Running Head: PROBLEM SIZE EFFECT

TITLE

Distinguishing between Cognitive Explanations of the Problem Size Effect in Mental Arithmetic via Representational Similarity Analysis of fMRI Data

Kerensa tiberghien[1], Bert De Smedt[2], Wim Fias[1], Ian M. Lyons[3]

[1]Ghent University, Department of Experimental Psychology

[2]University of Leuven, Faculty of Psychology and Educational Sciences

[3]Georgetown University, Psychology Department

Funding Information: This research was supported by Interuniversity Attraction Poles Program of the Belgian Federal Government (P7/11), Ghent University (GOA 01G01108) to Fias, Belgian Fund for Scientific Research - Flanders (project G063817N) to De Smedt and Fias, Banting Postdoctoral Fellowship to Lyons (National Sciences and Engineering Research Council, Canada), Departmental Start-Up Funds to Lyons (Georgetown University, Psychology Department) Travel Grant Faculty Mobility Fund Ghent University to Lyons. The authors declare no financial conflicts of interest.

Word Count (Main Text): 7384

ABSTRACT

Not all researchers interested in human behavior remain convinced that modern neuroimaging techniques have much to contribute to distinguishing between competing cognitive models for explaining human behavior, especially if one removes reverse inference from the table. Here, we took up this challenge in an attempt to distinguish between two competing accounts of the problem size effect (PSE), a robust finding in investigations of mathematical cognition. The PSE occurs when people solve arithmetic problems and indicates that numerically large problems are solved more slowly and erroneously than small problems. Neurocognitive explanations for the PSE can be categorized into representation-based and process-based views. Behavioral and traditional univariate neural measures have struggled to distinguish between these accounts. By contrast, a representational similarity analysis (RSA) approach with fMRI data provides competing hypotheses that can distinguish between accounts without recourse to reverse inference. To that end, our RSA (but not univariate) results provided clear evidence in favor of the representation-based over the process-based account of the PSE in multiplication; for addition, the results were less clear. Post-hoc similarity analysis distinguished still further between competing representation-based theoretical accounts. Namely, data favored the notion that individual multiplication problems are stored as individual memory traces sensitive to input frequency over a strictly magnitude-based account of memory encoding. Together, these results provide an example of how human neuroimaging evidence can directly inform cognitive-level explanations of a common behavioral phenomenon, the problem size effect. More broadly, these data may expand our understanding of calculation and memory systems in general.

Highlights

- The problem-size effect (PSE) is a common and robust behavioral effect in arithmetic
- Univariate fMRI does not but RSA does differentiate cognitive accounts of the PSE
- RSA data show problems are stored as memory traces sensitive to input frequency
- Data were inconsistent with a strictly magnitude-based account of memory encoding
- Human fMRI data can directly inform cognitive explanations of behavioral phenomena

INTRODUCTION

The problem size effect (PSE) is a well-known phenomenon in mental-arithmetic and probably the most studied effect in the history of mathematical cognition (e.g., Ashcraft, 1992; Campbell & Xue, 2001; Zbrodoff & Logan, 2005). Problems with large operands (8 + 9, 6 x 8) take longer to solve and produce more errors than problems with small operands (4 + 2; 3 x 4; Zbrodoff & Logan, 2005). At the neural level, the PSE has been examined primarily using univariate approaches, identifying various regions that show differences in univariate activity as a function of problem size, with increased activity for large compared to small problems in a large network of frontal, parietal and temporal regions (De Smedt et al., 2011; Jost et al., 2011; Grabner et al., 2013; Prado et al., 2013; Menon, 2015, for a review). As it is a robust effect that characterizes much of arithmetic processing, understanding the mechanisms that underpin the PSE can also expand our understanding of how the brain achieves simple arithmetical processing.

Cognitive explanations of the PSE tend to fall broadly into two different categories, which could be labeled as *representation-based* and *processing-based* accounts (Ashcraft & Guillaume, 2009). Behavioral and traditional univariate approaches to fMRI data have struggled to distinguish between these two competing accounts without recourse to reverse inference (Coltheart, 2006; Anderson, 2014). The present study uses human neuroimaging data to distinguish between these two accounts of the PSE via representational similarity analysis (RSA; Kriegeskorte et al., 2008; Davis & Poldrack, 2013). Rather than looking at the difference in neural activation between conditions, RSA allows one to look at the relative similarity in spatial patterns of neural activation *within* conditions. Interestingly and key to the current study, the representation-based account and processing based account make competing predictions on the similarity in spatial patterns of brain activity for small and large problems, as we elaborate in more detail below.

Representation-based accounts focus on how arithmetic facts are stored or *represented* in memory, and on the various factors that impact the representation and retrieval of this memory trace (e.g., Ashcraft, 1987; Siegler & Shrager, 1984; Campbell, 1995). In general, this view predicts that the representations of small problems will be more precise and less overlapping than the representations of large problems. Within the representation-based account, there are various explanations for this difference in representational overlap between small and large problems. One such explanation focuses on the order and frequency with which arithmetic problems are learned and practiced. More frequently taught problems (i.e., small problems) are predicted to have a higher strength of representation and are consequently more distinctive (i.e., less overlapping). This distinctiveness is also characterized by a recent history comprising fewer erroneous retrievals (Siegler & Shrager, 1984),

leading to a more 'peaked distribution of associations' (between stimulus and correct response), which in turn implies greater likelihood of retrieving the correct answer. By contrast, larger problems are learned later and encountered and practiced less frequently. As a result, they are stored in memory less distinctively with greater overlap in memory traces. Large problems will have a more widely spread distribution of associations due to a greater history of errors – i.e., a given problem is linked to a greater number of answers or outputs with a broader or less peaked distribution around the correct answer. Larger problems with a widely-spread distribution of associations will be solved more slowly and more erroneously because the correct answer is likely to achieve retrieval threshold less efficiently (Ashcraft, 1987; Ashcraft & Christy, 1995; McCloskey & Lindemann, 1992).

A second explanation within the representation-based account explains the PSE by recourse to representations of problem magnitude (Campbell, 1995). Specifically, a widespread 'tuning-curves' model of magnitude representation posits that larger quantities are represented less precisely, such that the representations of two large quantities (e.g., 8 and 9) are expected to overlap to a greater extent than two smaller quantities (e.g., 1 and 2), holding distance constant (Piazza et al., 2004; Lyons et al., 2015). In his network interference model of arithmetic facts, Campbell (1995) proposed that greater representational overlap for large problems in turn results in greater interference – and hence worse performance – on large relative to small problems.

Processing-based accounts focus on the various *processes* – typically in the form of different strategies – that are used to solve small and large problems (e.g., LeFevre et al., 1996; Campbell & Xue, 2001). In this account, small problems tend to rely on very similar processing strategies, such as fact retrieval. Large problems, on the other hand, are predicted to be solved by procedural strategies, which consists of a much more diverse set of processing strategies, such as decomposing a problem in a series of other problems [e.g., 8×7 can be decomposed into $(7^2) + 7$], rounding [e.g., $9 \times 7 = (10 \times 7) - 7$], transforming [e.g., $8 \times 5 = 2 \times (4 \times 5)$], and even counting. This account predicts that there should be greater similarity when processing small problems, as they involve more similar solution strategies; larger problems should be less similar due to a more heterogeneous set of processing strategies.

It is important to note that, with respect to the PSE, both representation and process-based accounts make similar predictions for behavior (poorer performance on larger relative to smaller problems) and univariate analysis of neural data (greater activity for large relative to small problems). One might make differing predictions for the different views with regard to observing the PSE in different sets of brain regions; however, it is difficult to see how to do this without obvious (and hence problematic) recourse to reverse inference. Crucially, however, representation-based and processing-based accounts make competing predictions with respect to similarity of patterns of neural responses for small and large problems *even within the same brain region* (thus largely circumventing the issue of reverse inference). Specifically, as reviewed above, *Representation-based* views posit that the

representations of individual numerically small problems (2 + 1, 3 × 4) are relatively distinct from one another, with more narrowly tuned distributions due to less overlapping representations. Consequently, the neural similarity of distributed patters of neural activity among small-problems should be low. Large-problems (8 + 9, 6 × 8) elicit broader, more overlapping distributions, and should thus show relatively high neural similarity with one another. If a given brain area shows greater similarity for *large relative to small problems*, this would provide support for a *representation-based account* of the PSE. By contrast, *Processing-based accounts* of the PSE rely on the notion that small-problems are solved via a narrow range of highly consistent strategies, which should lead to *high* similarity values among small-problems. Large-problems, being solved by a wider range of more variegated strategies should yield lower similarity values with one another. If a given brain area shows greater similarity for *small relative to large problems*, this would provide support for a *processing-based account*. To conclude, RSA allows us to characterize the manner in which each brain area's response corresponds to one PSE account or the other (or neither) based on its own pattern of neural responses, rather than relying on reverse inference (Poldrack, 2011; Davis & Poldrack, 2013). This approach can thus provide clear *elucidation of the neurocognitive mechanisms* underlying the PSE that is difficult to obtain in behavioral or traditional univariate approaches (Coltheart, 2006).

We assessed the PSE in multiplication and addition mental arithmetic. Behavioral data were acquired prior to scanning; fMRI data were acquired in a manner that isolated arithmetic computation from response-selection. Univariate contrasts were used to establish *whether* a given region was sensitive to relative problem-size. RSA was then used to assess *how* the underlying neural patterns were modulated by problem-size. Specifically, for multiplication and addition separately, we assessed the similarity among small-problems and the similarity among large-problems to test the abovementioned competing predictions. To summarize: Regions consistent with a representation-based account should show greater similarity among large- relative to small-problems; regions consistent with a process-based account of the PSE should show greater similarity among small- relative to large-problems.

METHODS

Participants

Thirty adults from Ghent University participated in the experiment (22 female, mean age = 24yrs, range: 18-27yrs, all right-handed). All participants had normal or corrected-to-normal vision and reported no history of neurological or psychiatric illness. Prior to taking part in the study, participants gave written consent and all participants were paid €40 for their participation. The study was approved by the Medical Ethical Committee of Ghent University and Ghent University Hospital.

Due to movement artifacts (4 participants), technical acquisition problems (1 participant) and diagnosis of dyslexia and dyscalculia (1 participant), 6 participants in total were excluded from further analyses, leaving a final sample of 24 participants.

Procedure

All experiments were controlled by E-Prime (Psychology Software Tools, Pittsburg, PA) and displayed on a 1600 x 900-resolution screen. Participants performed an arithmetic task both prior to and during scanning. The computer was placed on average 50 cm in front of the participant for the behavioral task. In the scanner, the experiment was presented using a Brainlogics 200MR digital projector, which was visible via a mirror attached to the head coil, with a viewing distance of 120cm.

Tasks

We should note that the data reported here are part of a larger dataset; all results reported here are unique and address hypotheses that do not overlap with prior publications arising from this dataset (Tiberghien et al., 2019). Both the pre-scan and fMRI arithmetic tasks included three operations: multiplication, addition and subtraction.[1] Operation order was fully randomized across participants, so the presence of subtraction problems should not yield systematic biases when considering just the multiplication and addition problems.

Pre-Scan Arithmetic Task

The pre-scan arithmetic task was a production task (i.e., a task where the participant needed to generate the answer) containing all permutations of two operands ranging from 0 to 10 (121 total problems) with three different arithmetic operations: addition, multiplication and subtraction (resulting in a grand total of 363 problems). All problems were presented once, with order randomized

[1] Subtraction was included for analysis projects not immediately related to the hypotheses being tested here. In particular, here we restrict the focus exclusively on multiplication and addition because it is not always clear whether the PSE in subtraction should consider the solution size or the operand size. This is doubly problematic in the current study, as all three operations were equated in terms of operands, meaning that, for instance, the 'large' operands in 9–8 generate a 'small' solution (1). Moreover, half of the subtraction problems yielded negative solutions. In multiplication and addition, this is not an issue because solutions will never be negative (all operands ≥ 0), and the solutions to problems with large operands will always be larger than problems with small operands.

across participants. A trial started with a fixation (i.e., three squares) presented for 3000msec followed by the arithmetic problem. The problem remained on the screen until the participant responded. Once the participant has said the response out loud, a voice-key recorded the onset of speech. Then, the experimenter recorded the response of the participant and recorded if the voice-key triggered correctly. That is, another sound (e.g., "ehm" or a cough) sporadically triggered the voice-key. If indeed the voice-key was falsely triggered on a certain problem, that problem reappeared later in the experiment. The inter-trial-interval was 1000msec. This procedure was repeated until the participant had solved all 363 problems with a valid registration of voice-key reaction time for each problem. Intermittent breaks were given after every 33 trials.

Arithmetic Task – fMRI Version

The arithmetic task inside the scanner was kept as similar to the pre-scan version as possible: all problems ranging from 0 to 10 were used in addition, multiplication and subtraction, resulting in a grand total of 363 problems. Trials were divided as evenly as possible across 6 separate runs, with trial (and hence also run) order randomized across participants. As in the pre-scan task, a trial started with a fixation presented for 3000msec followed by the first arithmetic problem. Here, the problem (e.g., 9 × 8) remained on the screen for 2600msec. Our goal was to separate the mental calculation aspects of arithmetic processing from response preparation and execution. Participants were instructed to mentally compute the answer during this period. For most trials, the problem was then replaced by fixation. For 10 percent of trials, a response was required, in which case the problem was replaced by two response possibilities for 1500msec. One number was the correct response, while the other number was the correct response +/-1. Participants pressed either a left or a right key (with left or right index finger) to indicate which was the correct answer. This was followed by fixation. Inter-trial interval was jittered (range = 1000 – 8194msec, mean = 3421msec) for all trials. Response events were modeled as events of no interest.

Accuracy for problems of the scanner-task was high (Multiplication: *M* = 93.9%, *SE* = 1.5%; Addition: *M* = 95.0%, *SE* = 1.2%) and comparable to the pre-scan task outside the scanner (Multiplication: *M* = 93.9%, *SE* = 0.8%; Addition: *M* = 98.5%, *SE* = 0.2%). Average response-times were in fact faster than those seen for the pre-scan behavioral task (Multiplication-pre-scan: *M* = 1150msec, *SE* = 61msec; Multiplication-in-scan: *M* = 681msec, *SE* = 19msec; Addition-pre-scan: *M* = 908msec, *SE* = 33msec; Addition-in-scan: *M* = 680msec, *SE* = 20msec), which is what one would expect if participants were computing the answer during presentation of the problem (2600msec) prior to appearance of the (occasional) verification probe. Further evidence that participants were engaging with the task as instructed is the presence of problem-size effects in the neural data (see Results). That is, because fMRI analyses focused exclusively on the calculation period (prior to response), it is difficult to explain

how a problem-size effect could have been observed during this period if participants were simply ignoring problem presentation and waiting until the sporadic presentation of response options to engage with the task.

fMRI Data Acquisition and Preprocessing

Images were collected with a 3T Siemens Magnetom Trio MRI system (Siemens Medical Systems, Erlangen, Germany) using a 32-channel radio frequency head coil. Participants were positioned headfirst and supine in the magnet bore. Subjects were instructed to move their heads as little as possible throughout the entire scanning session. A whole-brain high-resolution anatomical scan was acquired using a standard 3D MPRAGE sequence (voxel size = $1mm^3$). Functional images were collected using an echo-planar imaging (EPI) sequence: TR = 2600msec, TE = 28msec, flip angle = 80°. In-plane resolution was $3.3mm^2$ in a 64 × 64 matrix; slice thickness was 3.3 mm (44 slices, ascending-interleaved acquisition, no skip between slices), yielded a net field of view of 211.2 × 211.2 × 145.2mm, comprised of $3.3mm^3$ isometric voxels.

Structural and functional images were analyzed using Brain Voyager QX 20.4 (Brain Innovation, Maastricht, Holland). Functional data were interpolated to $3mm^3$ in size. Next, they were corrected for slice scan-timing using cubic spline interpolation, followed by 3D motion-correction (trilinear/sinc interpolation), and then high-pass filtered using a GLM procedure with a Fourier basis set. Excessive motion was deemed net drift > 3mm in a given run or > 1.5mm sudden movement; participants with runs exceeding these criteria were removed from analysis (n = 4). Participants' functional images were then co-registered to their respective anatomical scans using 12-parameter gradient-based affine alignment, and anatomical images were co-registered into Talairach space (Talairach & Tournoux, 1988). For univariate analyses, functional data were spatially smoothed at 3mm FWHM; for multivariate (RSA) analyses, unsmoothed data were used in order not to contaminate the patterns of activation across voxels. Multivariate analyses were conducted using Matlab (2016a). Our a priori whole-brain univariate statistical threshold was an uncorrected voxel-wise threshold of $p < .001$, subsequently cluster-corrected for multiple comparisons using a Monte Carlo simulation procedure (Forman et al., 1995) at $\alpha < .01$.

For both the behavioral and univariate imaging analyses, to equate the number of trials in each bin, we defined small problems as having *both* operands smaller in the range 1-4 and large problems having both operands in the range 6-9 (in keeping with previous studies; e.g. Campbell & Graham, 1985; Prado et al., 2011). Dichotomizing problem size in this manner allowed us to characterize behavioral/neural responses first *within* each category before contrasting *between* categories. This also aligned better with the RSA approach (described below).

RSA Analysis

Region of Interest (ROI) Selection

Our aim was to assess *in what manner* the underlying neural patterns were modulated by problem size, thereby potentially allowing us to distinguish between representation- and process-based accounts of the PSE. A reasonable precondition for distinguishing between accounts would be to establish *whether* a given region demonstrates sensitivity to relative problem size in the first place – i.e., via a standard univariate contrast.

A major advantage of this approach is that it significantly reduces the need for reverse inference in interpreting the RSA results in regions known (in this dataset) to be sensitive to the primary effect of theoretical interest: the PSE. By analogy, it would seem odd to examine the neural underpinnings of a behavioral 'effect' that was not actually present in the behavioral data, which is a distinct possibility when using a searchlight approach to RSA. To see this, one can imagine a searchlight analysis that yielded several regions showing a hypothesized similarity result. However, if only a subset of regions show a univariate PSE, how, then, should one interpret the regions showing a significant RSA PSE but not a univariate PSE? A standard response is to rely (implicitly or explicitly) on reverse inference, but we do not condone that practice as a primary method of drawing meaningful conclusions about cognitive theory. Because our goal was to use neural data to distinguish between different theoretical views, here we elected to restrict RSA for a given operation to regions showing a significant PSE (in either direction) for that operation via a whole-brain univariate contrast. This also allowed us to situate our univariate results with respect to those already published.

Furthermore, beyond the clear theoretical motivation outlined above, it is also important to point out that our univariate-to-multivariate approach does *not* constitute 'double-dipping'. First, from a theoretical standpoint, the questions being asked by the two analyses are distinct: the univariate contrast assesses *whether* a region is sensitive to relative problem size; the RSA analyses assess *in what manner* this sensitivity manifests. Second, this approach is not statistically biased. The ROIs are being identified via a within-subjects contrast *between* large (averaged together) and small (averaged together) problems. Because a within-subjects contrast takes into account the correlation between conditions, it is the case that one would expect to see inflated correlations *between* small and large problems. However, there is no reason to assume inflated correlations *within* problems of a given condition – i.e., small~small and large~large correlations. Most crucially, there is still less (statistical) reason to expect that the correlations within one condition should systematically differ from those within the other condition – i.e., that small~small correlations should be greater than large~large correlations, or vice versa. In this way, the ROI selection here is both theoretically driven and, crucially, *not* a form of 'double-dipping'.

RSA Model

An important prerequisite for this set of analyses is to assess similarity between problems within a given category (e.g., within small addition problems). This allows one to compute PSEs (large-addition similarity versus small-addition similarity). To assess similarity within a given category, one could compute the similarity between each pair of problems in that category (1 + 1 ~ 1 + 2, 1 + 2 ~ 1 + 3, etc.) and then average over these similarity value estimates (*r*-values). Each stimulus (problem) was presented only once in the current dataset, so to analyze each stimulus separately would require us to rely on activity estimates based on a single event. Similarity computations between individual events were therefore likely to carry a substantial amount of noise. Thus, we adopted a different approach based on the 'operand families' method developed by LeFevre and Morris (1999) that involves binning responses according to a specific operand value (similar to 'times tables'). Within each operation, we created a predictor that included all problems/events involving all that operand (e.g., 'Addition-9', or 'A09': 9 + 0, 9 + 1, 9 + 2… 9 + 10). We then used this predictor to extract average intensity values (for a given subject) that included all 21 addition events which involved a 9 in any fashion. This approach significantly increases the precision of our activity estimates for each operand as it now comprises 21 events instead of just one event. The end result, for each operation, was 11 activity estimates corresponding to the operands 0-10 (which we write, for addition, 'A00', 'A01'…'A10'; for multiplication, 'M00', 'M01'…'M10'). In this way, we generated activity estimates in each voxel for each predictor. In a given ROI, we extracted and vectorized the distributed activity pattern across functional voxels for each predictor. We then correlated these vectors (e.g., 'M00', 'M01'…'M10') with one another to generate an 11×11 correlation matrix. For present purposes, we were interested in PSEs, and to keep these analyses consistent with the behavioral and univariate analyses above, we then averaged similarity estimates between operands within each size category (small: 1~2, 1~3…3~4; large: 6~7, 6~8…8~9; we took only the lower triangle in each case given matrix symmetricity over the main diagonal). The above process was repeated for each subject separately. In this way, for each subject, we extracted within-category similarity estimates for each operation (addition, multiplication) at each size (small, large). Statistical tests were performed by comparing Fisher-*z*-transformed *r*-values (large – small) across 24 subjects.

One potential downside of this approach is that, for within operation similarity estimates, a pair of predictors/operands shared 2 of their 21 respective events. For example, the 'A01' and 'A02' both contain '2 + 1' and '1 + 2'. This will of course modestly inflate expected correlation (similarity) values above an expected mean value of 0. First, this inflation should impact all similarity values equally because all pairs of predictors/operands shared exactly 2 events; thus comparing similarity values against one another should not be biased by this inflation factor. On the other hand, comparing similarity values against 0 (i.e., to test if similarity in a given category differs from chance), will indeed

be biased. To correct for this, we ran 100,000 simulations assuming the same parameters as the current dataset but with randomly generated values instead, and found an average inflation factor of .093 (*r*-values were .093, on average, instead of the expected value of 0). Hence, .093 was subtracted from all within-operation similarity estimates. Complete similarity matrices can be found in Supplementary Information.

Another point is that these operand-based predictors contained information from both small and large problems. For instance, A01 contained information from 1 + 8 and 9 + 1. However, this in fact works against our hypotheses, as it should reduce the likelihood of finding a statistically significant PSE; by making our predictors a priori more similar in terms of size, finding differences in similarity as a function of size is thus made more difficult. Finally, we also demonstrate the validity of this modeling approach by using it to recompute the behavioral results (see Figure 1), which showed very similar PSEs to those found using the more traditional approach described above.

RESULTS

Behavioral Results

Traditional Approach

As expected for such simple arithmetic problems, accuracy on all operations was near ceiling (all problems: $M = 96\%$, $SE < .01\%$; addition: $M = 98\%$, $SE < .01\%$; multiplication: $M = 94\%$, $SE = .01\%$). Thus, we focus here instead on reaction times (RTs), which are summarized in Table 1. The problem size effect (PSE) was highly significant for both multiplication (Large–Small: $M = 719.03$msec, $SE = 95.66$; $t(23) = 7.52$, $p < .001$, $d = 1.53$) and addition (Large–Small: $M = 257.39$msec; $SE = 149.91$; $t(23) = 8.41$, $p < .001$, $d = 1.72$).

Table 1: Pre-Scan RTs

	Small	Large
Addition	781 (29)	1038 (45)
Multiplication	830 (30)	1549 (113)

Note: Mean reaction times (msec) and standard errors in parentheses.

Operand-Bin Method

To verify the operand-bin model used for the RSA model (see Methods), we binned behavioral RT data (pre-scan behavioral task) using this method. Figure 1 shows these data below. To calculate PSEs, we computed each individual participant's regression slope with RT as dependent variable and Operand Size (0-10) as the predictor. We then used a one-sample *t*-test on these slopes (betas) across participants ($N = 24$). This procedure was repeated for each operation separately. The results showed that there was a highly significant positive slope for both operations: Multiplication: $t(23) = 8.01$, $p =$

4E-08; Addition: $t(23) = 8.37$, $p = 2E\text{-}08$. Thus, behavioral data analyzed in the same manner as the RSA data yielded PSEs highly comparable to the more traditional means of computing PSEs used in the main text. This indicates the operand-bin model is capable of detecting the influence of problem-size (an assumption critical to the RSA approach adopted here).

Figure 1: RT data using the operand-bin model.

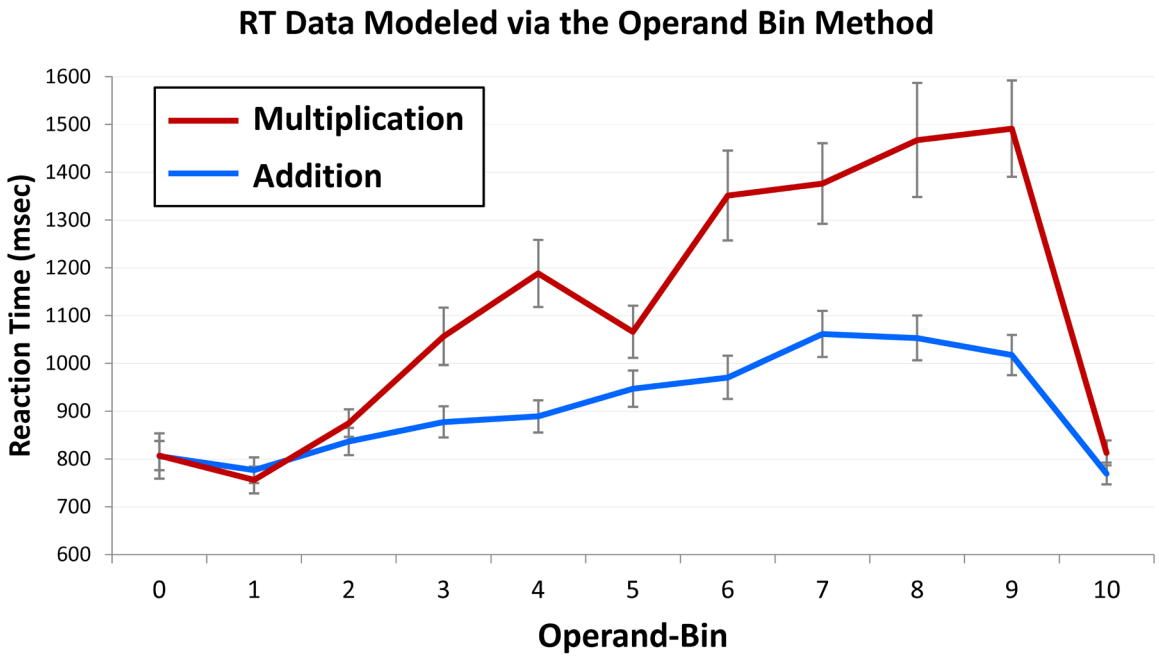


Figure 1 shows RT data modelled via the operand-bin method (identical to that used in the RSA model). Values here are averaged across participants. Error-bars indicate standard-errors.

Univariate fMRI Results

Multiplication

A standard voxel-wise GLM was run with separate predictors for small and large multiplication and addition. Small and large predictors comprised trials in the same manner as the behavioral results (other trials and response events were modeled as events of no interest). Brain areas that showed a PSE were identified for multiplication and addition separately by contrasting Large and Small predictors (agnostic to direction).

For multiplication, significant PSEs (large > small) were identified in a network including bilateral intraparietal sulci (IPS), left ventral temporal-occipital junction (LTOJv), multiple prefrontal regions, and several subcortical regions including dorsal striatum and cerebellum. Regions are shown in Figure 2, and a complete list of regions – along with region details, abbreviations and activity estimates – can be found in Table 2. Note that several very large regions clearly spanned multiple

cortical areas; these were each split into sub-regions using a standard *k*-means clustering algorithm (Lloyd, 1982) based on Talairach coordinates[2]. No regions showed the reverse pattern (small > large).

Table 2: Univariate PSE Region Details – Multiplication

	Talairach Coordinates			Size	Beta Estimates	
Region	x	y	z	(mm^3)	Large	Small
LDLPFC	-42.3	14.2	29.4	4803	.56 (.07)	.00 (.06)
RDLPFC	42.3	8.4	30.4	541	.44 (.07)	-.04 (.07)
LINSa	-31.4	18.4	6.6	4139	.52 (.07)	-.08 (.05)
RINSa	34.1	19.0	4.9	4020	.50 (.07)	-.08 (.05)
LSFSa	-31.9	51.8	19.9	498	.33 (.07)	-.10 (.06)
RSFSa	31.2	46.6	22.8	578	.34 (.07)	-.06 (.06)
LFEF	-28.1	-2.6	55.4	696	.53 (.07)	.03 (.05)
LIPSa	-34.0	-45.9	42.1	3680	.73 (.06)	.16 (.06)
LIPSp	-24.1	-63.5	42.9	4545	.78 (.08)	.17 (.06)
RIPSp	27.7	-56.8	43.2	3017	.53 (.06)	.03 (.06)
LPRC	-6.9	-68.7	41.3	695	.39 (.08)	-.11 (.07)
LTOJv	-46.8	-60.1	-6.4	1071	.58 (.07)	.16 (.06)
ACCd	1.3	19.4	34.2	5548	.50 (.06)	-.04 (.05)
PreSMA	-0.1	8.8	49.5	4155	.73 (.09)	.09 (.05)
PCC	-1.4	-23.4	27.0	924	.38 (.05)	-.08 (.07)
LSTRId	-15.2	5.3	7.0	2305	.47 (.05)	.01 (.07)
RSTRId	15.0	7.0	8.7	1555	.50 (.06)	.05 (.07)
LTHALdm	-10.7	-6.1	14.7	1841	.38 (.05)	-.07 (.06)
RTHALdm	9.4	-6.5	10.2	802	.34 (.05)	-.08 (.06)
RCBMd	30.6	-59.9	-20.9	1530	.52 (.06)	.05 (.06)
RCBMv	29.4	-62.0	-38.9	787	.47 (.07)	-.06 (.05)
MCBMp	2.99	-73.8	-18.9	2228	.34 (.06)	-.09 (.05)

Note: Region Abbreviations: leading L=Left, leading R=right, leading M=middle; a=anterior, p=posterior, d=dorsal, v=ventral, m=medial; DLPFC=dorsolateral prefrontal cortex, INS=insula, SFS=superior frontal sulcus, FEF=frontal eye field, IPS=intraparietal sulcus, PRC=precuneus, TOJ=temporal-occipital junction, ACC=anterior cingulate cortex, PreSMA=pre-supplementary motor area, PCC=posterior cingulate cortex, STRI=striatum, THAL=thalamus, CBM=cerebellum.

[2] Regions split in this way were as follows: a large frontal midline region was split into pre-supplementary motor area (PreSMA) and dorsal anterior cingulate cortex (ACCd); a large region running the length of the left IPS was split into anterior (LIPSa) and posterior (LIPSp) regions; large bilateral subcortical regions were each split into dorsal striatum (STRId) and dorso-medial thalamus (THALdm).

Figure 2: Univariate PSE Regions – Multiplication

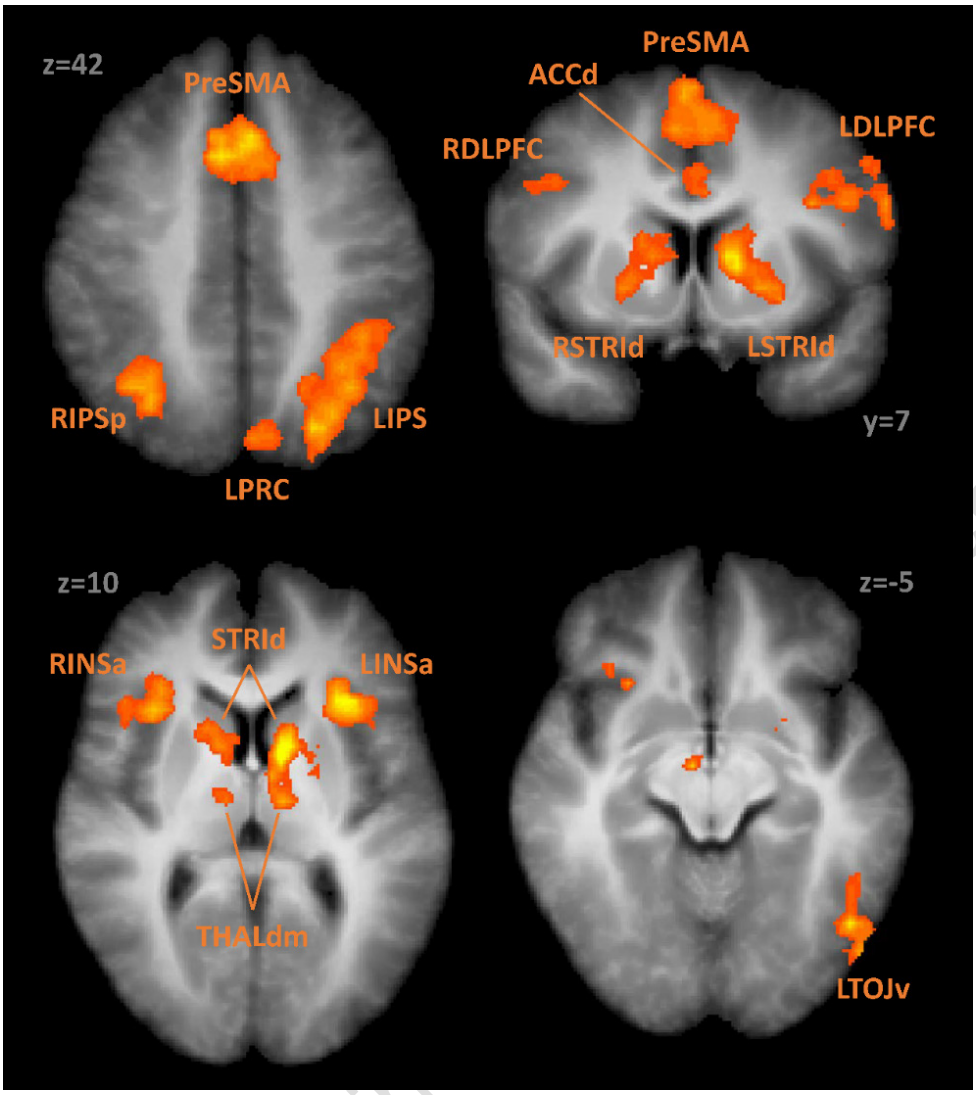


Figure 2 shows whole-brain univariate PSE Regions for Multiplication. See Table 2 for region abbreviations. All regions: Large > Small.

Addition

For addition, no regions showed a significant PSE using the more conservative a priori whole-brain threshold (voxelwise $p < .001$, cluster-corrected at $\alpha < .01$). However, several regions were significant at the slightly more liberal threshold of voxelwise $p < .005$, cluster-corrected at $\alpha < .01$. Because (1) our primary focus was not on the univariate but the RSA results, and (2) this slightly more liberal threshold is still not unreasonable in the field (Cunningham & Koscik, 2017; Slotnick, 2017), we deemed it advisable to avoid a potentially obvious Type II error and so proceeded with this lower threshold for these addition problems. Still, caution may be warranted when interpreting the results for addition. Several regions showed a standard PSE (large > small) for addition, including PreSMA and bilateral IPS and LTOJv. Note that each of these regions overlapped with similar regions seen for multiplication. Two regions showed a reversal of the PSE (small > large): left AGd and RIFGa. Regions are shown in

Figure 3, and a complete list of regions – along with region details, abbreviations and activity estimates – can be found in Table 3.

Figure 3: Univariate PSE Regions – Addition

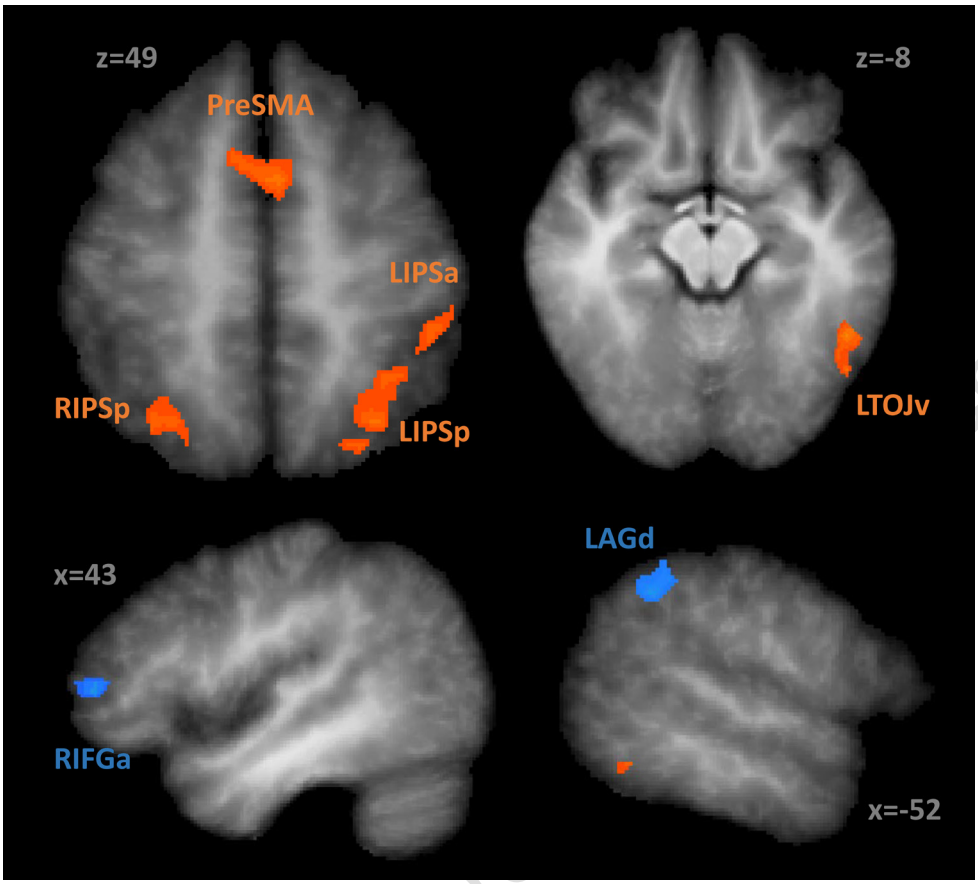


Figure 3 shows whole-brain univariate PSE Regions for Addition. See Table 3 for region abbreviations. Orange: large>small; blue: small>large.

Table 3: Univariate PSE Region Details – Addition

	Talairach Coordinates			Size	Beta Estimates	
Region	x	y	z	(mm^3)	Large	Small
LIPSa	-42.8	-34.2	46.1	572	.50 (.07)	.11 (.06)
LIPSp	-28.3	-54.0	49.0	1034	.49 (.07)	.12 (.06)
RIPSp	26.2	-56.7	49.0	686	.39 (.07)	.01 (.06)
LTOJv	-47.3	-53.9	-8.1	401	.38 (.08)	.02 (.05)
PreSMA	1.4	6.5	49.2	645	.43 (.06)	.08 (.06)
LAGd	-52.3	-42.0	38.7	493	-.24 (.05)	.08 (.05)
RIFGa	44.7	42.7	7.9	403	-.30 (.06)	.06 (.05)

Note: Region Abbreviations: leading L=Left, leading R=right; a=anterior, p=posterior, d=dorsal, v=ventral; IPS=intraparietal sulcus, TOJ=temporal-occipital junction, PreSMA=pre-supplementary motor area, AG=angular gyrus, IFG=inferior frontal gyrus.

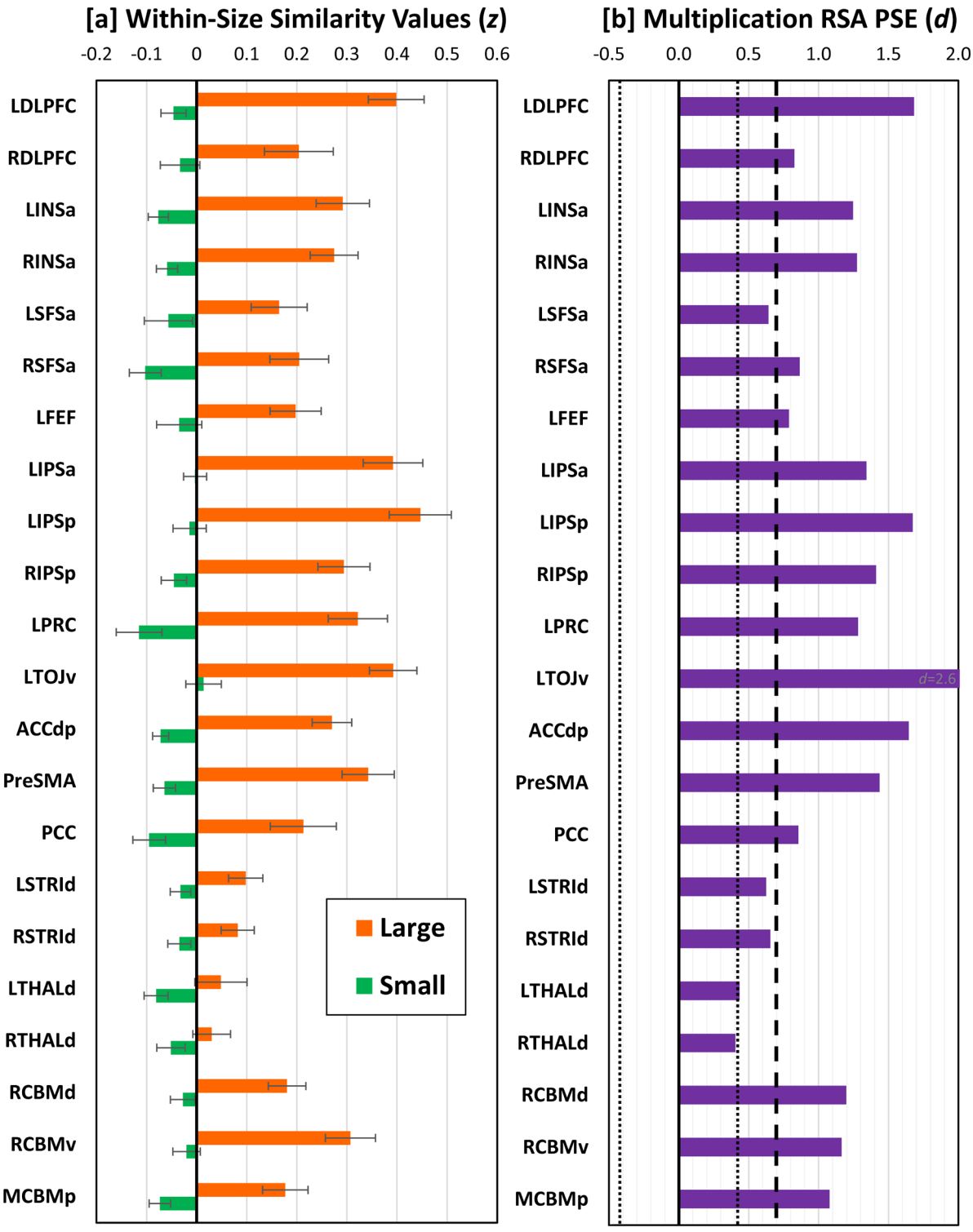
Representational Similarity Analysis (RSA) Results

In this section, we assessed similarity within each small and large condition (separately, for multiplication and addition). By contrasting the relative similarity among large problems with the similarity between small problems, we thus tested whether the regions showing a significant univariate PSE exhibited similarity patterns more consistent with a representation- or process-based account of the PSE. RSA for a given operation was conducted in the corresponding regions from the previous section (i.e., those in Table 2 for multiplication and those in Table 3 for addition). Because *r*-values are bounded between -1 and 1 and are not normally distributed, all statistical tests are computed over Fisher-*z*-transformed *r*-values [*z*=atanh(*r*), values reported in Figures 4a and 5a are thus *z*-values as well]. Finally, here we adopt a more network-based view in that arithmetic is best characterized by a network of regions working in concert (Menon, 2015). Furthermore, the primary goal of the present paper was to use neuroimaging data to distinguish between two classes of theoretical explanations formulated at the cognitive level for a behavioral and neural (univariate) phenomenon. For these reasons, and to protect against reverse inference, we focus on overall patterns of results as opposed to interpreting each ROI in isolation.

Figure 4a shows RSA results for *multiplication*. In all 22 regions, large~large similarity tended to be higher than small~small similarity. Figure 4b shows results from directly contrasting large and small similarity values (shown as effect-sizes). 21 of 22 regions showed a significant large > small similarity-based PSE at $p < .05$ (higher than the dotted line in Figure 4b), and 17 of 22 regions showed an effect that was significant at the more stringent threshold of $p < .0023$ (higher than the bold dashed line in Figure 4b; this *p*-value was determined by correcting for 22 comparisons via the Dunn- Šidák method; Šidák, 1967). The probability of 21 of 22 regions obtaining significance at $p < .05$ by chance is roughly 1E-26; the probability of 17 of 22 regions obtaining significance at $p < .0023$ by chance is roughly 4E-41. We thus consider the evidence in favor of the representation-based account of the PSE in *multiplication* to be strong.

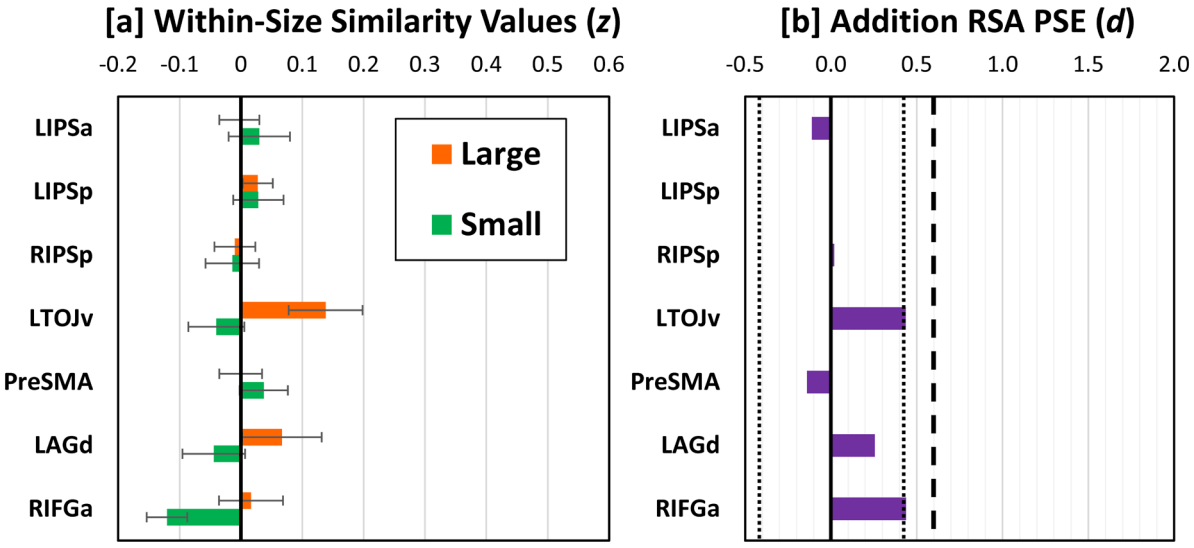
Figure 5a shows RSA results for *addition*. Results for addition were far less conclusive than for multiplication. In just 2 of the 7 regions was large~large similarity significantly higher than small~small similarity. Further, Figure 5b shows that these two regions, LTOJv and RIFGa, showed a similarity-based PSE that only passed only the more liberal threshold of $p < .05$. The probability of 2 of 7 regions obtaining significance at $p < .05$ by chance is .041, which, while significant, is perhaps underwhelming. Moreover, one of the two regions (RIFGa) in fact showed a reversed PSE in the univariate results, which somewhat clouds the interpretation of the PSE in the RSA results. Furthermore, it is worth remembering that the univariate addition PSE was revealed only once we lowered the threshold. We thus consider the evidence in favor of the representation-based account of the PSE in *addition* to be somewhere between marginal and unconvincing.

Figure 4: Similarity-Based PSE – Multiplication



Caption: Figure 4a shows *multiplication* large~large and small~small similarity values (Fisher-z values). Figure 4b shows the similarity-based PSE (large – small) expressed as effect-sizes (*d*). Dotted line: p = .05; bold dashed line: p = .0023 (correcting for 22 comparisons).

Figure 5: Similarity-based PSE – Addition



Caption: Figure 5a shows *addition* large~large and small~small similarity values (Fisher-z values). Figure 5b shows the similarity-based PSE (large – small) expressed as effect-sizes (*d*). Dotted line: $p = .05$; bold dashed line: $p = .0023$ (correcting for 22 comparisons).

Post-hoc RSA

Our main conclusion from the previous section was that the PSE in multiplication is best explained by a representation-based account, wherein smaller problems are represented more distinctly (lower similarity) relative to larger problems (higher similarity). That said, even within the representation-based view of the PSE, there is not always agreement about the ultimate source of the PSE, as one could further distinguish between *memory*-based and *magnitude*-based versions of representation-centered accounts of the PSE. Memory-based accounts place explanatory emphasis on the frequency of input and, relatedly, the order in which the input is learned; in other words, the quality of retrieval association between stimulus and response. Smaller problems are encountered more frequently and learned earlier, and so the quality (in the form of narrower or more 'peaked' distributions) of associations is higher relative to large problems, which are encountered less frequently and learned later, thus incurring lower quality (broader and more overlapping) associations (e.g., Ashcraft, 1987; Ashcraft & Christy, 1995; De Visscher & Noël, 2014; McCloskey & Lindemann, 1992; Siegler & Shrager, 1984). Magnitude-based accounts place explanatory emphasis on the magnitudes of the quantities in question: the neural encoding of larger quantities occurs via neural tuning curves that are wider and thus less finely tuned than the neural encoding of smaller quantities, which have more exact (narrower and less overlapping) neural tuning curves (e.g., Nieder & Dehaene, 2009; Campbell, 1995). Note that both accounts are representation-based (one focusing on the strength and specificity of the memory-association-trace and the other focusing on numerical magnitude), and both predict the similarity-based PSE observed for multiplication in the previous section. Knowing the results of the previous

section, is there a post-hoc test capable of distinguishing between memory- and magnitude-based accounts?

To answer this question, it is informative to look at the multiplication problems involving the operand 10 ('ten-problems'). In the behavioral data using the operand-bin model (see Figure 1), multiplication shows a steep slope (the larger the operand, the larger the reaction time); however, the slope drops quite sharply at the operand-bin '10'. Thus, these 'ten-problems' are performed better than any other of the large operand-bins (ranging from 6 to 9). Multiplication problems including a '10' (e.g., 4 x 10) are often solved via direct memory retrieval. Hence, these problems tend to be performed faster and less erroneously than problems with other operands (e.g., Hinault, tiberghien, & Lemaire, 2015; Lemaire & Reder, 1999; Siegler, 1988). Given that these problems have exceptional patterns of behavioral performance relative to other problems, the question is thus how ten-problems manifest at the neural level.

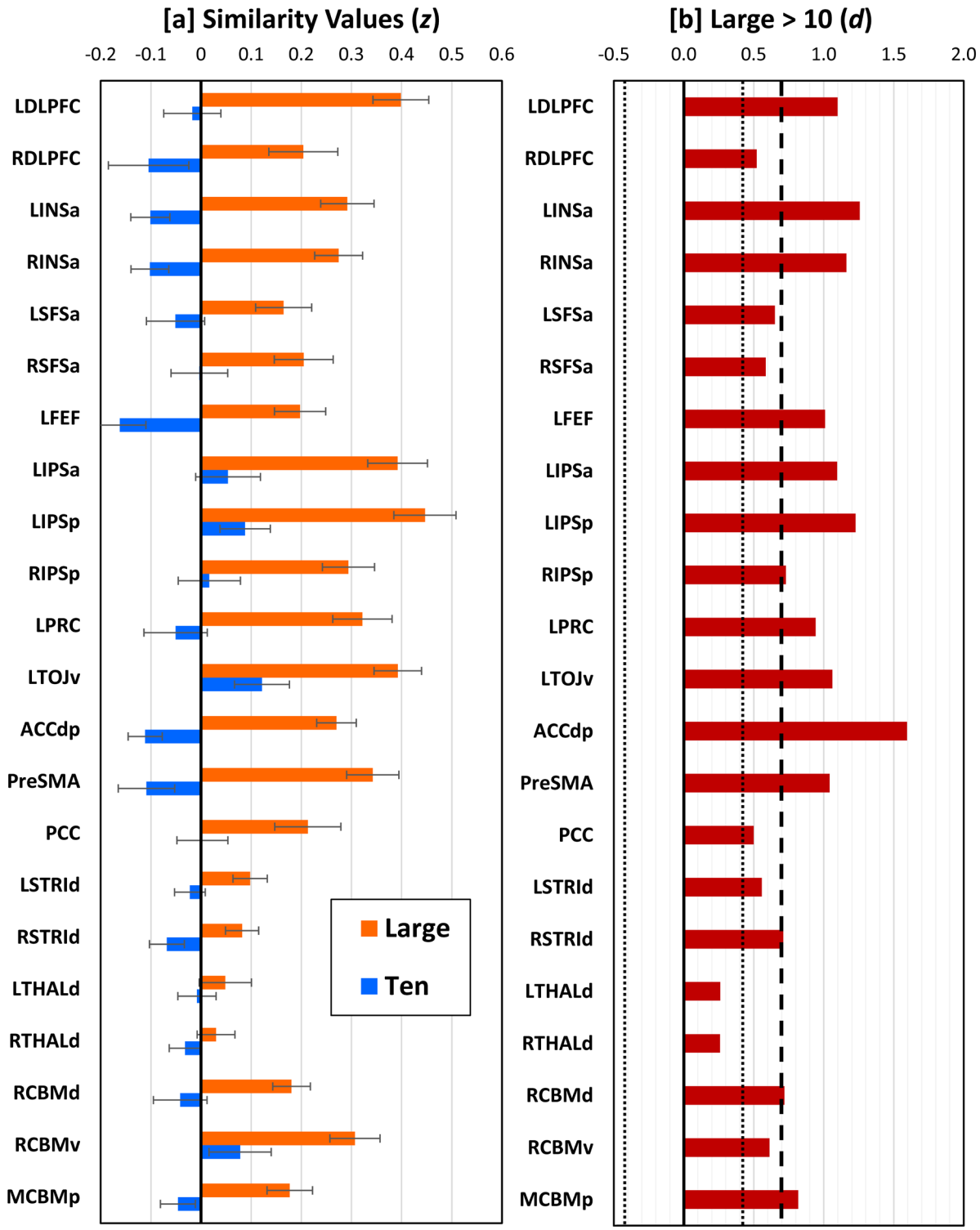
More specifically, if a given problem's representation is purely based on the underlying problem *magnitude*, then the similarity between ten-problems and the problems we previously classified as 'large' (6-9) should be just as high if not higher than that seen for large~large similarity values computed in the previous section. In contrast, a memory-based representation account would emphasize repeated practice with operand-10 problems that leads to highly distinct memory traces for these problems. Akin to what was observed for small problems above, a memory-based view would predict *lower* similarity between ten-problems and 'large' problems relative to large~large similarity values.

To test this explicitly, as in the previous section, large problems were an average of the correlations of problems with operands 6 – 9 within a participant. Ten-problems were an average of the correlations between operand-10 problems and the other large operands (operand-6 through operand-9), again computed within a participant. Average similarity values (*z*-transformed *r*-values) for 'Large' and 'Ten' problems are shown in Figure 6a (orange and blue bars, respectively). Next, we directly contrasted these values using a paired sample *t*-tests (Large – Ten); results are shown as effect-sizes in Figure 6b.

From Figure 6a, it is evident that the similarity between ten-problems and large-problems tended to be relatively low – in fact lower than large-large similarity values in all 22 regions. Looking at the direct contrast in Figure 6b, 20 of 22 regions showed a significant large > ten difference (higher than the dotted line in Figure 6b), and 18 of 22 regions showed an effect that was significant at the more stringent threshold of $p < .0023$ (higher than the bold dashed line in Figure 6b). The probability of 20 of 22 regions obtaining significance at $p < .05$ by chance is roughly 2E-24; the probability of 18 of 22 regions obtaining significance at $p < .0023$ by chance is roughly 2E-44. We thus consider the

evidence in favor of the *memory-based* (as opposed to the magnitude-based) representational account of the PSE in multiplication to be strong.

Figure 6: Assessing 'Ten Problems'



Caption: Figure 6a shows multiplication large~large and ten~large similarity values (Fisher-z values). Figure 6b shows the contrast (large – ten) expressed as effect-sizes (*d*). Dotted line: $p = .05$; bold dashed line: $p = .0023$ (correcting for 22 comparisons).

DISCUSSION

The problem size effect (PSE) in arithmetic is one of the most robust effects in mathematical cognition (e.g., Ashcraft, 1992; Zbrodoff & Logan, 2005) where arithmetic performance tends to be better for numerically small- relative to large-problems, and neural activity estimated in a univariate fashion tends to be higher for large relative to small-problems (Menon, 2015). The PSE is generally explained through two different accounts: a representation-based and a process-based account. By means of representational similarity analyses, we sought to disentangle the two accounts.

One type of explanation tends to focus on how a given item is stored or *represented* in memory, and on the various factors that impact this memory trace. One such explanation highlights the frequency with which arithmetic problems are taught in school. More frequently taught problems (i.e., small problems) are predicted to have a higher strength of representation. By contrast, large problems are stored in memory with lower strength, which results in slower and more erroneous responses (Ashcraft, 1987; Ashcraft & Christy, 1995; McCloskey & Lindemann, 1992). In a similar vein, Siegler and Shrager (1984) suggested that problems with a lesser history of error (e.g., smaller problems) will have a more peaked 'distribution of associations', leading to a larger likelihood of retrieving the correct answer. Large problems will have a more widely spread distribution of associations (a given stimulus is linked to a greater number of answers or outputs with a broader or less peaked distribution around the correct answer). Problems with a widely-spread distribution of associations are solved more slowly and more erroneously because the correct answer is likely to achieve retrieval threshold less efficiently. In a computationally similar model, the network interference model by Campbell (1995) also explains the PSE by means of representations, but via representations of magnitude. Large magnitudes have broader distributions which are present when solving arithmetic problems. These broader distributions lead to greater representational overlap which in turn results in greater interference – and hence worse performance – on large relative to small problems.

A second type of explanation for the PSE tends to focus on the various *processes* – typically in the form of different strategies – that are used to solve small and large problems. A prominent example of this type of explanation focuses expressly on strategic variation (LeFevre et al., 1996; Campbell & Xue, 2001). In this view, small problems tend to rely on similar processing strategies (e.g., retrieval), whereas large problems require a more diverse set of processing strategies (e.g., estimation, calculation, transformation, etc.). The reduced efficiency and increased variability of strategies for large problems leads to poorer performance. In a similar vein, Campbell and Xue (2001) argued that there are three strategy-related sources of the PSE: (1) less frequent retrieval use for large relative to

small problems, (2) lower retrieval efficiency (i.e., speed and accuracy) for large relative to small problems and (3) lower procedural efficiency for large relative to small problems.

Behavioral approaches have struggled to distinguish between these two accounts, and, to date, the neural basis of the PSE has been examined more or less exclusively using univariate approaches which tend to show results that more or less mirror the behavioral results (e.g., regions showing differences in activity as a function of problem size (e.g., De Smedt et al., 2011; Jost et al., 2011; Grabner et al., 2013; Prado et al, 2013). Moreover, interpretations of these activity differences as they pertain to specific cognitive explanations underlying the outward effects (behavioral or neural) have largely been driven by reverse inference, which is highly problematic when interpreting fMRI data (Poldrack, 2006, 2011). Thus, it is unclear how such univariate-based approaches might notably shift the needle with respect to distinguishing between representation- and process-based accounts of the PSE.

Using an RSA-based approach, our data clearly provided evidence for a representation-based account (and against a process-based account) of the PSE in multiplication by showing higher similarity among large relative to small problems. According to this view, large problems should have more overlapping (and thus more similar) distributions. Specifically, greater similarity for large problems indicates that large problems may be less distinguished from one another in terms of their respective neural patterns. Further evidence that multiplication processing is characterized by the representation and retrieval of specific memory traces (memory-based account) was seen in that large multiplication problems showed little or no similarity with ten-problems. This is presumably because these problems can be solved by means of retrieval that can override the underlying magnitude (e.g., Hinault, tiberghien, & Lemaire, 2015; Lemaire & Reder, 1999; Masse & Lemaire, 2001; Siegler, 1988). In other words, ten-problems failed to follow the standard problem-size progression one would expect based on magnitude alone, thus essentially serving as the 'exception that proves the rule'. Multiplication problems are stored as individual representations which typically become less distinct as problem-size increases; but some large problems that can be solved via special rules or that occur much more frequently than expected based on their size can break this trend, and in that case, they are represented quite distinctly even with respect to other large multiplication problems.

On a broader scale, we contend that these data provide an example of how modern neuroimaging techniques may contribute to distinguishing between competing cognitive models for explaining human behavior, even after largely removing reverse inference from the table. An interesting upshot of this perspective is that it requires one to think of the neural data as a dependent variable that must stand on its own merit, and not something that is somehow privileged simply because the data are derived from a neural source per se. Here, merit arises because, when analyzed in a particular fashion, the data afford an opportunity to distinguish between competing theoretical

predictions not provided by other dependent variables, such as reaction time data. Demonstrating that this variable (in this case similarity between neural activity patterns) – clearly favored one set of predictions over another in an obvious majority of candidate brain regions (Figure 4) was sufficient to accomplish the primary aim of the study without overt recourse to reverse inference. That said, it is important to acknowledge that some dependent variables can be hierarchical in nature. For instance, functional characterization of specific brain areas highlighted here may be possible in conjunction with existing or future work, and doing so may be of interest especially for future studies designed expressly to test such functionally specific predictions in specific brain areas.

A methodological consideration when interpreting the RSA results for multiplication is the possibility that greater similarity for large problems may be driven simply by the fact that some or all of the regions in question are not involved in processing small problems to begin with. In this view, there is no systematic signal associated with small problems in the regions revealed here, leading to similarity values close to 0, which in turn speciously drives greater apparent similarity for large relative to small problems. There are multiple aspects of the current data that are inconsistent with this view, however. First, several regions (ACCdp, LINSa, LPRC, LTHALd, MCBMp, PCC, PreSMA, RINSa, RSFSa; Figure 4, left) showed significant dissimilarity (negative similarity) for small problems, indicating the presence of systematic signal in the pattern-based data among small problems, and yet none of these areas showed univariate activity significantly different from (above or below) baseline (Table 2). The presence of systematic relations between variables within the small category is thus not dependent upon whether univariate responses differed significantly from baseline. Approaching the question from the opposite direction, several regions did show significant positive univariate activity for small problems (LIPSa, LIPSp, LTOJv; Table 2), and yet none of these showed significant similarity for small problems (Figure 4, left). Finally, for addition, we saw activity substantially different from baseline for large problems in all 7 regions (Table 3), and yet only 1 of the 7 (LTOJv) showed significant similarity among large problems (Figure 5, left). In other words, there are multiple reasons to doubt a close contingency between univariate activity levels and presence of sufficient signal to detect meaningful correlations via a similarity-based approach.

Finally, it is important to acknowledge that RSA results were less clear for addition. This may indicate that addition problems are represented with highly distinct representations and therefore showed no similarity (as in the representation-based account). If these problems are indeed highly distinct, then most of the additions are likely to be solved by retrieval. Indirect evidence favoring this view can also be seen in that addition RTs were closer to the range of small multiplication RTs, and variability (standard errors) in addition RTs was overall relatively low (also more similar to small multiplication problems; see Table 1 and Figure 1). However, this interpretation is admittedly

speculative; hence, at minimum, the current study underscores the need for future studies to take into account arithmetic operation when probing the neural bases of arithmetic in general.

To conclude, not all researchers interested in human behavior remain convinced that neuroscience – in particular modern neuroimaging techniques – have much to contribute to distinguishing between competing cognitive models for explaining human behavior (e.g., Coltheart, 2006), an issue that becomes all the more acute once one takes seriously the notion of removing reverse inference from the table (Anderson, 2014). Here, we took up this challenge in an attempt to distinguish between competing accounts of the problem size effect (PSE) in arithmetic. Our results provide clear evidence in favor of a representation-based over a process-based account of the PSE in multiplication. Post-hoc analysis further distinguished between different accounts within the broader family of representation-based accounts – specifically, results favored a memory-based account of the PSE over a strictly magnitude-based account. For addition, results were not nearly as conclusive, though a cautious interpretation would slightly favor a representation-based account for addition as well. Because behavioral and univariate fMRI analyses can struggle to distinguish between theoretical accounts of the PSE in arithmetic, the fact that we found such clear evidence for a representation-based account of the PSE in multiplication is an example of how investigating neural data can contribute directly to our cognitive interpretation of a well-known behavioral phenomenon. More broadly, this work may prove useful for understanding the origins of atypical mathematical development such as dyscalculia (Butterworth et al., 2011), as difficulties in arithmetic are the crucial feature of children with dyscalculia.

REFERENCES

Anderson ML (2014). *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Ashcraft MH (1987). Children's Knowledge of simple arithmetic. A developmental model and simulation. In C. J. Brainerd, R. Kail, & J. Bisanz (Eds.) *Formal methods in developmental research* (pp. 302-338). New York: Springer-Verlag.

Ashcraft MH (1992). Cognitive arithmetic – a review of data and theory. *Cognition, 44*(1-2), 75-106.

Ashcraft MH, Christy KS (1995). The frequency of arithmetic facts in elementary texts: Addition and multiplication in grades 1-6. *J Res Maths Edu*, 396-421.

Ashcraft MH, Guillaume MM (2009). Mathematical cognition and the problem size effect. *Psychol Learn Motiv*, *51*, 121-151.

Berteletti I, Man G, Booth JR (2015). How number line estimation skills relate to neural activations in single digit subtraction problems. *NeuroImage*, 107:198-206.

Butterworth B, Varma S, Laurillard D (2011). Dyscalculia: From brain to education. *Science, 332*(6033), 1049-1053.

Campbell JID, Graham DJ (1985). Mental multiplication skill: Structure, process and acquisition. *Canadian J Psychol, 39,* 338-366.

Campbell JID (1995). Mechanisms of simple addition and multiplication: A modified network-interference theory and simulation. *Math Cognit, 1*(2), 121-164.

Campbell JID, Xue Q (2001). Cognitive arithmetic across cultures. *J Exp Psychol Gen*, *130*(2), 299-315.

Coltheart M (2006). What has Functional Neuroimaging told us about the Mind (so far)? (Position Paper Presented to the European Cognitive Neuropsychology Workshop, Bressanone, 2005). *Cortex*, 42(3):323-31.

Davis T, Poldrack RA (2013). Measuring neural representations with fMRI: practices and pitfalls. *Ann NY Acad Sci*, *1296*(1), 108-134.

De Smedt B, Holloway ID, Ansari D (2011). Effects of problem size and arithmetic operation on brain activation during calculation in children with varying levels of arithmetical fluency. *NeuroImage*, *57*(3), 771-781.

De Visscher A, Noël MP (2014). The detrimental effect of interference in multiplication facts storing: typical development and individual differences. *J Exp Psychol Gen*, 143(6):2380-400.

Forman SD, Cohen, JD., Fitzgerald, M, Eddy, WF, Mintun, MA, & Noll, DC(1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magnetic Resonance in Medicine*, *33*(5), 636-647.

Grabner RH, Ansari D, Koschutnig K, Reishofer G, Ebner F (2013). The function of the left angular gyrus in mental arithmetic: evidence from the associative confusion effect. *Hum Brain Mapp*, 34(5):1013-24.

Groen GJ, Parkman JM (1972). A chronometric analysis of simple addition. *Psychol Rev*, *79*(4), 329.

Hinault T, Tiberghien K, Lemaire P (2015). Age-related differences in plausibility-checking strategies during arithmetic problem verification tasks. J Gerontol B Psycho Sci Soc Sci, 71(4):613-21.

Jost K, Khader PH, Burke M, Bien S, Rösler F (2011). Frontal and parietal contributions to arithmetic fact retrieval: a parametric analysis of the problem-size effect. *Hum Brain Mapp*, 32(1):51-9.

Kriegeskorte N, Mur M, Bandettini P (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Front Syst Neurosci*, *2*.

LeFevre JA1, Morris J (1999). More on the relation between division and multiplication in simple arithmetic: evidence for mediation of division solutions via multiplication. *Mem Cognit*, 27(5):803-12.

LeFevre JA, Sadesky GS, Bisanz J (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *J Exp Psychol Learn Mem Cognit*, *22*(1), 216.

LeFevre JA, DeStefano D, Penner-Wilger M, Daley KE (2006). Selection of procedures in mental subtraction. *Canad J Exp Psychol*, *60*(3), 209.

Lemaire P, Reder L (1999). What affects strategy selection in arithmetic? The example of parity and five effects on product verification. *Mem Cognit*, 27(2):364-82.

McCloskey M, Lindemann AM (1992). MATHNET: Preliminary results from a distributed model of arithmetic fact retrieval. *Adv Psychol*, *91*, 365-409.

Menon V (2015). Arithmetic in the child and adult brain. In Cohen Kadosh R, Dowker A (Eds.), *The Oxford handbook of numerical cognition* (pp. 502-530). Oxford library of psychology.

Nieder A, Dehaene, S (2009). Representation of number in the brain. *Ann Rev Neurosci*, 32, 185-208.

Poldrack RA (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cognit Sci*, 10(2):59-63.

Poldrack RA (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72(5):692-7.

Prado J, Lu J, Liu L, Dong Q, Zhou X, Booth JR (2013). The neural bases of the multiplication problem-size effect across countries. *Front Hum Neurosci*, 7:189.

Siegler RS, Shrager J (1984). Strategy choices in addition and subtraction: How do children know what to do. *Orig Cognit Skills*, *23*(1), 229-293.

Siegler RS (1988). Strategy choice procedures and the development of multiplication skill. *J Exp Psychol Gen*, 117(3):258.

Šidák Z (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*, 62(318):626-33.

Talairach, J, & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging*. New York: Thieme Medical Publishers Inc.

Tiberghien K, Sahan MI, De Smedt B, Fias W, Lyons IM (2019). Disentangling Neural Sources of Problem Size and Interference Effects in Multiplication. *J Cogn Neurosci*, 31(3):453-467.

Zbrodoff NJ, Logan GD (2005). What everyone finds: The problem-size effect. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 331-346). New York: Psychology Press.