms, peak = 1,808, lasting until 2,392 ms, $p < 0.05$, corrected).

Since the visual and magnitude models partially correlated with each other, we next investigated the unique variance explained by each model, while regressing out the effect of the other model.

For operand 1, the magnitude model did not reach statistical significance at any time point after regressing out the visual model. Conversely, the visual model had two small significant values at ~ 224 ms and ~544 ms after controlling for the magnitude model ($p < 0.05$, corrected). In contrast, the magnitude model remained significant for operand 2 after regressing out the effect of the visual model from 1,672 to 2,360 ms ($p < 0.05$, corrected). Conversely the visual model also remained significant after controlling for the magnitude model, but for a shorted period, from 1,696 to 2,103 ms ($p < 0.05$, corrected).

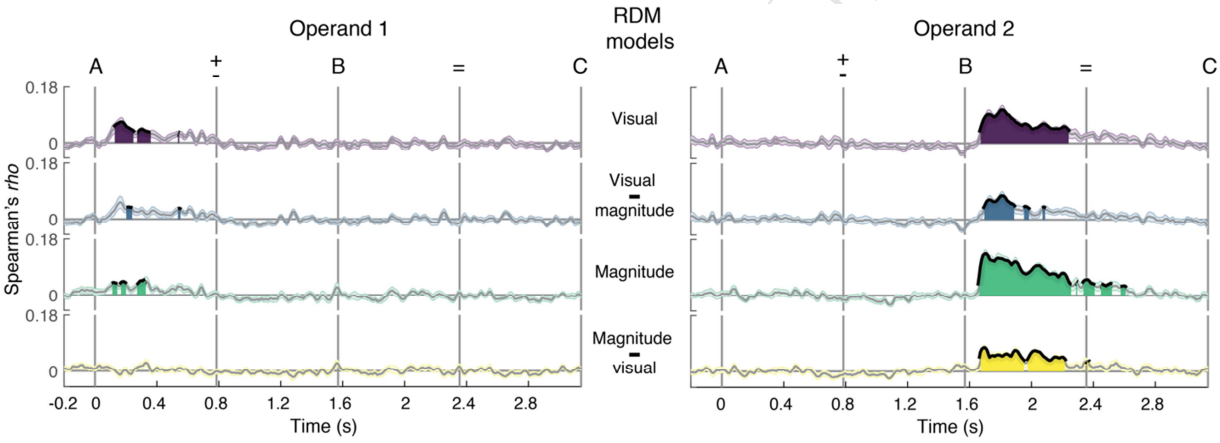


Figure 5. Representational geometries of the operands
A series of RSA models (see Figure 4) were used to investigate the temporal dynamics of the representation of operand 1 and operand 2. Correlations between the theoretical and observed dissimilarity matrix were performed at each time sample. We first correlated the RSA for single predictors (visual and magnitude, lines 1 and 3). Next, to test the unique variance explained by each model, we partialled out the effect of the other model (visual – magnitude and magnitude – visual; lines 2 and 4). Gray horizontal lines indicate theoretical chance level. Thick lines and filled areas represent time periods in which the second-level statistical tests across subjects revealed a correlation coefficient significantly above 0 (cluster corrected for multiple comparison, $p < 0.05$).

Overall, the RSA corroborates the decoding results, by showing that the representational geometry can be better retrieved from MEG signals for operand 2 compared to operand 1. Crucially, the RSA revealed that both visual and magnitude dimensions of the operands are coded at about the same time. While the dominant

dimension for operand 1 was visual, both visual and magnitude dimensions could be independently retrieved from operand 2, but with a predominance of the magnitude dimension.

3.11. Inability to decode the internally computed result

We next searched the data for a representation of the internally computed correct result (i.e. A+B or A-B, depending on the operation) (Figure 6). The visual model had no significant effect across the entire trial. The magnitude model was transiently significant, but only right after the presentation of the operation sign, that is, before the actual calculation could have started. An additional estimator using Ridge regression corroborated this finding (see Figure A.1). This result was probably driven by the correlation between the magnitude and the operation sign, since our experimental design had additions and subtractions with matched operands, thus additions produced overall higher results and subtractions smaller results (see Methods). Confirming this intuition, after regressing out the effect of the operation model, the magnitude model did not explain any unique variance, whereas conversely the effect of operator was virtually unchanged when regressing out the magnitude model of the result and it even showed a transient reactivation after the presentation of the equal sign, similarly to the decoding analysis (see Figure 2).

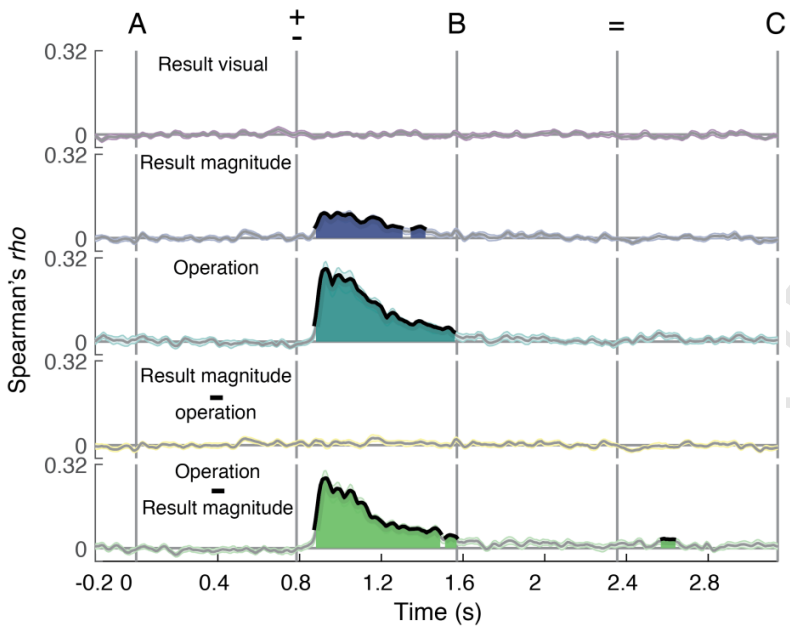


Figure 6. Attempting to decode the internally computed result
We first correlated the RSA for single predictors of the result (visual and magnitude, lines 1 and 2). Next, to test the unique variance explained by each model, we partialled out the effect of the operation (line 3) from the magnitude model (line 4) and vice versa (line 5). Gray horizontal lines indicate theoretical chance level. Thick lines and filled areas represent time periods in which the second-level statistical tests across subjects revealed a correlation coefficient significantly above 0 (cluster corrected for multiple comparison, $p < 0.05$).

Because we were surprised at our inability to decode the internal computed result, we performed several additional analyses, but none were successful. Here we briefly describe the rationale behind each strategy. First, we explored event related fields (ERF) at the univariate level, using Fieldtrip cluster-based method. Within the time window between operand 2 and the equal sign, or between operand 2 and proposed result (when subjects are supposedly performing the exact calculation), the cluster-based permutation test did not reveal any cluster with a significant correlation with the correct result. We first did this analysis while grouping together additions and subtractions, then replicated it while analyzing them separately, and also within each group of MEG sensors, to no avail.

As regards multivariate analyses, we first attempted to predict whether the internally computed result was 3, 4, 5 or 6. The rationale, as explained in the Methods section, was that experimental design used additions and subtractions matched by

operands, thus imposing an inhomogeneity on the distribution of results. Therefore, for the main decoding analysis, we only used the most homogeneously distributed results (numbers 3-6), which overall represented 62.48 % of the trials. As described earlier, this analysis did not result in any significant decoding score. We reasoned that if the brain signals associated with the computed results are weak, it might be better to first train the decoder on an explicitly presented number using a large training set, and only test its generalization to the internally computed result. This was done by training the model to decode the operand 1 during the first 800 ms, and then testing its generalization to the internally computed result. At no time point prior to the presentation of the correct result did we find any significant cross-generalization classification score.

We also trained a classifier to decode the proposed result (when it was correct) time-locked to the proposed result (for 800 ms) and tested if it could generalize backwards to the correct result at the time window between the operand 2 and the proposed result. This model only included 30% of the trials and learning was not above chance for decoding the proposed result, therefore no generalization could be tested on the internally computed result.

Another possibility is that the result is coded in the spectral domain, perhaps within a specific frequency band. To explore that, we used a searchlight approach in time, sensor space and frequency (using the Matlab Cosmo MVPA Toolbox (Oosterhof, Connolly, & Haxby, 2016)). We fitted a series of linear discriminant analysis (LDA) estimators (instead of SVM, for computational simplicity) with our standard cross-validation scheme to classify the main variables of interest (operand 1, operation, operand 2 and result), with the following procedure. First, we selected two frequency bands (low: 1 – 34 Hz and high: 34 - 100 Hz). Next, we selected one sensor (only gradiometers) to be the center of the “sphere” and included its 10 closest neighbor sensors. The matrix of features was therefore composed of single-trial MEG frequency power signals (X, shape = n_{trials} x ($10_{sensors}$ x $1_{time\ sample}$ x $1_{frequencies}$)). No significant classifications scores were found in the high frequencies. As can be seen in Figure A.4, the operand 1, operation and operand 2 could be decoded generally from occipital-parietal sensors, at a short time window following their respective onsets and mostly between 3 to 20 Hz, a frequency band which corresponds to event-related signals and

is a classical finding for visually presented stimuli (King, Pescetelli, & Dehaene, 2016). However, no sign of above-chance classification was found for the result in any group of sensors at the time point between the operand 2 and the proposed result and in any frequency.

Finally, we reasoned that, if the computation time varied on a trial-by-trial basis, the brain response induced by the internally computed result could be brief and diluted in time, thus obscuring its decodability when the trials were time-locked to operand onset. We tried to overcome this timing issue by computing the Fourier spectrum of the low-pass signal in the low frequency range (2 – 34 Hz) using the entire time window from B to C, then feeding the classifier with a feature matrix of single-trial MEG frequency power (X, shape = n_{trials} x $n_{frequencies}$). The logic is that once phase information is removed, the Fourier spectrum is invariant for temporal delays. No significant classification was found. Additionally, we tested a classifier based on Riemannian geometry using a covariance matrix estimation that integrates the temporal information (ERPCov, see Methods). This pipeline was applied to classify the operand 1, operation, operand 2 and result, in two time windows (0 – 800 ms and 800 – 1,600). Results are summarized in Figure A.5. As can be seen, the ERPCov classifier boosted the classification accuracies for operand 1, operation and operand 2 (especially in the 0 – 800 ms window), but yielded no significant classification accuracy for the result. Therefore, we conclude that in the current dataset, the internally computed result could not be decoded from MEG signals.

4. Discussion

By combining time-resolved multivariate pattern analysis (MVPA) to MEG signals, we obtained a comprehensive picture of the unfolding processing stages underlying arithmetic calculations. Our verification task, using sequentially presented addition and subtraction problems, allowed us to investigate the main components of mental arithmetic: encoding of the operands, processing of the operation sign, calculation, decision of correctness, and finally response preparation and execution. Overall Global Field Power (GFP) revealed that the activity was sustained during the entire trial, with additional transient peaks at ~150 ms after each stimulus. MEG topographies showed

that the evoked responses evolved across time from posterior to anterior sensors, both after each stimuli onset and also across the entire trial, which fits nicely with previous electrophysiological findings on arithmetic processing (Dehaene, 1996) and visual object processing in general (Cichy, Pantazis, & Oliva, 2014; King et al., 2016; Sergent, Baillet, & Dehaene, 2005).

4.1. A cascade of partially overlapping processing stages in mental arithmetic

Crucially, we could decode a series of calculation features, revealing a cascade of partially overlapping brain states during the solution of a problem as simple as 3+2=5. First, we could transiently decode the identity of the operand 1 between 112 - 640 ms after stimuli onset. Next, the operation (addition vs. subtraction) could be decoded from 95 ms after the onset of the operation sign, dropping somewhat 700 ms after the sign, but remaining above chance until the offset of operand 2, with a subsequent transient recovery after the onset of the equal sign (significant decoding for ~2,000 ms). The high initial classification score is most likely due to the visual difference between the operation signs, but could also reflect task-specific preparation, such as operator priming (Fayol & Thevenot, 2012) as well as visual-spatial mechanisms or spatial-numerical associations (Hartmann et al., 2015; Masson & Pesenti, 2014; Mathieu et al., 2017, 2016), as would follow from the idea that calculation is essentially a movement along the mental number line (Knops, Thirion, et al., 2009; Knops, Viarouge, & Dehaene, 2009; Pinheiro-Chagas, Dotan, et al., 2017). In line with this hypothesis, we found that a classifier trained on discriminating subtractions vs. additions cross-generalized and accurately discriminated smaller vs. larger numbers, respectively, but only at the time of presentation of operand 2, which is probably the stage in which subjects are calculating or manipulating quantities. In fact, the identity of the operand 2 could be decoded for an extended time window, ranging from 102 ms after stimulus onset till the offset of the equal sign, thus partially overlapping with the decoding of operation for about 1,000 ms. This results fits with our recent behavioral findings that the operator sign transiently affected the decision about the location of the result of arithmetic calculations on a number line (a plus sign attracted the finger to the right [larger results] and a minus sign

to the left [smaller results]) around the time that subjects were processing operand 2 (Pinheiro-Chagas, Dotan, et al., 2017). The existence of a code that is partially common across the elaboration of an arithmetical sign and a number also comes from behavioral data showing that both stimuli (an arithmetical sign and a number) trigger shifts in spatial attention that are consistent with a left-to-right oriented representation, thus facilitating target detection (Fischer, Castel, Dodd, & Pratt, 2003; Mathieu et al., 2016). Nevertheless, our cross-generalization result alone could also be explained by the fact that subtractions normally produce smaller numbers than additions, without postulating the existence of a spatially organized mental number line.

Importantly, operand 2 was classified with a higher accuracy as compared to operand 1 (Figure 2 and Figure A.5), suggesting that more intense and more stable brain activity occurred after operand 2 than after operand 1. This is understandable given that, at this stage subjects were able to start calculating, a process whose duration depends on the size of the *min* operand (or the smallest operand) (Groen & Parkman, 1972; Pinheiro-Chagas, Dotan, et al., 2017; Uittenhove et al., 2016), which in the present experiment is precisely operand 2. These results also fit with recent neurophysiological findings. An ECoG study using an essentially identical verification task (Hermes et al., 2015) showed that neuronal populations in the ventral temporal cortex (VTC) have stronger activity following operand 2 as compared to operand 1, with an averaged time course very similar to our Figure 1B. This finding was interpreted as suggesting that the VTC activity is modulated by task demands, in this case the actual manipulation of numbers, which can only happen after operand 2 (Hermes et al., 2015). A more recent ECoG study revealed that in addition to the number form area (NFA) in the ventral temporal cortex (VTC), which selectively responds to numerical digits independently of the presentation context (Shum et al., 2013, for reviews see Hannagan, Amedi, Cohen, Dehaene-Lambertz, & Dehaene, 2015; Price, Yeo, Wilkey, & Cutting, 2016), there are neuronal populations in the posterior inferior temporal gyrus (pITG) (just adjacent to the NFA), that respond slightly later (~10 ms) and exhibit more sustained activity than the NFA. Crucially, these lateral sites respond only when numerals are presented in the context of a calculation or, in the case of the sequentially presented verification task, only for operand 2 and the proposed result, but not for

operand 1 (Daitch et al., 2016). Thus, these results provide a plausible psychophysiological basis for our finding that operand 2 can be decoded with a higher accuracy as compared to operand 1.

Although it was not the aim of our paper to arbitrate between different cognitive models of mental arithmetic, our results impose some restrictions to fact-retrieval models (Ashcraft, 1982; Campbell, 1995). These models assume that single-digit additions and subtractions are solved by directly retrieving the result from a long-term memory representation of arithmetic facts, without relying on any calculation, procedure or quantity manipulation. Therefore, they do not provide any prediction or explanation for the higher accuracy in decoding the operand 2, nor for the cross-generalization from subtractions vs. additions to smaller vs. larger numbers. Conversely, these results can easily be accommodated by a models that postulates that single-digit additions and subtractions are computed by a stepwise displacement on a spatially organized mental number line (Knops, Viarouge, & Dehaene, 2009; Uittenhove et al., 2016), starting with the larger number and incrementally adding or subtracting the smaller number (Pinheiro-Chagas, Dotan, et al., 2017).

4.2. The representational geometries of the operands

Although those ECoG studies were very informative about the fine-grained spatial-temporal dynamics of calculations, they did not provide any direct indication about the nature of the underlying representations of the operands. Here, to investigate this question, we applied time-resolved representational similarity analysis (RSA). Our results indicated that while for operand 1 the dominant dimension represented was visual, for operand 2 both visual and magnitude dimensions explained unique variance in the MEG signal. A similar conclusion, corroborating this finding, could be drawn from the results of the multivariate regression analysis (Figure A.3), in which only the estimator for operand 2 achieved above chance performance. Although a natural prediction for operand 2 would be that the visual dimension precedes and partially overlaps with the magnitude dimension, we observed an effect of the two dimensions starting practically at the same time, at ~100 ms after stimuli onset, but the magnitude dimension was predominant and lasted longer (Figure 2). As ECoG suggested that the

difference in latency between NFA and both pITG and IPS is very small (~14 ms), it is possible that we did not have a high enough signal-to-noise ratio to separate in time the visual and magnitude dimensions with MEG. Further ECoG studies specially designed for this purpose could provide a definitive answer. It is also important to note that in our experiment, the magnitude of the operand 2 defines the problem-size, the size of the *min* operand that needs to be added or subtracted, and which is known to be a major determinant of calculation duration and difficulty (Groen & Parkman, 1972; Pinheiro-Chagas, Dotan, et al., 2017). Therefore, the decoding of operand 2 and its correlation with the magnitude model of the RSA could be a combination of the quantity representation and the calculation process itself. Future experiments should aim at disentangling these two processes.

4.3. Parsing the processing stages of arithmetic decision-making

At the decision stage (Figure 2, time-locked to C), we found a fast and highly overlapping dynamics of identifying the proposed result (from 92 till 400 ms), judging whether it was correct or incorrect (from 172 ms till the end of the trial) and finally pressing the response button (from 196 ms till the end of the trial). The last two stages were better observed when time-locking the signal to the RT. We could see a slow ramping in the decoding of the correctness starting at -428 ms before the RT and persisting until the end of the trial, followed by a fast and sharp increase of classification score for the response button at -212 ms before the RT (Figure 2). It is important to highlight that those three features (proposed result, correctness and response button) are orthogonal to each other in our experimental design, so the classifiers could no rely on a single feature to perform above chance level.

The proposed result was transiently decoded after its onset, but we did not observe a distance effect for the incorrect trials (absolute distances = 1 - 4) in both behavioral and electrophysiological levels (no significant decoding scores), which is at odds with previous positive findings (Avancini, Galfano, & Szucs, 2014; Avancini et al., 2015; Dehaene, 1996). We believe that this null finding was probably due to a combination of the small distances used (1 - 4), and the slow pace of our experimental design. As a result, subjects probably had the correct result in mind for at least 1 s

before the proposed result appeared, and could perform a fast symbolic same-different judgement without showing any influence of numerical distance.

4.4. Temporal dynamics of the decoding patterns

The decoding patterns of the calculation features observed in this experiment are far from trivial and deserve attentive consideration. Due to the sequential structure of our task, a series of informations had to be maintained in working memory. For example, to correctly perform the task, subjects needed to keep in mind the operand 1 at least until the operand 2 was presented. Yet, surprisingly, the classification score for operand 1 rapidly decreased to chance level after stimulus offset and remained so until the end of the trial. A similar result was observed in a series of working memory studies in which the information could not be decoded in a sustained way during the memory maintenance period, suggesting that, contrary to previous suggestions, working memory may not be encoded by a stable pattern of sustained activity (LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2013; Sprague, Ester, & Serences, 2016; Trübutschek et al., 2017; Wolff et al., 2017). A slightly different decoding time course was observed for the operation and operand 2: both features remained decodable for a much longer time (above 1,000 ms), although again with a drastic drop in accuracy after 800 ms. Finally, remember that we could not decode the internally generated result, even though subjects were instructed to compute it and keeping it “in mind” during the delay prior to the appearance of the proposed result.

Several theories may explain either the complete absence of decodable sustained activity, or the strong decrease in the decoding performance, during the various delay periods of our arithmetic task. First, instead of stable sustained neural firing, information might be maintained in working memory through occasional gamma and beta bursts (Lundqvist et al., 2016) which would therefore be diluted in time and which our MEG signals might not be sensitive enough to capture. Second, the coding scheme to store information in working memory may not be through persistent neuronal firing, but through short-term synaptic changes (Mongillo, Barak, & Tsodyks, 2008), so called ‘silent states’ (Stokes, 2015; Trübutschek et al., 2017) and therefore may not be directly measurable with conventional neuroimaging methods. Finally, a third possibility

is that the neural coding schemes changes across successive stages from an easily decodable spatial code based on large cortical columns in posterior areas, to a more microscopic and sparse code in the prefrontal cortex and other associated areas, based on overlapping neural populations and orthogonal vectors (Mante, Sussillo, Shenoy, & Newsome, 2013), which may therefore not be detectable with MEG. All three possibilities are plausible, and fine-grained electrophysiological recordings will be needed to separate them.

The temporal generalization analysis (King & Dehaene, 2014) revealed that the underlying codes of the main calculation features are highly dynamic along the trial, as indicated by a diagonal generalization-across-time matrix showing that they remained decodable only when train and test time were similar. (Figure A.1). The sole exception was around 200-400 ms after the presentation of the operation sign and operand 2, where a thicker diagonal, closer to a square pattern of generalization, suggested a more stable neural code. Such a succession of diagonal and then square pattern has been systematically observed in several studies (Crouzet, Busch, & Ohla, 2015; King et al., 2016; Marti, King, & Dehaene, 2015; Stokes, Wolff, & Spaak, 2015; Trübutschek et al., 2017) and has been interpreted (King et al., 2016; Trübutschek et al., 2017) as compatible with classical cascade models (McClelland, 1979), suggesting that information is encoded by an initial cascade of successive neural codes, followed by a more sustained (though still transient) activity during later decision or working-memory stages. It also corroborates a series of functional and anatomical findings on the highly hierarchical organization of the cortex (Chaudhuri, Knoblauch, Gariel, Kennedy, & Wang, 2015; Cichy & Teng, 2016; Felleman & Van Essen, 1991; King et al., 2016; Rajalingham, Schmidt, & DiCarlo, 2015). Because of the series of mental transformation involved in our task, some of the features could be discarded along the way and substituted by their transformed or combined version. For example, the operands probably underwent a series of visual processing stages before their symbolic identity was established. Similarly, during calculation, operand 1, operand 2 and the operation sign were probably transformed into an internal representation of the computed result after the presentation of operand 2, and from this stage on, the task required only that result to be maintained in working memory for later comparison with the proposed result.

4.5. The search for the neural correlates of the internally computed result

With this idea in mind, we systematically searched for a neural signature of this internally computed result. Surprisingly, however, none of our attempts were successful. Could this finding arise from limitations in our experimental design? One potential weak point is that our task did not allow to establish the precise moment when the calculation was completed, which probably varied on a trial-by-trial basis. Our hypothesis, however, was that the activity induced or evoked by the correct result would last until the proposed result appeared, so that we would decode it without necessarily time-locking the signal to the peak of activation generated by the correct result. For instance, although RT systematically varies across trials, we did not need to time lock the response button press to RT to achieve above-chance classification score when decoding the response side (Figure 2). This strategy did not work, however, for the internally computed result. To overcome this potential timing limitation, we tried some decoding models which received as input the induced oscillatory activity in a wide frequency range and one model that used Riemannian geometry (Barachant & Congedo, 2014) and embeds the temporal information of the signal by concatenating along the sensor axis the averaged ERF (across trial) of each class, and is therefore well suited to capture both evoked and induced responses. Although the latter estimator indeed boosted decoding scores for the other calculation features of interest (operand 1, operation and operand 2), it showed no improvement to decode the correct result.

After testing several robust state-of-the-art decoding models, we therefore conclude that the internally computed result of a simple arithmetic calculation is not as easily decodable from MEG signals as the externally presented stimuli. This finding could originate from the same three explanations listed above to account for the vanishing of the codes for operand 1 and 2: brief bursts of gamma or beta activity; short-term synaptic codes; or overlapping neural microcodes. Follow-up studies could try to use a larger number of trials to train the classifiers.

Additionally, it is also possible that, since we used a verification task, subjects did not need to calculate in every trial. They could use a range of rule-based strategies, such as comparing the parity and size between the operands and the proposed result.

However, subjects were explicitly asked to calculate in order to judge as fast as possible the correctness of the proposed result. Moreover, a series of findings indicate that they did indeed engage in calculation. First, we could decode the additions vs. subtractions ~2,000 ms after the onset of the operation sign, overlapping with the decoding of the operand 2 for about 1,000 ms. Second, we found higher classification accuracies for decoding the operand 2 vs. operand 1, and the magnitude model in the RSA was dominant for predicting operand 2. These results suggest that the actual calculation process was initiated after the onset of operand 2 and lasted until the offset of the equal sign (around 1,200 ms), when both operation type and operand 2 could no longer be decoded.

Finally, it is possible that subjects were purely retrieving the result of the problems from long-term memory, as suggested by fact-retrieval models (Ashcraft, 1982; Campbell, 1995). But even in this case, we would still expect to be able to decode the correct result, since it had to be internally generated and maintained in working memory somehow to be compared with the proposed result.

MEG decoding has been successfully applied to characterize the spatial-temporal dynamics of several cognitive functions, such dual-task interference (Marti et al., 2015), attention (Brandman & Peelen, 2017; Kaiser, Azzalini, & Peelen, 2016), working memory (King et al., 2016; Trübutschek et al., 2017; Wolff, Jochim, Akyürek, & Stokes, 2017), reward value (Bach, Symmonds, Barnes, & Dolan, 2017), taste perception (Crouzet et al., 2015), object recognition and categorization (Carlson et al., 2011, 2013; Cichy et al., 2014; Isik et al., 2014), written and spoken language (Chan, Halgren, Marinkovic, & Cash, 2011; Kocagoncu, Clarke, Devereux, & Tyler, 2017), etc. However, virtually all of these studies either used classifiers that could rely on activity evoked by low-level sensory properties of the stimuli (or mental imagery), or probed classical semantic categories that are known to be anatomically segregated. Therefore, evidence for within category time-resolved decoding at the single-trial level of abstract internally generated mental objects is still lacking. One reason might be that such mental objects, like the result of a calculation, are represented in a highly distributed fashion, difficult to measure with non-invasive methods that have a relatively low signal-

to-noise ratio, therefore suggesting the existence of different neural substrates for externally and internally generated codes.

5. Conclusion

Despite its inability to decode the internally computed result, the present study provides a first picture of the series of successive processing stages and mental transformations that unfold during a simple arithmetic calculation. The results reveal a highly dynamic coding profile and a cascade of partially overlapping brain states during elementary arithmetic, therefore increasing our understanding of the neurocognitive underpinnings of high level symbolic cognition.

6. Acknowledgments

This research was sponsored by INSERM, CEA, and the Bettencourt-Schueller Foundation. S. Dehaene is supported by an advanced grant “NeuroSyntax” from the European Research Council (ERC). Pedro Pinheiro-Chagas gratefully acknowledges a Science Without Borders Fellowship from the Brazilian National Council for Scientific and Technological Development (CNPq) (nr. 246750/2012-0). We are grateful to all subjects who participated in the study and to Veronique Joly-Testault, Laurence Laurier and all the NeuroSpin team who helped recruiting them. We would also like to thank Valentina Borghesani, Marco Buiatti and Darinka Trübutschek for methodological advice. And Sebastian Marti, Fosca Al-Roumi, Maxime Maheu and Lina Teichmann for helpful discussions.

7. References

Amalric, M., & Dehaene, S. (2016). Origins of the brain networks for advanced mathematics in expert mathematicians. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(18), 4909–4917. http://doi.org/10.1073/pnas.1603205113

Amalric, M., & Dehaene, S. (2017). Cortical circuits for mathematical knowledge: evidence for a major subdivision within the brain's semantic networks. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *373*(1740), 20160515. http://doi.org/10.1098/rstb.2016.0515

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review*, *2*(3), 213–236. http://doi.org/10.1016/0273-2297(82)90012-0

Ashcraft, M. H., & Battaglia, J. (1978). Cognitive arithmetic: Evidence for retrieval and decision processes in mental addition. *Journal of Experimental Psychology: Human Learning & Memory*, *4*(5), 527–538. http://doi.org/10.1037/0278-7393.4.5.527

Avancini, C., Galfano, G., & Szucs, D. (2014). Dissociation between arithmetic relatedness and distance effects is modulated by task properties: An ERP study comparing explicit vs. implicit arithmetic processing. *Biological Psychology*, *103*, 305–316. http://doi.org/10.1016/j.biopsycho.2014.10.003

Avancini, C., Soltész, F., & Szucs, D. (2015). Separating stages of arithmetic verification: An ERP study with a novel paradigm. *Neuropsychologia*, *75*, 322–329. http://doi.org/10.1016/j.neuropsychologia.2015.06.016

Bach, D. R., Symmonds, M., Barnes, G., & Dolan, R. J. (2017). Whole-Brain Neural Dynamics of Probabilistic Reward Prediction. *The Journal of Neuroscience*, *37*(14), 3789–3798. http://doi.org/10.1523/JNEUROSCI.2943-16.2017

Barachant, A., & Congedo, M. (2014). A plug & play P300 BCI using information geometry. Retrieved from http://arxiv.org/abs/1409.0107

Brandman, T., & Peelen, M. V. (2017). Interaction between scene and object processing revealed by human fMRI and MEG decoding. *The Journal of Neuroscience*, 582–17. http://doi.org/10.1523/JNEUROSCI.0582-17.2017

Butterworth, B., Zorzi, M., Girelli, L., & Jonckheere, A. R. (2001). Storage and retrieval of addition facts: the role of number comparison. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *54*(4), 1005–1029. http://doi.org/10.1080/713756007

Campbell, J. I. D. (1994). Architectures for numerical cognition. *Cognition*, *53*(1), 1–44. http://doi.org/10.1016/0010-0277(94)90075-2

Campbell, J. I. D. (1995). Mechanisms of simple addition and multiplication: A modified network-interference theory and simulation. *Mathematical Cognition*, *1*(2), 121–164.

Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., & Turret, J. (2011). High temporal resolution decoding of object.pdf. *Journal of Vision*, *11*(2011), 1–17. http://doi.org/10.1167/11.10.9.Introduction

Carlson, T. A., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1–1. http://doi.org/10.1167/13.10.1

Chan, A. M., Halgren, E., Marinkovic, K., & Cash, S. S. (2011). Decoding word and category-specific spatiotemporal representations from MEG and EEG. *NeuroImage*, *54*(4), 3028–3039. http://doi.org/10.1016/j.neuroimage.2010.10.073

Chaudhuri, R., Knoblauch, K., Gariel, M. A., Kennedy, H., & Wang, X. J. (2015). A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. *Neuron*, *88*(2), 419–431. http://doi.org/10.1016/j.neuron.2015.09.008

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–62. http://doi.org/10.1038/nn.3635

Cichy, R. M., & Teng, S. (2016). Resolving the neural dynamics of visual and auditory scene processing in the human brain : a methodological approach. *Philos.Trans. R. Soc. B*, *372*, 1--11. http://doi.org/10.1098/rstb.2016.0108

Crouzet, S. M., Busch, N. A., & Ohla, K. (2015). Taste quality decoding parallels taste sensations. *Current Biology*, *25*(7), 890–896. http://doi.org/10.1016/j.cub.2015.01.057

Daitch, A. L., Foster, B. L., Schrouff, J., Rangarajan, V., Kasikci, I., Gattas, S., … Parvizi,

J. (2016). Mapping human temporal and parietal neuronal population activity and functional coupling during mathematical cognition. *Proceedings of the National Academy of Sciences*, *113*(46), 201608434. http://doi.org/10.1073/pnas.1608434113

Dehaene, S. (1996). The Organization of Brain Activations in Number Comparison: Event-Related Potentials and the Additive-Factors Method. *Journal of Cognitive Neuroscience*, *8*(1), 47–68. http://doi.org/10.1162/jocn.1996.8.1.47

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three Parietal Circuits for Number Processing. *Cognitive Neuropsychology*, *20*(3–6), 487–506. http://doi.org/10.1080/02643290244000239

Diedrichsen, J., & Kriegeskorte, N. (2017). *Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis*. *PLoS Computational Biology* (Vol. 13). http://doi.org/10.1371/journal.pcbi.1005508

Dotan, D., & Dehaene, S. (2013). How do we convert a number into a finger trajectory? *Cognition*, *129*(3), 512–529. http://doi.org/10.1016/j.cognition.2013.07.007

Dotan, D., & Dehaene, S. (2015). The origins of logarithmic number-to-position mapping. *In Press*, *123*(6), 1–52. http://doi.org/10.1037/rev0000038

Eger, E., Michel, V., Thirion, B., Amadon, A., Dehaene, S., & Kleinschmidt, A. (2009). Deciphering cortical number coding from human brain activity patterns. *Current Biology : CB*, *19*(19), 1608–15. http://doi.org/10.1016/j.cub.2009.08.047

Eger, E., Pinel, P., Dehaene, S., & Kleinschmidt, A. (2015). Spatially invariant coding of numerical information in functionally defined subregions of human parietal cortex. *Cerebral Cortex*, *25*(5), 1319–1329. http://doi.org/10.1093/cercor/bht323

Fayol, M., & Thevenot, C. (2012). The use of procedural knowledge in simple addition and subtraction problems. *Cognition*, *123*(3), 392–403. http://doi.org/10.1016/j.cognition.2012.02.008

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47. http://doi.org/10.1093/cercor/1.1.1-a

Fischer, M. H., Castel, A. D., Dodd, M. D., & Pratt, J. (2003). Perceiving numbers

causes spatial shifts of attention. *Nature Neuroscience*, *6*(6), 555–556. http://doi.org/10.1038/nn1066

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, (7 DEC). http://doi.org/10.3389/fnins.2013.00267

Groen, G. J., & Parkman, J. M. (1972). A chronometric analysis of simple addition. *Psychological Review*, *79*(4), 329–343. http://doi.org/10.1037/h0032950

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2016). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *26*(3), 1–21. http://doi.org/10.1162/jocn_a_01068

Hannagan, T., Amedi, A., Cohen, L., Dehaene-Lambertz, G., & Dehaene, S. (2015). Origins of the specialization for letters and numbers in ventral occipitotemporal cortex. *Trends in Cognitive Sciences*. http://doi.org/10.1016/j.tics.2015.05.006

Hartmann, M., Mast, F. W., & Fischer, M. H. (2015). Spatial biases during mental arithmetic: Evidence from eye movements on a blank screen. *Frontiers in Psychology*, *6*(JAN). http://doi.org/10.3389/fpsyg.2015.00012

Harvey, B. M., Klein, B. P., Petridou, N., & Dumoulin, S. O. (2013). Topographic Representation of Numerosity in the Human Parietal Cortex. *Science*, *341*(6150), 1123–1126. http://doi.org/10.1126/science.1239052

Hermes, D., Rangarajan, V., Foster, B. L., King, J.-R., Kasikci, I., Miller, K. J., & Parvizi, J. (2015). Electrophysiological Responses in the Ventral Temporal Cortex During Reading of Numerals and Calculation. *Cerebral Cortex (New York, N.Y.: 1991)*. http://doi.org/10.1093/cercor/bhv250

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102. http://doi.org/10.1152/jn.00394.2013

Kaiser, D., Azzalini, D. C., & Peelen, M. V. (2016). Shape-independent object category responses revealed by MEG and fMRI decoding. *Journal of Neurophysiology*, jn.01074.2015. http://doi.org/10.1152/jn.01074.2015

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural

images from human brain activity. *Nature*, *452*(7185), 352–355. http://doi.org/10.1038/nature06713

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. http://doi.org/10.1016/j.tics.2014.01.002

King, J.-R., Pescetelli, N., & Dehaene, S. (2016). Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information. *Neuron*, *92*(5), 1122–1134. http://doi.org/10.1016/j.neuron.2016.10.051

Knops, A., Thirion, B., Hubbard, E. M., Michel, V., & Dehaene, S. (2009). Recruitment of an area involved in eye movements during mental arithmetic. *Science (New York, N.Y.)*, *324*(5934), 1583–5. http://doi.org/10.1126/science.1171599

Knops, A., Viarouge, A., & Dehaene, S. (2009). Dynamic representations underlying symbolic and nonsymbolic calculation: evidence from the operational momentum effect. *Attention Perception Psychophysics*, *71*(4), 803–821. Retrieved from http://app.psychonomic-journals.org/content/71/4/803.full.pdf

Knops, A., Viarouge, A., Dehaene, S., Cea, F., Universit, F., & When, F. (2009). Dynamic representations underlying symbolic and nonsymbolic calculation: evidence from the operational momentum effect. *Attention Perception Psychophysics*, *71*(4), 803–821. http://doi.org/10.3758/APP

Kocagoncu, E., Clarke, A., Devereux, B. J., & Tyler, L. K. (2017). Decoding the Cortical Dynamics of Sound-Meaning Mapping. *The Journal of Neuroscience*, *37*(5), 1312–1319. http://doi.org/10.1523/JNEUROSCI.2858-16.2016

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*. http://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., Ruff, D. D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of mand and monkey. *Neuron*, *60*(6), 1126–1141. http://doi.org/10.1016/j.neuron.2008.10.043.Matching

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding Attended Information in Short-term Memory: An EEG Study.

Journal of Cognitive Neuroscience, *25*(1), 127–142. http://doi.org/10.1162/jocn_a_00305

Lehmann, D., & Skrandies, W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and Clinical Neurophysiology*, *48*(6), 609–621. http://doi.org/10.1016/0013-4694(80)90419-8

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L. L., Buschman, T. J. J., & Miller, E. K. K. (2016). Gamma and Beta Bursts Underlie Working Memory. *Neuron*, *90*(1), 152–164. http://doi.org/10.1016/j.neuron.2016.02.028

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. http://doi.org/10.1038/nature12742

Marti, S., King, J.-R., & Dehaene, S. (2015). Time-Resolved Decoding of Two Processing Chains during Dual-Task Interference. *Neuron*, *88*(6), 1297–1307. http://doi.org/10.1016/j.neuron.2015.10.040

Masson, N., & Pesenti, M. (2014). Attentional bias induced by solving simple and complex addition and subtraction problems. *The Quarterly Journal of Experimental Psychology*, *67*(8), 1514–1526. http://doi.org/10.1080/17470218.2014.903985

Mathieu, R., Epinat-duclos, J., Léone, J., Fayol, M., & Thevenot, C. (2017). Developmental Cognitive Neuroscience Hippocampal spatial mechanisms relate to the development of arithmetic symbol processing in children. *Developmental Cognitive Neuroscience*, (August 2016), 0–1. http://doi.org/10.1016/j.dcn.2017.06.001

Mathieu, R., Gourjon, A., Couderc, A., Thevenot, C., & Prado, J. (2016). Running the number line: Rapid shifts of attention in single-digit arithmetic. *Cognition*, *146*, 229–239. http://doi.org/10.1016/j.cognition.2015.10.002

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychol Rev*, *86*(4), 287–330. http://doi.org/10.1037/0033-295X.86.4.287

McCrink, K., Dehaene, S., & Dehaene-Lambertz, G. (2007). Moving along the number line: operational momentum in nonsymbolic arithmetic. *Perception &*

Psychophysics, *69*(8), 1324–33. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18078224

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*, *319*(5869), 1543–1546. http://doi.org/10.1126/science.1150769

Nieder, A. (2016). The neuronal code for number. *Nature Reviews. Neuroscience*, *17*(6), 366–382. http://doi.org/10.1038/nrn.2016.40

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*. http://doi.org/10.1155/2011/156869

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, *10*, 27. http://doi.org/10.3389/fninf.2016.00027

Pantazis, D., Fang, M., Qin, S., Mohsenzadeh, Y., Li, Q., & Cichy, R. M. (2017). Decoding the orientation of contrast edges from MEG evoked and induced responses. *NeuroImage*, 1–31. http://doi.org/10.1016/j.neuroimage.2017.07.022

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://doi.org/10.1007/s13398-014-0173-7.2

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. http://doi.org/10.1016/j.jneumeth.2006.11.017

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555. http://doi.org/10.1016/j.neuron.2004.10.014

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A Magnitude Code Common to Numerosities and Number Symbols in Human Intraparietal Cortex. *Neuron*, *53*(2), 293–305. http://doi.org/10.1016/j.neuron.2006.11.022

Pinhas, M., & Fischer, M. H. (2008). Mental movements without magnitude? A study of spatial biases in symbolic arithmetic. *Cognition*, *109*(3), 408–415.

http://doi.org/10.1016/j.cognition.2008.09.003

Pinheiro-Chagas, P., Daitch, A., Parvizi, J., & Dehaene, S. (2017). Brain mechanisms of arithmetic: a crucial role for ventral temporal cortex. *Under Review*.

Pinheiro-Chagas, P., Dotan, D., Piazza, M., & Dehaene, S. (2017). Finger tracking reveals the covert processing stages of mental arithmetic. *Open Mind: Discoveries in Cognitive Science*, 1–12. http://doi.org/10.1162/OPMI_a_00003

Price, G. R., Yeo, D. J., Wilkey, E. D., & Cutting, L. E. (2016). Prospective relations between resting-state connectivity of parietal subdivisions and arithmetic competence. *Developmental Cognitive Neuroscience*. http://doi.org/10.1016/j.dcn.2017.02.006

Rajalingham, R., Schmidt, K., & DiCarlo, J. J. (2015). Comparison of Object Recognition Behavior in Human and Monkey. *Journal of Neuroscience*, *35*(35), 12127–12136. http://doi.org/10.1523/JNEUROSCI.0573-15.2015

Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, *8*(10), 1391–1400. http://doi.org/10.1038/nn1549

Shum, J., Hermes, D., Foster, B. L., Dastjerdi, M., Rangarajan, V., Winawer, J., … Parvizi, J. (2013). A brain area for visual numerals. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(16), 6709–6715. http://doi.org/10.1523/JNEUROSCI.4558-12.2013

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron*, *91*(3), 694–707. http://doi.org/10.1016/j.neuron.2016.07.006

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315. Retrieved from http://www.sciencedirect.com/science/article/pii/0001691869900559

Stokes, M. G., Wolff, M. J., & Spaak, E. (2015). Decoding Rich Spatial Information with High Temporal Resolution. *Trends in Cognitive Sciences*. http://doi.org/10.1016/j.tics.2015.08.016

Taulu, S., & Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine and*

Biology, *51*(7), 1759–1768. http://doi.org/10.1088/0031-9155/51/7/008

Trübutschek, D., Marti, S., Ojeda, A., King, J.-R., Mi, Y., Tsodyks, M., & Dehaene, S. (2017). A theory of working memory without consciousness or sustained activity. *eLife*, *6*, 1–46. http://doi.org/10.7554/eLife.23871

Uittenhove, K., Thevenot, C., & Barrouillet, P. (2016). Fast automated counting procedures in addition problem solving: When are they used and why are they mistaken for retrieval? *Cognition*, *146*, 289–303. http://doi.org/10.1016/j.cognition.2015.10.008

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–871. http://doi.org/10.1038/nn.4546

Zbrodoff, J. N., & Logan, G. D. (2005). What everyone finds: the problem size effect. In J. I. D. Campbell (Ed.), *The Handbook of Mathematical Cognition* (pp. 331–345). New York: Psychology Press.

8. Appendix A

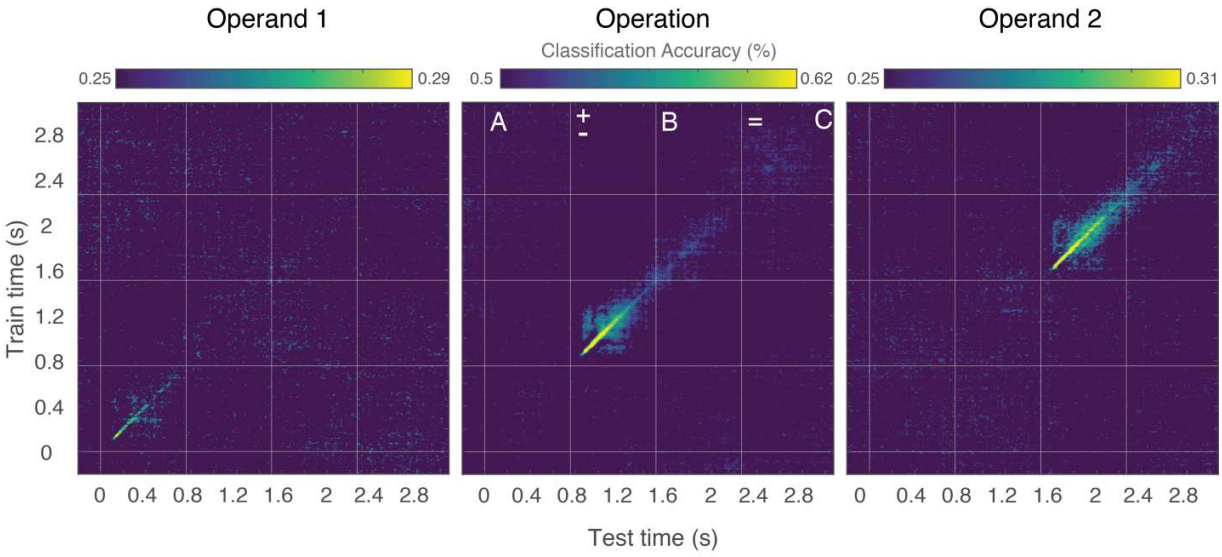


Figure A.1. Generalization across time matrices during calculation
To characterize the dynamics of the mental representations underlying calculation, we tested how the classifiers of the main calculation features generalized in time. Results indicate a succession of dynamical internal codes (diagonal pattern). The plots only show the classification accuracies that were significantly above chance (second-level statistical tests across subjects, with $p < 0.05$, not corrected for multiple comparison).

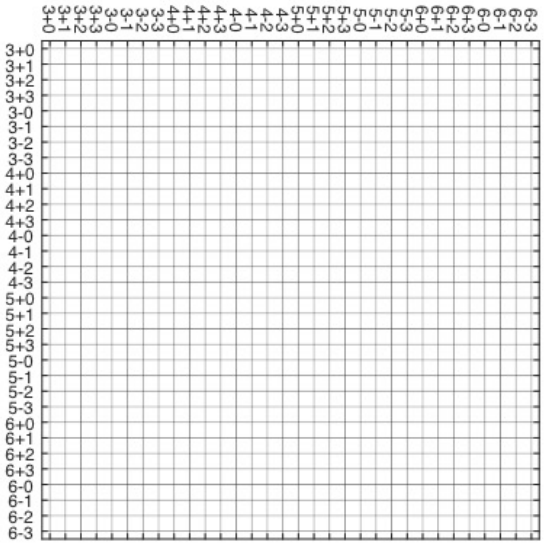


Figure A.2. Matrix structure used for the RSA

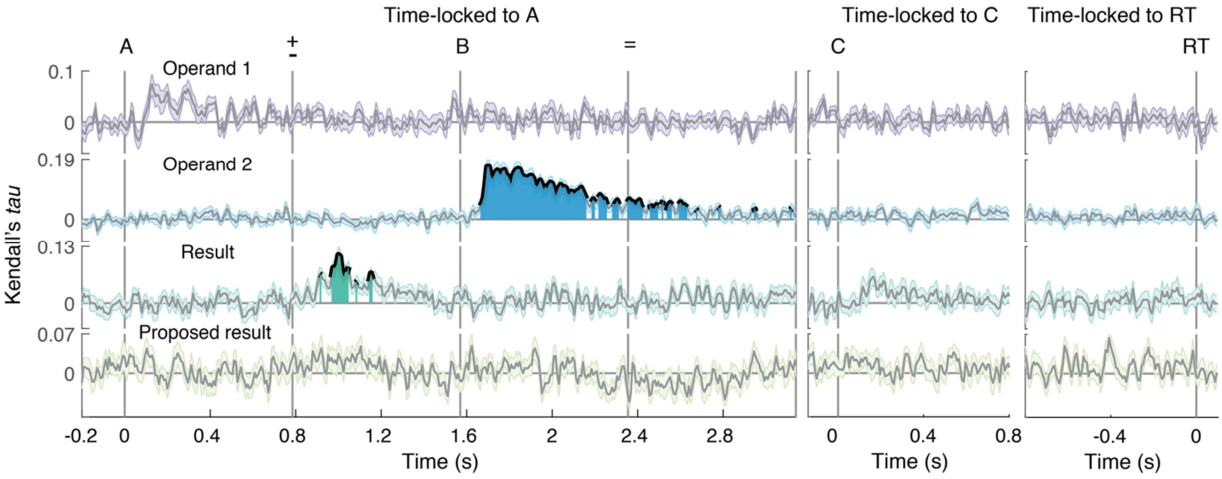


Figure A.3. Decoding the calculation features using regression

Ridge regression estimators were used to decode main calculation features, thus looking for a monotonic relationship between brain signals and the corresponding numbers. Trials were time-locked to three events of interest: onset of the operand 1 (A), onset of the proposed result (C) and RT. Gray horizontal lines indicate theoretical chance level. Thick lines and filled areas represent time periods in which the second-level statistical tests across subjects revealed a Kendall's *tau* significantly above 0 (cluster corrected for multiple comparison, $p < 0.05$).

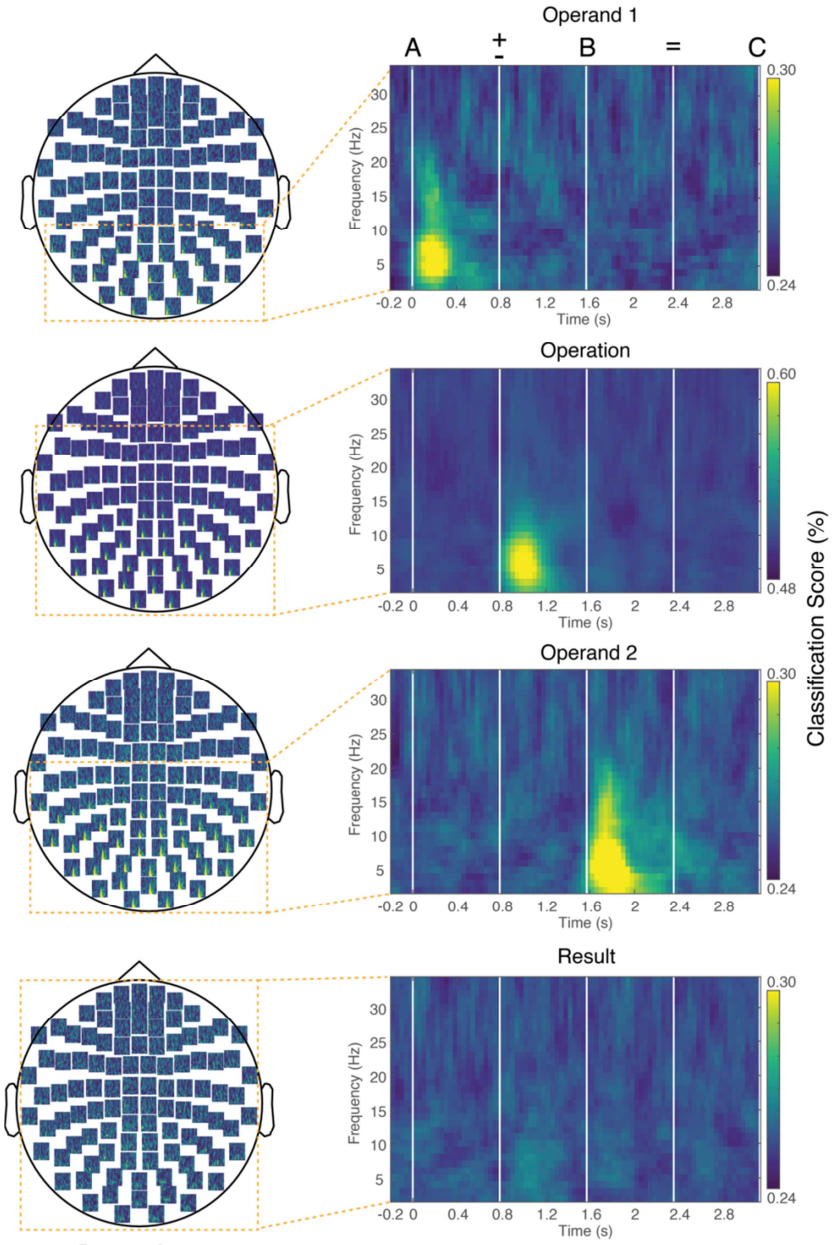


Figure A.4. Searchlight decoding in time, sensor space and frequency
To exhaustively explore the decoding of the main calculation features, in time, sensor space and frequency, we fitted a series of linear discriminant analysis (LDA) in a searchlight approach. The topoplots on the left show the classification accuracy (color axis) in each 'sphere' of 10 neighbor sensors, for each time (x axis) and low frequency samples (2 - 34 Hz) (y axis). Averaged 'spheres' qualitatively chosen after visual inspection are zoomed in the right plots. Operand 1, operation and operand 2 could be generally from occipital-parietal sensors, at a short time window following their respective onsets and mostly between 3 to 20 Hz. For the result, we found no indication of decoding accuracy above chance level in the time window between B and C (when exact calculation is expected to happen).

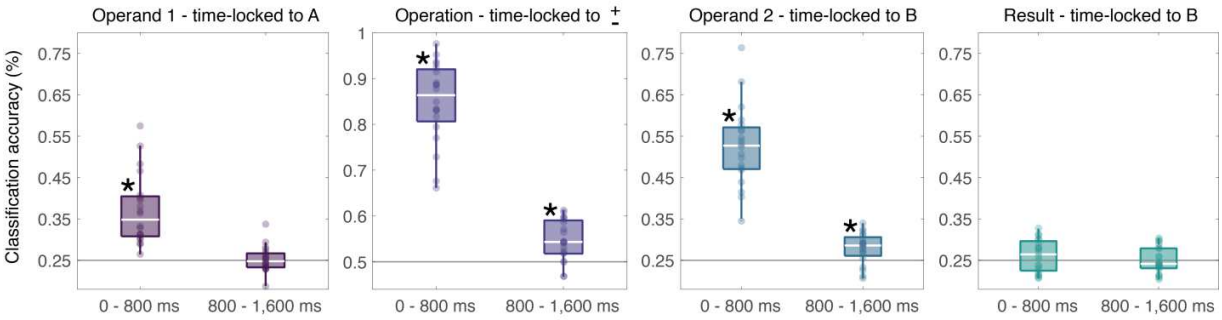


Figure A.5. Decoding calculation features with Riemannian geometry
The ERPCov classifier was used to decode the main calculation features in two time windows (0 – 800 ms and 800 – 1,600). Boxplots represent classification scores across subjects (individual dots). Gray horizontal lines indicate theoretical chance level. (* = $p <$ 0.01, second-level 1-sampled t-test). Classification accuracies were highly boosted as compared to time by time SVM, specially for the operation and operand 2 in the time window of 0 – 800 ms. No significant classification accuracy was found for the internally computed result.