Quantitative modeling demonstrates format-invariant representations of mathematical problems in the brain

Tomoya Nakai[1,2*], Shinji Nishimoto[2,3,4]

1. Lyon Neuroscience Research Center (CRNL), INSERM U1028 - CNRS UMR5292, University of Lyon, Lyon, France

2. Center for Information and Neural Networks, National Institute of Information and Communications Technology, Suita, Japan

3. Graduate School of Frontier Biosciences, Osaka University, Suita, Japan

4. Graduate School of Medicine, Osaka University, Suita, Japan

*Corresponding author (Email: nakai.tomoya@neuro.mimoza.jp)

Full postal address:
Centre Hospitalier Le Vinatier (Bât. 452), 95 Bd Pinel, 69500 Bron, France

Key words: fMRI, mathematics, IPS, encoding, RSA

Abstract

Mathematical thoughts can be described in either symbolic form or natural language. Previous studies have reported that activation overlaps exist for these two types of mathematical problems, but it is unclear whether they are based on similar brain representations. Furthermore, quantitative modeling of mathematical problem solving has yet to be attempted. In the present study, eight subjects were subjected to 3 h of functional magnetic resonance experiments involving math word and math expression problems, and a read word condition without any calculations was used as a control. To evaluate the brain representations of mathematical problems quantitatively, we constructed voxel-wise encoding models. Both intra- and cross-format encoding modeling significantly predicted brain activity in the bilateral intraparietal sulcus (IPS), even after subtraction of the control condition. Representational similarity analysis and principal component analysis revealed that mathematical problems with different formats had similar cortical organization in the bilateral IPS. These findings support the idea that our mathematical thoughts are represented in the brain in a format-invariant manner.

Introduction

Our mathematical thoughts can be presented in different formats. In symbolic form [math expression (ME)], mathematical problems are represented as a combination of digits and operators (e.g., "3 × 2 = 6"); in linguistic form [math word (MW)], the same mathematical problems are described in natural language (e.g., "There are 3 eggs in each box. How many eggs are in 2 boxes?"). Although most mathematical problems are represented in symbolic ME format, the MW format is still "one of the most common materials in school curricula for teaching students to transfer mathematical knowledge into real-world contexts" (1). Studies with a large sample of schoolchildren have found that performance in MW problems is related to later performance in more advanced algebra (2, 3), suggesting that a shared cognitive basis of mathematical problem solving may exist across different formats.

In several neuroimaging studies, the format-independency of mathematical problem solving has been examined using univariate methods (4–7). Sohn *et al.* reported that MW problems induce activity in the left prefrontal cortex, whereas ME problems induce activity in the bilateral parietal cortex (4). Newman *et al.* found evidence for shared activation of MW and ME problems in the bilateral intraparietal sulcus (IPS) and left frontal cortex (7). Zhou *et al.* tested MW problems with different complexities and found activations that included the left inferior frontal gyrus and dorsomedial prefrontal cortex (6). Chang *et al.* found that, compared with sentence comprehension, MW problems induced higher activation in the IPS and dorsolateral prefrontal cortex (5). Collectively, these studies suggest the involvement of the IPS and prefrontal cortex in both MW and ME problem solving. However, it remains unclear whether the two types of mathematical problems share brain representations. Furthermore, quantitative modeling of mathematical problem solving has yet to be attempted.

To address these issues, we used a voxel-wise encoding modeling approach (8). Encoding models predict brain activity in a quantitative manner based on a combination of features extracted from the presented stimuli. Researchers have adopted this approach to examine brain responses to visual (9–12), auditory (13, 14), semantic (15–17), and emotional (18, 19) stimuli, and to assess many complex cognitive functions (20). In addition, the approach allows for quantitative evaluation of the generalizability of models constructed under a certain format to data in another format (16). Moreover, encoding models enable visualization of multivoxel patterns with different categorical information and evaluation of their similarity across different formats.

In the current study, we performed 3-h functional magnetic resonance imaging (fMRI) experiments using eight subjects (**Figure 1A**) who each attempted to solve a series of

mathematical problems in MW condition; (see Methods for detail) and ME condition formats. As a control, we also included a read word (RW) condition, in which the subjects read the same sentences used in the MW condition but did not conduct calculations. Using sparse operator features, we modeled brain activity specific to each mathematical problem (**Figure 1B**). By applying the cross-modal modeling technique developed in our previous study (16), we defined a format invariance (FI) index, which quantifies how much information can be captured in a format-invariant manner in each brain region. In addition, representational similarity analysis (RSA) and principal component analysis (PCA) revealed representational relationships among different mathematical operations. Thus, through quantitative evaluations, this study provides new evidence of brain representations in math problem solving.

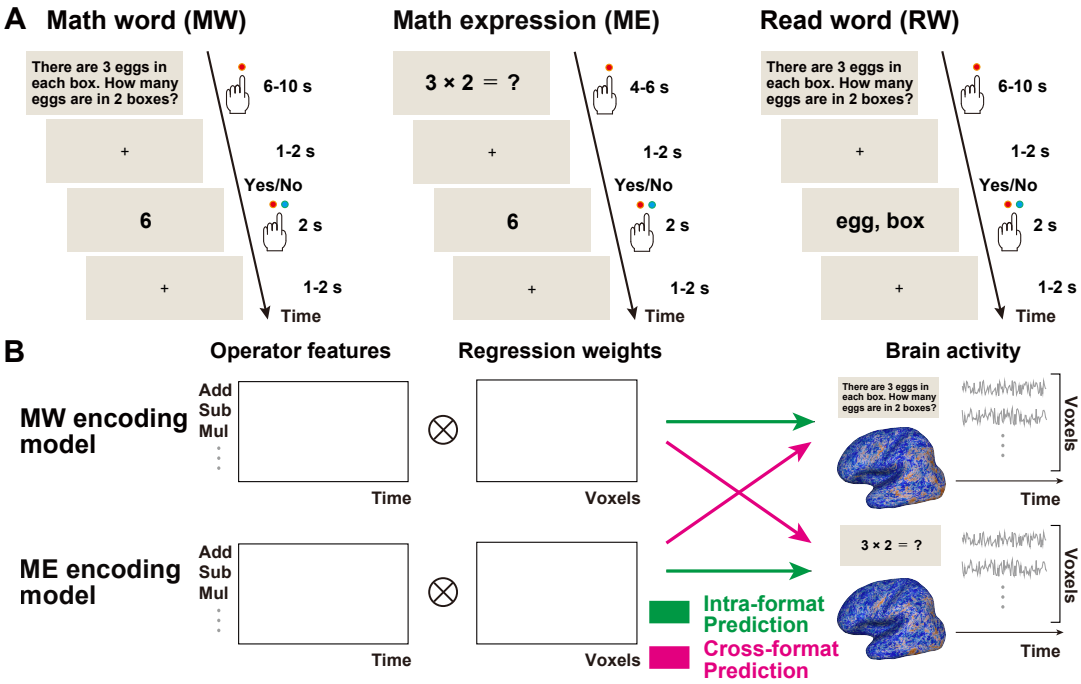

Figure 1. **Experimental design**. (**A**) In the math word (MW) and math expression (ME) conditions, mathematical problems were presented either in a natural language format for 6–10 s or in a symbolic format for 4–6 s. Subjects pressed the button on the left when they had solved a given problem. They pressed the button on the left or right to indicate whether the probe digit stimulus (presented for 2 s) did or did not match their answer, respectively. In the control read word (RW) condition, the same sentences were presented as those presented in the MW condition. The subjects pressed the button on the left when they remembered all nouns that appeared in the given sentence. The probe stimuli consisted of two nouns; the subjects pressed the buttons on the left or right to indicate whether both nouns were or were not included in the

previous sentences, respectively. (**B**) Intra- and cross-format encoding models were constructed to predict brain activity from operator features. The feature matrix and brain response matrix are transposed for the purpose of visualization.

Results

Encoding models predict brain activity in an intra- and a cross-format manner for each subject

To confirm that the encoding models successfully captured brain activity during the math problem solving, we performed a series of intra-format encoding modeling predictions and examined the prediction accuracy using a test dataset with the same format as that of the training dataset (**Figure 1B**, yellow). First, we trained an encoding model using the MW training dataset and predicted brain activity using the MW test dataset. Using operator features, the encoding model significantly predicted the activity in large brain networks, including the bilateral inferior frontal, parietal, and occipital cortices ($P < 0.05$, FDR corrected; 40.1% ± 8.2% of voxels were significant; **Figure 2A** blue). Similarly, we trained encoding models using the ME training dataset and predicted brain activity using the ME test dataset. This encoding model significantly predicted a smaller part of the region, including the bilateral inferior frontal, bilateral parietal, and bilateral occipital cortices, compared with that predicted by the MW intra-format encoding model (21.5% ± 9.4% of voxels were significant; **Figure 2B** blue).

To determine whether the encoding models were generalizable to mathematical problems in other formats, we performed cross-format encoding model analyses. Specifically, we applied an encoding model trained with the MW training dataset to the ME test dataset; prediction accuracy was significant in the bilateral inferior frontal, bilateral parietal, and bilateral occipital cortices (12.2% ± 4.8% of voxels were significant; colored red or white in **Figure 2A**). Similarly, we applied the encoding model trained with the ME training dataset to the MW test dataset; prediction accuracy was also significant in the bilateral inferior frontal, bilateral parietal, and bilateral occipital cortices (21.2% ± 8.2% of voxels were significant; colored red or white in **Figure 2B**). In summary, our encoding models successfully predicted brain activity induced in large brain regions, including the parietal and frontal cortices, across different formats of mathematical problems.

Format-invariant processing of mathematical operations measured in a quantitative manner

To quantify how much information a certain model captures for each format of mathematical

problem, we calculated FI indices by combining the intra-format and cross-format prediction accuracies. FI was adopted based on modality invariance indices, which have been used previously to quantify modality-invariant processing of semantic information across text and speech modalities (16). We found that FI values were significant for the bilateral inferior frontal, bilateral parietal, and bilateral occipital cortices (19.7% ± 9.9% of voxels were significant; **Figure S1**). To evaluate the consistency of our results across test subjects, we calculated the ratio of significant voxels in each anatomical region of interest (ROI) of each subject and averaged across subjects. We focused on ROIs where >20% of voxels had a significant FI value. Suprathreshold FI values were found in large brain regions, including the bilateral prefrontal, parietal, and occipital cortices (**Figure 2C**).

The original encoding models predict not only arithmetic-related regions but also regions related to sensorimotor and semantic information. To exclude regions of noninterest from our analyses, we applied encoding models trained in MW and ME conditions to the control RW condition. The average prediction accuracies of these two models were regarded as baseline accuracies. The baseline prediction accuracy was subtracted from the original intra- and cross-format prediction accuracies, and FI values were calculated. Consequently, we found significant FI values predominantly in the IPS (6.7% ± 4.1% of voxels were significant across the cortex; **Figure 2D**), indicating that format-invariant representations of mathematical problems are robust across different problem structures.

One concern regarding our analyses was whether the encoding model results could be explained solely by the problem complexity factor or also by cognitive load. To examine this further, we first constructed additional encoding models using only single- or double-operators (**Figure S2**); for both operator types, we found significant FI values in the bilateral IPS (10.4% ± 5.5% and 17.2% ± 8.1% of voxels were significant across the cortex, respectively), indicating that model predictability for the bilateral IPS was not a superficial effect due to differences between problem complexities. To further investigate our concern, we performed encoding modeling after excluding cortical voxels, which can be predicted by features of noninterest (i.e., visual, motor, orthography, and cognitive load; see **Supplementary Information** for detail). We calculated intramodal predictions using a leave-one-run-out method within the training dataset. Voxels that were consistently predicted in either the MW or ME condition were excluded from FI analysis; we found significant FI values in the bilateral IPS even after excluding these voxels (15.4% ± 7.2% of voxels were significant; **Figure S3**), indicating that the FI values in the abovementioned analyses were not explained by the effect of sensorimotor processing or general cognitive load.

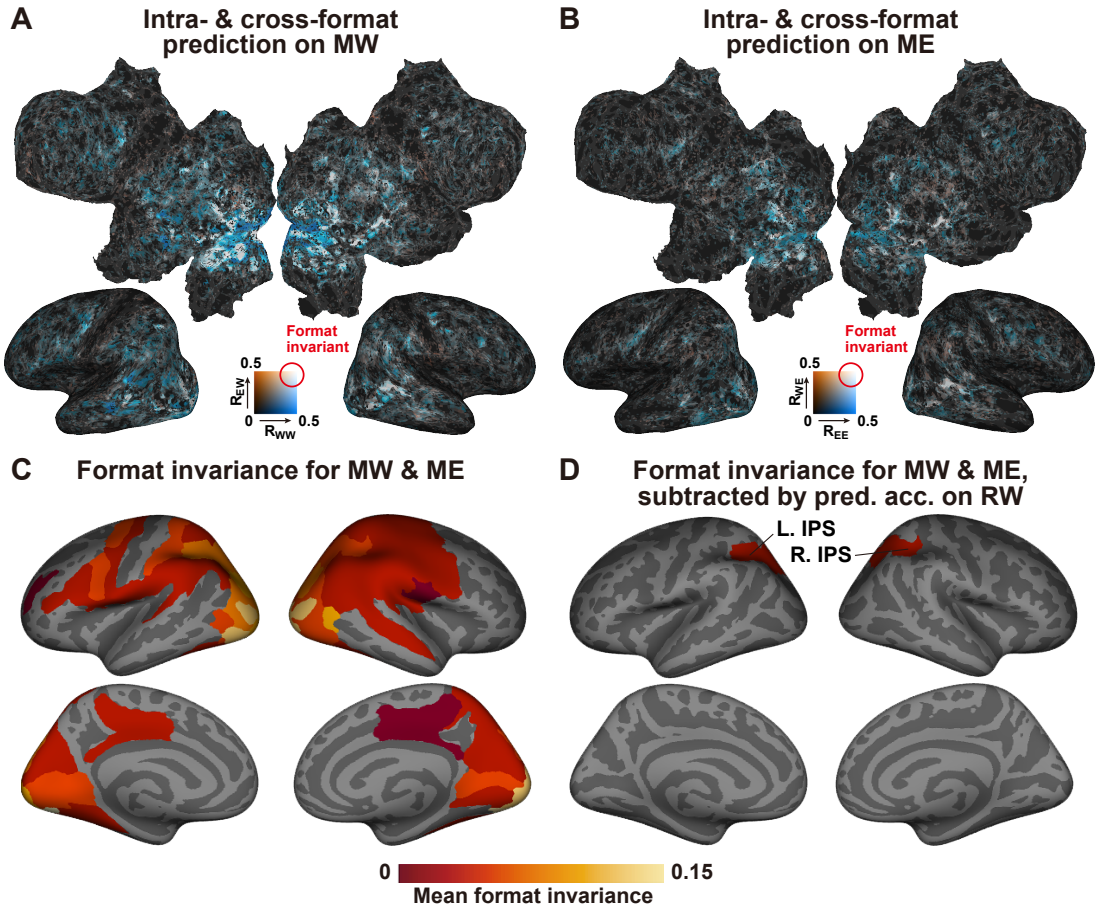

Figure 2. Intra- and cross-format predictions. (**A**) Comparison of the prediction accuracies when applied to the MW test dataset using operator features. The intra-format prediction accuracies of the MW model (denoted as R_{WW}, blue) and the cross-format ME model (denoted as R_{EW}, orange) are mapped onto the cortical surface of subject ID01. Regions in which activity was predicted, regardless of the stimulus modality, are shown in white. Only significant voxels ($P < 0.05$, FDR corrected) are shown. (**B**) Comparison between the prediction accuracies of the ME test dataset. Intra-format prediction accuracy, using the ME model (denoted as R_{EE}, blue) and cross-format prediction accuracy, using the MW model (denoted as R_{WE}, orange) are mapped onto the cortical surface. (**C**) Group average of format invariance (FI) mapped onto the template brain. FI values were averaged in each anatomical ROI and averaged across eight subjects. Only ROIs with >20% significant voxels are shown. (**D**) Group average of FI with subtraction of the prediction accuracy in the RW condition. ME, math expression condition; MW, math word condition; RW, read word condition; IPS, intraparietal sulcus; L, left hemisphere; R, right hemisphere.

Similarities in brain representations related to different mathematical operations

To investigate representational differences between different mathematical operations, we performed RSA based on the encoding model weights. Based on our previous observations, we focused on the anatomical ROIs of the bilateral IPS. In each anatomical ROI, we calculated a representational similarity matrix (RSM) for both MW and ME problems using Pearson's correlation distance (**Figure 3A**), and we thereby compared the resultant RSMs between the two formats. We found a positive correlation between the representational similarities of MW and ME problems in the bilateral IPS (Spearman's correlation coefficients: left IPS, $\rho = 0.446 \pm 0.097$; right IPS, $\rho = 0.462 \pm 0.169$; **Figure 3B, 3C**). This positive correlation existed even after excluding voxels predicted under the control RW condition (left IPS, $\rho = 0.324 \pm 0.056$; right IPS, $\rho = 0.430 \pm 0.170$; data of two subjects were not calculable because <10 voxels remained after voxel exclusion; **Figure 3D**), indicating that the similarity of the representational relationships among different operators provides the basis of format invariance.

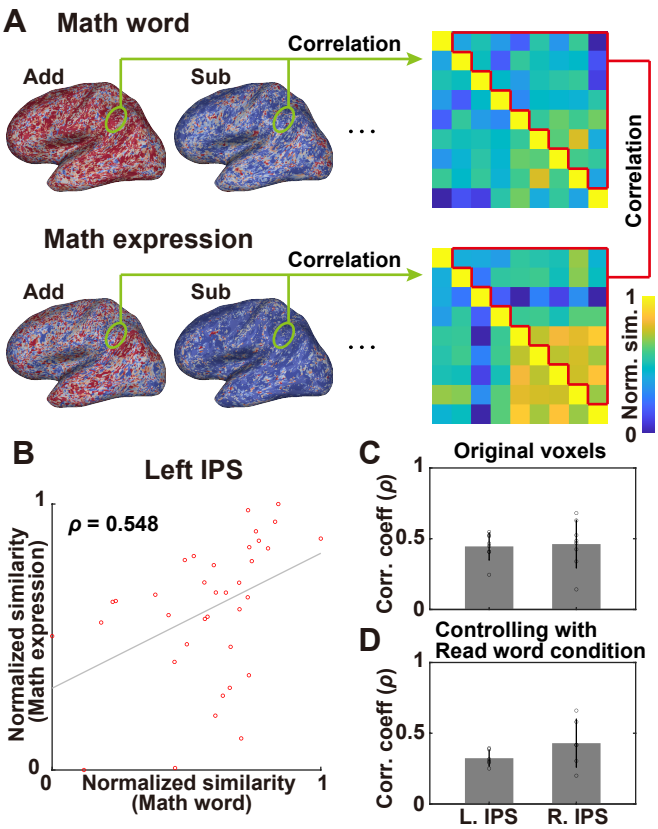

Figure 3. Representational similarity analysis. (**A**) A representational similarity matrix (RSM) was calculated using correlation distances between weight vectors extracted from the target region of interest (ROI). Spearman's correlation analysis including the upper triangular

parts of the RSM in the math word (MW) and math expression (ME) conditions was conducted for each target ROI. (**B**) Scatter plot of the normalized similarities in the MW and ME conditions in the left intraparietal sulcus (IPS) for subject ID01. (**C, D**) Spearman's correlation coefficients of all subjects are shown for the bilateral IPS using all voxels (**C**) and after excluding voxels predicted in the read word condition (**D**). Data for each subject are indicated by dots, and error bars represent standard deviation.

Visualization of brain representations related to different mathematical operations

Both FI analysis and RSA revealed shared brain representations for MW and ME problems. To visualize the related organizations, we applied PCA to the weight matrices of encoding models. PCA was also based on the bilateral IPS voxels. We found that nine tested problems had similar organizations in the MW and ME conditions (**Figure 4A**, **B**). The representational map visualized for the first and second PCs (PC1 and PC2) showed that double-operator problems were located in the bottom-right area, whereas single-operator problems were located in the upper-left area. Such organization was common across the MW and ME formats, and the mathematical problems were similarly organized even when the voxels predicted under the control RW condition were excluded (**Figure S4**). To evaluate the similarity between the representational space of MW and ME problems in a quantitative manner, we calculated Spearman's correlation coefficients between the PCA loading vectors of two formats. For the top two PCs, the PCA loading vectors were positively correlated between MW and ME problems ($\rho > 0.7$; PC1 and PC2 were significant using a permutation test, $P < 0.05$; **Figure 4C**). This effect was also observed after controlling for the RW condition ($\rho > 0.7$ for PC1–3; PC1 and PC2 were significant using a permutation test, $P < 0.05$; **Figure 4D**). Thus, the nine tested mathematical problems were organized in similar brain representational spaces regardless of their format.

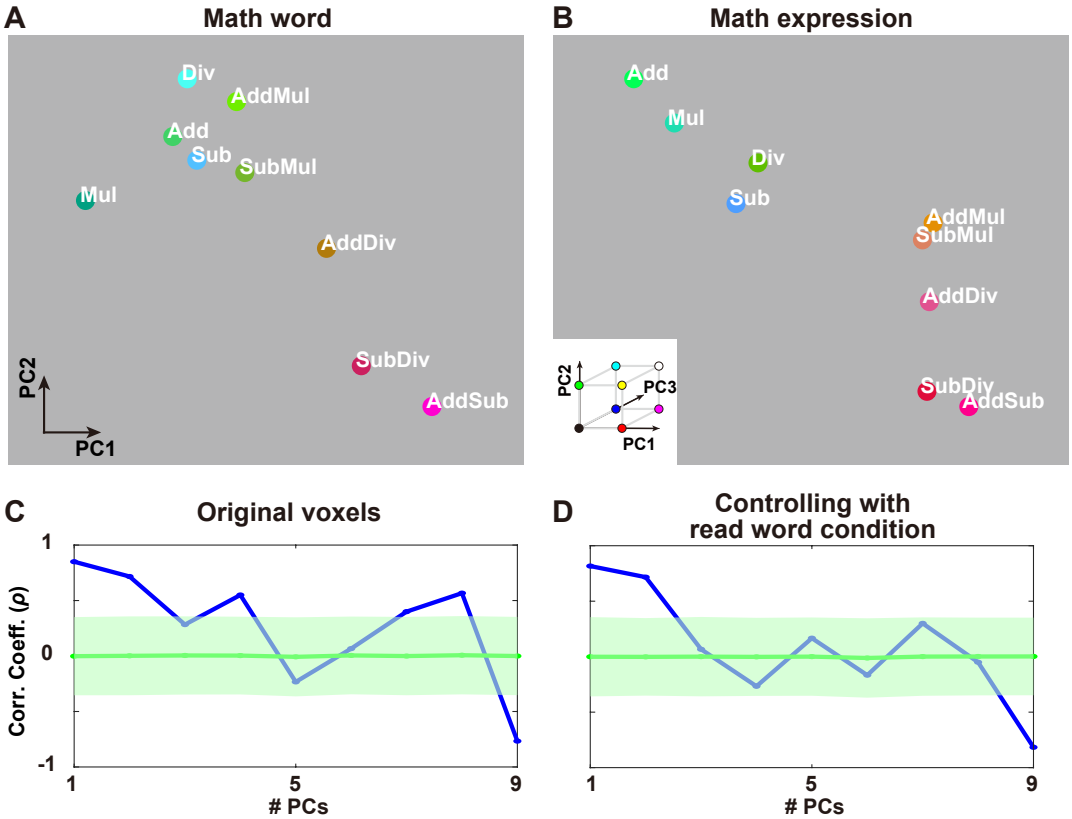

Figure 4. Principal component analysis. (**A, B**) Visualization of the representational space of mathematical operations in the MW (**A**) and ME (**B**) conditions. Colors indicate the loadings of the top three principal components (PC1, red; PC2, green; PC3, blue) of the operator feature weights (concatenated across subjects), which are mapped onto the two-dimensional cognitive space based on the loadings of PC1 and PC2. All operation types are shown in white. (**C, D**) Spearman's correlation coefficients from analyses of the PC loadings in the MW and ME conditions (blue line) plotted for the top nine PCs using all voxels (**C**) and after excluding voxels predicted in the read word condition (**D**). Mean correction coefficients (cyan line) and standard deviations (shaded areas) were obtained using 5,000 permutations of the operator labels of weight matrices.

Discussion

To the best of our knowledge, all previous neuroimaging studies on mathematical problem solving have used univariate analysis to assess the format-independency of mathematical operations (4–7). Thus, it had not been made clear whether the activation overlap of mathematical problems with different formats reflected shared cortical representations. In contrast to univariate analysis, multivariate analysis (i.e., RSA) can test whether activation

patterns are similar across different formats. Notably, training and test datasets had not previously been analyzed separately nor had the generalizability of quantitative models to new data been evaluated. In the current study, we demonstrated for the first time that multivariate analyses and encoding model approaches are useful for studying the format-invariance of mathematical problems.

The bilateral IPS is involved in numerical cognition (21–24) and in processing more complex mathematical problems (25, 26). Although format-invariance of quantity information has been questioned by several researchers (27–30), we provide evidence that complex mathematical problems are processed in this region in a format-invariant manner. Without controlling for the RW condition, we also found FI in the bilateral inferior frontal gyrus (IFG) (**Figure 2C**). The involvement of the left IFG in mathematical problems has been reported in previous studies (25, 26, 31), and the functional and anatomical connectivity between the IFG and IPS may play an essential role in the development of mathematical ability (5, 32–34). Thus, our findings on the bilateral IFG and IPS are consistent with the results of other mathematical cognition studies.

Our PCA visualization showed that single- and double-operator problems were distinctly clustered, which is consistent with the result in a previous study comparing single- and double-operator problems (35). Researchers have argued that activations in the bilateral IPS during mathematical problem solving are modulated by problem difficulty as measured using reaction times (RTs) (5). Although our PCA results seemed to reflect a difference between single- and double-operator problems (**Figure S5**), the representational space we obtained after controlling for the RW condition also had a similar organization between single- and double-operator problems (**Figure S4**). In particular, certain double-operator problems were closely located based on the shared single-operator, e.g., the AddSub and SubDiv conditions and the AddMul and SubMul conditions were paired, respectively. Our results for operator-specific representations are consistent with previous reports of double dissociation between different operators (36) and operator-selective neurons (37), which would contribute to the format-invariant predictability of cortical activations.

Although the current study paves the way toward quantitative modeling of complex mathematical thoughts, we acknowledge that it has several limitations. First, because we chose experimental stimuli from existing datasets (38, 39), we could not perfectly balance the number of letters and digit sizes. Although we excluded such effects by subtracting baseline data determined using the RW task with the same visual stimuli (**Figure 2D**) and by explicitly regressing out features of noninterest (**Figure S3**), the between-condition difference may still

have had some influence on the representational organization of the tested math problems. Second, although we validated our results for each subject (**Figure 3C, D; Figure S2**) and the sample size was consistent with that used in previous studies (9, 10, 15, 18, 20), the relatively small number of subjects tested may limit the generalizability of our results to the broader population and other age groups. Third, the categorical model of mathematical problems (i.e., operator features) may not fully capture the complex processes involved in math problem solving, especially the quantity information used in the problems. Therefore, further studies are warranted to fully clarify the individual differences and detailed subprocesses involved in mathematical thought. Nevertheless, based on the current work, we conclude that the bilateral IPS network subserves format-invariant representations of mathematical problems.

Methods

Subjects

Eight healthy college students (aged 20–23 years, three females, all with normal vision), denoted as ID01–ID08, participated in this study. The number of subjects was determined based on previous encoding model studies (9, 10, 15, 18, 20). The small number of subjects is compensated for by a large number of samples for each subject (i.e., 3 h of tests). Subjects were all right-handed (laterality quotient = 80–100) as assessed using the Edinburgh inventory (40). Written informed consent was obtained from all subjects prior to their participation in the study, and the experiment was approved by the ethics and safety committee of the National Institute of Information and Communications Technology in Osaka, Japan.

Stimuli and testing procedure

Subjects performed arithmetic problems in two different formats, namely MW and ME problems (see **Figure S5** for the behavioral results). We selected MW problems with a single operation of addition (Add), subtraction (Sub), multiplication (Mul), and division (Div) from the IL dataset (38), and MW problems with two operations, including addition and subtraction (AddSub), addition and multiplication (AddMul), addition and division (AddDiv), subtraction and multiplication (SubMul), and subtraction and division (SubDiv), from the CC dataset (39). Each condition consisted of 35 instances. Both of the abovementioned datasets have been widely used in previous machine learning studies on MW problems (41). For our MW problems, the original English sentences were translated into Japanese by the first author (T.N.). Corresponding arithmetic expression problems were created based on the MW problems; thus, these problems gave the same results.

In the MW condition, an arithmetic word problem, (e.g., “There are 3 eggs in each box. How many eggs are in 2 boxes?”) was presented for 6 or 10 s [6 s for the single-operator conditions (Add, Sub, Mul, and Div) and 10 s for the double-operator conditions (AddSub, AddMul, AddDiv, SubMul, and SubDiv)]. Subjects had access to a two-button response pad in their left hand, and they were instructed to perform a calculation based on the presented problem and press the left button when they had calculated an answer. After 1–2 s during which a fixation cross stimulus was displayed, a probe digit stimulus was presented for 2 s (e.g., “6”). Subjects were asked to press the left button if the presented digit matched their answer and the right button if it did not match their answer. The next trial started after 1–2 s, during which time a fixation cross stimulus was shown. The number of letters used in each condition was as follows: 61.3 ± 13.3 (Add), 57.0 ± 6.6 (Sub), 51.4 ± 11.2 (Mul), 66.9 ± 12.0 (Div), 89.4 ± 13.4 (AddSub), 85.6 ± 13.1 (AddMul), 91.2 ± 16.3 (AddDiv), 83.8 ± 7.2 (SubMul), and 91.4 ± 13.6 (SubDiv).

In the ME condition, an arithmetic expression problem (e.g., “3 × 2 = ?”) was presented for 4 or 6 s [4 s for the single-operator conditions (Add, Sub, Mul, and Div), and 6 s for the double-operator conditions (AddSub, AddMul, AddDiv, SubMul, and SubDiv)]. As in the MW condition, the subjects were instructed to perform a calculation based on the presented problem and to press the left button when they had calculated an answer. The fixation cross stimulus was then presented for 1–2 s, after which a probe digit stimulus was presented for 2 s (e.g., “6”). Subjects were again asked to press the left or right button if the presented digit matched or did not match their answer, respectively. The next trial began after the fixation cross stimuli was again displayed for 1–2 s. The number of letters used in each condition was as follows: 5.5 ± 0.5 (Add), 6.5 ± 0.5 (Sub), 5.0 ± 0.2 (Mul), 5.9 ± 0.4 (Div), 9.3 ± 0.9 (AddSub), 8.9 ± 0.3 (AddMul), 10.1 ± 0.7 (AddDiv), 9.6 ± 0.5 (SubMul), and 10.9 ± 0.5 (SubDiv).

In the RW condition, an arithmetic word problem was presented for 6 or 10 s as described above for the MW condition. The set of sentences presented was the same as that used in the MW condition (e.g., “There are 3 eggs in each box. How many eggs are in 2 boxes?”). However, subjects were instructed to carefully read the sentence and press the left button when they had memorized all non-numerical nouns. After a fixation cross stimulus was shown for 1–2 s, two words were presented for 2 s (e.g., “eggs” and “boxes”). Subjects were asked to press the left button if both nouns were included in the original sentence and the right button if one of the two nouns were not included in the sentence. Two nouns were presented in the answer-matching phase to ensure that the subjects memorized all the nouns that appeared in the sentence. The next trial started after a fixation cross stimuli was displayed for 1–2 s.

Stimuli were presented on a projector screen inside the scanner (21.0 × 15.8° visual angle at 30 Hz). During scanning, subjects wore MR-compatible ear tips. The experiment was performed for three days, with 5–7 runs performed each day. Presentation software (Neurobehavioral Systems, Albany, CA, USA) was used to control the stimulus presentation and collection of behavioral data. To measure button responses, optic response pads with two buttons were used (HHSC-2x2, Current Designs, Philadelphia, PA, USA).

In total, 16 runs were conducted, of which 6 training and 2 test runs were in the MW condition and another 6 training and 2 test runs were in the ME condition. The data from each pair of test runs were averaged to increase the signal-to-noise ratio. Each run contained 90 trials. The duration of a single run was 595 s for the MW condition and 455 s for the ME condition. At the beginning of each run, 10 s of dummy scans were acquired, during which the fixation cross was displayed, and these dummy scans were later omitted from the final analysis to reduce noise. We also obtained 10 s of scans at the end of each run, during which the fixation cross was displayed; however, these scans were included in the analyses.

MRI data acquisition

The experiment was conducted using a 3.0 T scanner (MAGNETOM Prisma; Siemens, Erlangen, Germany) with a 64-channel head coil. We scanned 72 interleaved axial slices that were 2-mm thick without a gap, parallel to the anterior and posterior commissure line, using a T2*-weighted gradient echo multiband echo-planar imaging sequence [repetition time (TR) = 1,000 ms; echo time (TE) = 30 ms; flip angle (FA) = 62°; field of view (FOV) = 192 × 192 mm^2; resolution = 2 × 2 mm^2; multiband factor = 6]. We obtained 605 volumes for the MW and RW conditions and 465 volumes for the ME condition, with each set following 10 dummy images. For anatomical reference, high-resolution T1-weighted images of the whole brain were also acquired from all subjects with a magnetization-prepared rapid acquisition gradient echo sequence (TR = 2,530 ms; TE = 3.26 ms; FA = 9°; FOV = 256 × 256 mm^2; voxel size = 1 × 1 × 1 mm^3).

fMRI data preprocessing

Motion correction in each run was performed using the statistical parametric mapping toolbox (SPM12; Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm/). All volumes were aligned to the first EPI image for each subject. Low-frequency drift was removed using a median filter with a 120-s window. Slice timing correction was performed against the first slice of each scan. The response for each voxel

was then normalized by subtracting the mean response and scaling it to the unit variance. We used FreeSurfer (https://surfer.nmr.mgh.harvard.edu/) to identify the cortical surfaces from the anatomical data and to register these to the voxels of the functional data. For each subject, the voxels identified in the cerebral cortex (59,499–75,980 voxels per subject) were used in the analysis.

Operator features

The operator features were composed of one-hot vectors, which were assigned values of 1 or 0 for each time bin during the stimulus presentation of arithmetic problems, indicating whether one of the nine tested conditions (Add, Sub, Mul, Div, AddSub, AddMul, AddDiv, SubMul, and SubDiv) was performed in that period; thus, nine operator features were used.

Encoding model fitting

In the encoding model, the cortical activity in each voxel was fitted with a finite impulse response model that captured the slow hemodynamic response and its coupling with neural activity (9, 42). The feature matrix $\mathbf{F_E}$ [T × 6N] was modeled by concatenating sets of [T × N] feature matrices with six temporal delays of 2–7 s (T = number of samples; N = number of features). The cortical response $\mathbf{R_E}$ [T × V] was then modeled by multiplying the feature matrix $\mathbf{F_E}$ by the weight matrix $\mathbf{W_E}$ [6N × V] (V = number of voxels):

$$\widehat{\mathbf{R}}_{\mathbf{E}} = \mathbf{F}_{\mathbf{E}}\mathbf{W}_{\mathbf{E}}.$$

We used an L2-regularized linear regression with the training dataset to obtain the weight matrix $\mathbf{W_E}$. The training dataset consisted of 3,630 (3,630 s) and 2,790 (2,790 s) samples for the MW and ME conditions, respectively, whereas a training dataset was not provided for the RW condition. The optimal regularization parameter was assessed using 10-fold cross-validation where the 11 different regularization parameters ranged from 1 to 2^{10}.

The test dataset consisted of 605 samples (605 s) for the MW and RW conditions and 465 samples (465 s) for the ME condition (each repeated twice). Two repetitions of the test dataset were averaged to increase the signal-to-noise ratio. Statistical significance (one-sided) was computed by comparing estimated correlations to the null distribution of correlations between two independent Gaussian random vectors with the same length as the test dataset. The statistical threshold was set at $P < 0.05$ and corrected for multiple comparisons using the FDR procedure (43). For data visualization on the cortical maps, we used pycortex (44) and fsbrain (45).

Quantification of format invariance

To quantify how the unimodal models explained brain activity in each voxel regardless of the presentation modality, we defined FI values. In our previous study, we used a similar measurement to quantify how linguistic information is processed in a modality-invariant manner (16). To quantify FI based on prediction accuracy, we used the geometric mean of prediction accuracy rather than the weight correlation. This was justified because models with similar weight values have similar predictive performance. FI consisted of two components: D_W and D_E. D_W was defined as the degree of predictability for the MW test dataset regardless of the training modality:

$$D_W = \sqrt{R_{EW} \cdot R_{WW}},$$

where R_{WW} and R_{EW} are the intramodal prediction accuracy for the MW model and the cross-modal prediction accuracy for the ME model when applied to the test dataset for the MW condition, respectively. Similarly, D_E was defined as the degree of predictability calculated for the ME test dataset regardless of the training format:

$$D_E = \sqrt{R_{WE} \cdot R_{EE}},$$

where R_{EE} and R_{WE} are the intra-format prediction accuracy using an encoding model trained with the ME condition and the cross-format prediction accuracy trained with the MW condition when applied to the test dataset for the ME condition, respectively. For all voxels with negative prediction accuracies, the prediction accuracy was reset to 0 to avoid obtaining imaginary values. FI was then calculated for each voxel as a geometric mean between D_W and D_E as follows:

$$FI = \sqrt{D_W \cdot D_E}.$$

FI values ranged from 0 to 1; high FI indicates that the target features are represented in a format-invariant manner, whereas an FI of zero indicates that the target voxel does not have a shared representation for MW and ME problems. To calculate the null distribution of FI values, a set of four correlation coefficients produced using Gaussian random vector pairs were used instead of R_{WW}, R_{EE}, R_{EW}, and R_{WE}. The statistical threshold was set at $P < 0.05$ and corrected for multiple comparisons using the FDR procedure (43).

RSA

For each of the target anatomical ROIs, *operator similarity* was calculated using the Pearson correlation distance between weight vectors of different operators extracted from the weight matrix of operator feature encoding models. We selected the voxels that showed significant FI values and averaged six time delays for each operator. An RSM was calculated based on all

combinations of nine operator types. The upper triangular part of the RSM was rearranged into a single vector for both the MW and ME conditions, and Spearman's correlation coefficients for correlations between the two rearranged vectors were calculated.

PCA

For each subject, we performed PCA on the weight matrix of the operator feature model concatenated across the eight subjects. We selected the voxels that showed significant FI values and averaged six time delays for each operator. We applied PCA to the [9*V] weight matrices (V = number of voxels) of the MW and ME encoding models. To show the structure of the representational space of mathematical operations, nine operators were mapped onto the two-dimensional space using the loadings of the first and second PCs, i.e., PC1 and PC2, as the x-axis and y-axis, respectively. In this space, the operators were colored red, green, and blue based on the relative PCA loadings in PC1, PC2, and PC3, respectively. To investigate the relationship between the components obtained using the MW and ME models, we calculated Spearman's correlation coefficients for correlations between the PCA scores for each component of two models (e.g., between PC1 of the MW model and PC1 of the ME model). To obtain a null distribution of such correlation coefficients, we randomly permutated the categorical labels of the MW and ME models 5,000 times before applying PCA. *P* values of the actual correlation coefficients were evaluated according to the number of random correlation coefficients larger than the actual value across all permutations.

Data and code availability

The source data and analysis code used in the current study are available upon reasonable request.

Reference

1. C.-T. Ng, T.-C. Lung, T.-T. Chang, Operation-Specific Lexical Consistency Effect in Fronto-Insular-Parietal Network During Word Problem Solving. *Front. Hum. Neurosci.* **15**, 631438 (2021).

2. L. S. Fuchs, *et al.*, Contributions of domain-general cognitive resources and different forms of arithmetic development to pre-algebraic knowledge. *Dev. Psychol.* **48**, 1315–1326 (2012).

3. S. R. Powell, L. S. Fuchs, Does Early Algebraic Reasoning Differ as a Function of Students' Difficulty with Calculations versus Word Problems? *Learn. Disabil. Res. Pract.* **29**, 106–116 (2014).

4. M.-H. Sohn, *et al.*, Behavioral equivalence, but not neural equivalence--neural evidence of alternative strategies in mathematical thinking. *Nat. Neurosci.* **7**, 1193–1194 (2004).

5. T.-T. Chang, T.-C. Lung, C.-T. Ng, A. W. S. Metcalfe, Fronto-insular-parietal network engagement underlying arithmetic word problem solving. *Hum. Brain Mapp.* **40**, 1927–1941 (2019).

6. X. Zhou, *et al.*, The semantic system is involved in mathematical problem solving. *Neuroimage* **166**, 360–370 (2018).

7. S. D. Newman, G. Willoughby, B. Pruce, The effect of problem structure on problem-solving: an fMRI study of word versus number problems. *Brain Res.* **1410**, 77–88 (2011).

8. T. Naselaris, K. N. Kay, S. Nishimoto, J. L. Gallant, Encoding and decoding in fMRI. *Neuroimage* **56**, 400–410 (2011).

9. S. Nishimoto, *et al.*, Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21**, 1641–1646 (2011).

10. K. N. Kay, T. Naselaris, R. J. Prenger, J. L. Gallant, Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).

11. T. Çukur, S. Nishimoto, A. G. Huth, J. L. Gallant, Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.* **16**, 763–770 (2013).

12. A. G. Huth, S. Nishimoto, A. T. Vu, J. L. Gallant, A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).

13. T. Nakai, N. Koide-Majima, S. Nishimoto, Correspondence of categorical and feature-based representations of music in the human brain. *Brain Behav.* **11**, e01936 (2021).

14. V. De Angelis, *et al.*, Cortical processing of pitch: Model-based encoding and decoding of auditory fMRI responses to real-life sounds. *Neuroimage* **180**, 291–300 (2018).

15. A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, J. L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).

16. T. Nakai, H. Q. Yamaguchi, S. Nishimoto, Convergence of modality invariance and attention selectivity in the cortical semantic circuit. *Cereb. Cortex* (2021) https:/doi.org/10.1093/cercor/bhab125.

17. S. Nishida, A. Blanc, N. Maeda, M. Kado, S. Nishimoto, Behavioral correlates of cortical semantic representations modeled by word vectors. *PLoS Comput. Biol.* **17**, e1009138 (2021).

18. T. Horikawa, A. S. Cowen, D. Keltner, Y. Kamitani, The neural representation of visually evoked emotion Is high-dimensional, categorical, and distributed across transmodal brain regions. *iScience* **23**, 101060 (2020).

19. N. Koide-Majima, T. Nakai, S. Nishimoto, Distinct dimensions of emotion in the human brain and their representation on the cortical surface. *Neuroimage* **222**, 117258 (2020).

20. T. Nakai, S. Nishimoto, Quantitative models reveal the organization of diverse cognitive functions in the brain. *Nat. Commun.* **11**, 1142 (2020).

21. M. Piazza, P. Pinel, D. Le Bihan, S. Dehaene, A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron* **53**, 293–305 (2007).

22. J. F. Cantlon, E. M. Brannon, E. J. Carter, K. A. Pelphrey, Functional imaging of numerical processing in adults and 4-y-old children. *PLoS Biol.* **4**, e125 (2006).

23. E. Eger, P. Sterzer, M. O. Russ, A.-L. Giraud, A. Kleinschmidt, A supramodal number representation in human intraparietal cortex. *Neuron* **37**, 719–725 (2003).

24. R. Cohen Kadosh, *et al.*, Are numbers special? The comparison systems of the human brain investigated by fMRI. *Neuropsychologia* **43**, 1238–1248 (2005).

25. T. Nakai, K. Okanoya, Cortical collateralization induced by language and arithmetic in non-right-handers. *Cortex* **124**, 154–166 (2020).

26. T. Nakai, K. L. Sakai, Neural mechanisms underlying the computation of hierarchical tree structures in mathematics. *PLoS One* **9**, e111439 (2014).

27. I. M. Lyons, D. Ansari, S. L. Beilock, Qualitatively different coding of symbolic and nonsymbolic numbers in the human brain. *Hum. Brain Mapp.* **36**, 475–488 (2015).

28. I. M. Lyons, S. L. Beilock, Characterizing the neural coding of symbolic quantities. *Neuroimage* **178**, 503–518 (2018).

29. J. Bulthé, B. De Smedt, H. P. Op de Beeck, Format-dependent representations of symbolic and non-symbolic numbers in the human cortex as revealed by multi-voxel pattern analyses. *Neuroimage* **87**, 311–322 (2014).

30. J. Bulthé, B. De Smedt, H. P. Op de Beeck, Visual number beats abstract numerical magnitude: format-dependent representation of Arabic digits and dot patterns in human parietal cortex. *J. Cogn. Neurosci.* **27**, 1376–1387 (2015).

31. T. Nakai, K. Okanoya, Neural Evidence of Cross-domain Structural Interaction between Language and Arithmetic. *Sci. Rep.* **8**, 12873 (2018).

32. J. M. Tsang, R. F. Dougherty, G. K. Deutsch, B. A. Wandell, M. Ben-Shachar, Frontoparietal white matter diffusion properties predict mental arithmetic skills in children. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 22546–22551 (2009).

33. R. W. Emerson, J. F. Cantlon, Early math achievement and functional connectivity in the fronto-parietal network. *Dev. Cogn. Neurosci.* **2 Suppl 1**, S139-51 (2012).

34. M. Rosenberg‐Lee, S. Ashkenazi, T. Chen, Brain hyper‐connectivity and operation‐specific deficits during arithmetic problem solving in children with developmental dyscalculia. *Developmental* (2015).

35. V. Prabhakaran, B. Rypma, J. D. Gabrieli, Neural substrates of mathematical reasoning: a functional magnetic resonance imaging study of neocortical activation during performance of the necessary arithmetic operations test. *Neuropsychology* **15**, 115–127 (2001).

36. S. Dehaene, L. Cohen, Cerebral pathways for calculation: double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex* **33**, 219–250 (1997).

37. E. F. Kutter, J. Boström, C. E. Elger, A. Nieder, F. Mormann, Neuronal codes for arithmetic rule processing in the human brain. *Curr. Biol.* (2022) https:/doi.org/10.1016/j.cub.2022.01.054.

38. S. Roy, D. Roth, Solving General Arithmetic Word Problems in *EMNLP*, (2015), pp. 1743–1752,.

39. S. Roy, T. Vieira, D. Roth, Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics* **3**, 1–13 (2015).

40. R. C. Oldfield, The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).

41. S. Mandal, S. K. Naskar, Solving Arithmetic Mathematical Word Problems: A Review and Recent Advancements in *Information Technology and Applied Mathematics*, (Springer Singapore, 2019), pp. 95–114.

42. K. N. Kay, S. V. David, R. J. Prenger, K. A. Hansen, J. L. Gallant, Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. *Hum. Brain Mapp.* **29**, 142–156 (2008).

43. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).

44. J. S. Gao, A. G. Huth, M. D. Lescroart, J. L. Gallant, Pycortex: an interactive surface visualizer for fMRI. *Front. Neuroinform.* **9**, 23 (2015).

45. T. Schäfer, C. Ecker, fsbrain: an R package for the visualization of structural neuroimaging data. *bioRxiv*, 2020.09.18.302935 (2020).

Acknowledgments

We thank MEXT/JSPS KAKENHI [grant numbers JP17K13083 and JP18H05091 in #4903 (Evolinguistics) for T.N., and JP15H05311 and JP18H05522 for S.N.] as well as JST CREST (JPMJCR18A5) and ERATO (JPMJER1801) (for S.N.) for their partial financial support of this study. The funders played no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Author contributions

T.N. and S.N. designed the study; T.N. collected and analyzed the data; T.N. and S.N. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Quantitative models demonstrate of format-independent representations of mathematical operations in the brain

Tomoya Nakai, Shinji Nishimoto

Supplementary Information

Encoding model fitting excluding regressors of noninterest

To evaluate the possible effect of sensorimotor features and general cognitive load on the model predictions, we performed an additional encoding model fitting by excluding the features of noninterest. To this end, we concatenated motion energy (MoE; visual), button response (BR; motor), reaction time (RT) (cognitive load), and letter (orthographic) features (the following subsections provide details on each feature). The concatenated features were used as a feature matrix for the encoding modeling. The original training dataset was divided into five subtraining runs and a single subtest run. Encoding models were trained using the subtraining dataset, and prediction accuracy was calculated using the remaining subtest dataset. L2-regularized linear regressions were applied in the same manner as that described in the "Encoding model fitting" subsection in the Methods. This procedure was repeated for all runs in the original training dataset. Cortical voxels significantly predicted in at least five out of six repetitions were regarded as nonmath voxels and excluded from the analysis of target encoding models.

MoE features

We employed an MoE model used in previous studies (Nishimoto *et al.* 2011; Nakai and Nishimoto 2020; Koide-Majima *et al.* 2020) that can be found in a public repository (https://github.com/gallantlab/motion_energy_matlab). First, movie frames and pictures were spatially downsampled to 96 × 96 pixels. The RGB pixel values were then converted into the Commission International de l'Eclairage (CIE) LAB color space, and the color information was discarded. The luminance (L*) pattern was passed through a bank of three-dimensional spatiotemporal Gabor wavelet filters; the outputs of the two filters with orthogonal phases (quadrature pairs) were squared and summed to yield local MoE. Subsequently, MoE was compressed with a log-transformation and temporally downsampled to 0.5 Hz. Filters were tuned to six spatial frequencies (0, 1.5, 3.0, 6.0, 12.0, and 24.0 cycles per image) and three temporal frequencies (0, 4.0, and 8.0 Hz) without directional parameters. Filters were positioned on a square grid that covered the screen. The adjacent filters were separated by 3.5 standard

deviations of their spatial Gaussian envelopes. To reduce the computational load, the original MoE features, which had 1,395 dimensions, were reduced to 300 dimensions using principal component analysis.

BR feature

The BR feature was constructed based on the number of button responses per second. One BR feature was used.

Letter feature

The letter feature was constructed based on the number of letters that appeared in each stimulus. One letter feature was used.

RT feature

For each time bin during the presentation of mathematical problems, the RT of that trial was assigned as a cognitive load of the trial. One RT feature was used.

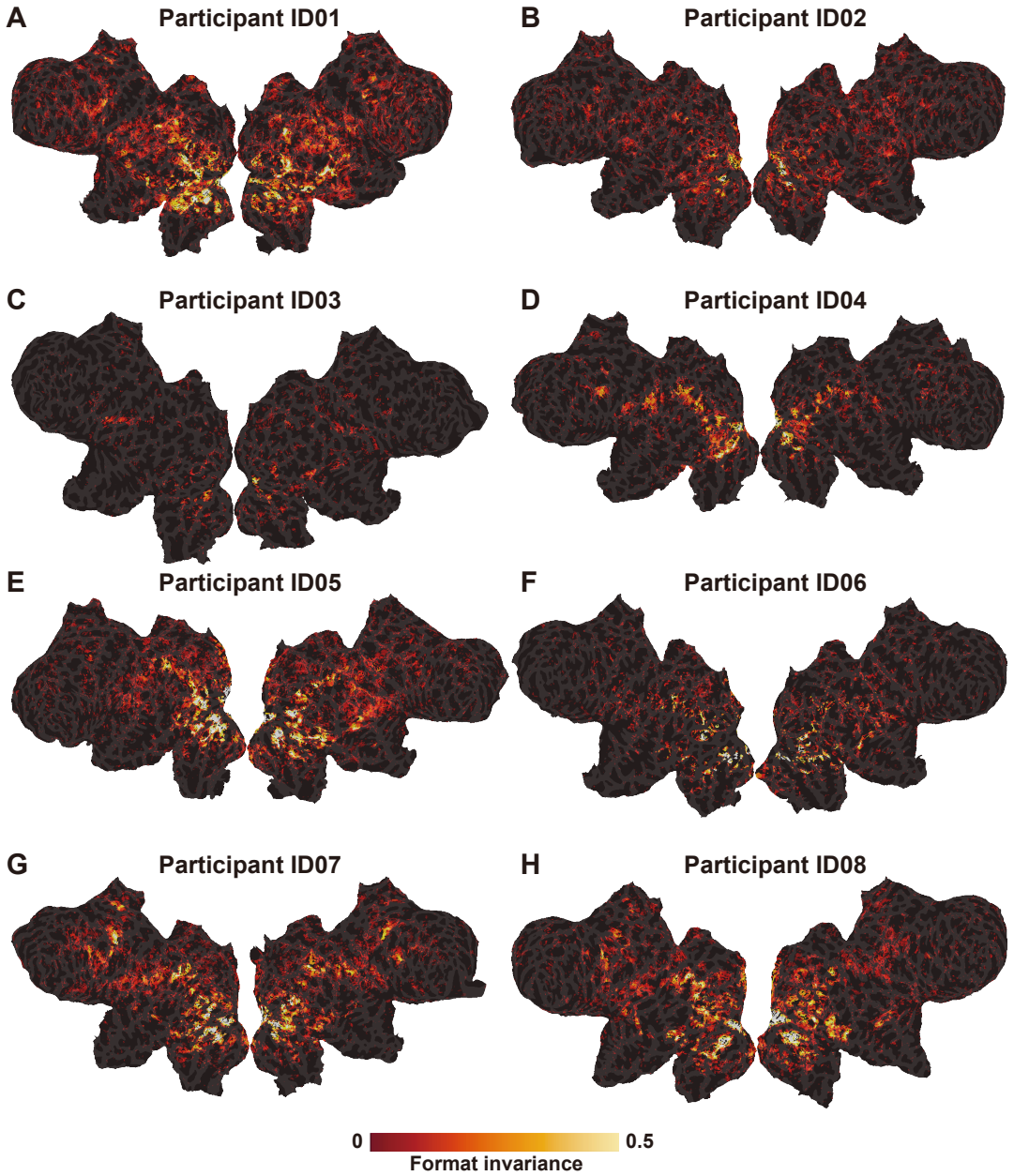

Figure S1. **Format invariance map for each subject.** Cortical maps of format invariance for subjects ID01–ID08 mapped onto cortical surfaces. Only significant voxels ($P < 0.05$, FDR corrected) are shown.

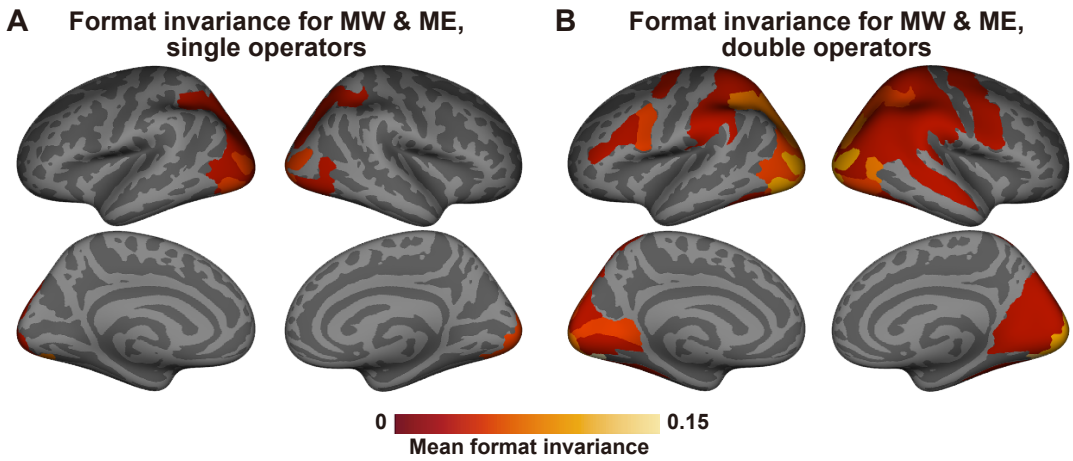

Figure S2. **Format invariance analysis of single- and double-operators.** Group averages of format invariance (FI) for (**A**) single- and (**B**) double-operator problems mapped onto the template brain. FI values were averaged in each anatomical region of interest (ROI) and averaged across eight subjects. Only ROIs with >20% significant voxels are shown. MW, math word; ME, math expression.

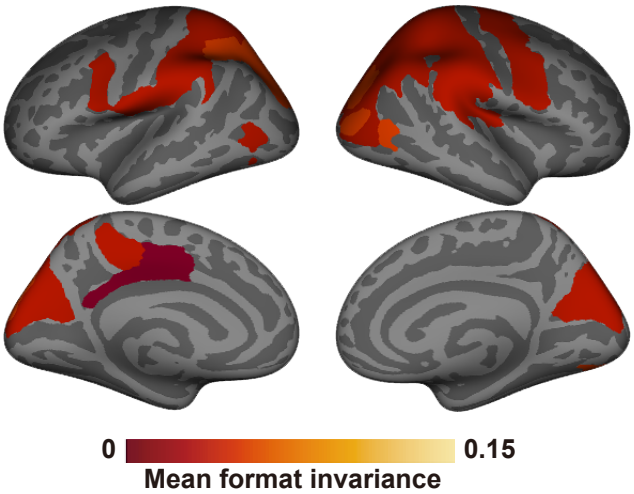

Figure S3. **Format invariance analysis after excluding the effects of regressors of noninterest.** Group averages of format invariance after excluding voxels significantly predicted by regressors of noninterest, either in the math word or math expression problems. The regressors of noninterest consisted of the visual, motor, orthographic, and cognitive load features.

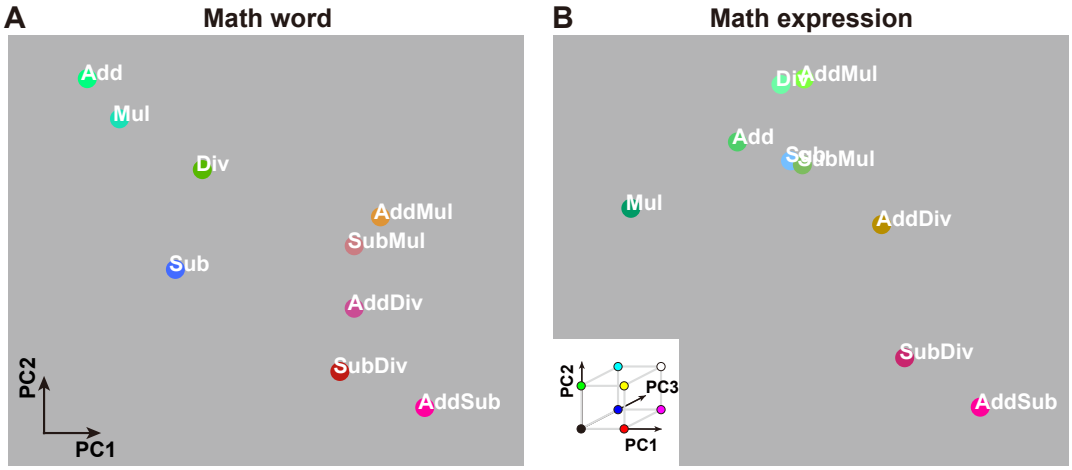

Figure S4. Principal component analysis including the read word condition control. (**A, B**) Visualization of the representational space of mathematical operations in (**A**) the math word and (**B**) the math expression conditions after excluding voxels predicted in the read word condition. Colors indicate the loadings of the top three principal components (PC1, red; PC2, green; PC3, blue) of the operator feature weights (concatenated across subjects), which were mapped onto the two-dimensional cognitive space based on the loadings of PC1 and PC2.

Behavioral results

The subjects were able to solve both MW and ME conditions with >90% accuracy (MW condition, 92.7% ± 4.0%; ME condition, 94.8% ± 2.5%; Cohen's $d = 0.62$; **Figure S5A** and **C**, respectively). The subjects solved the single-operator (Add, Sub, Mul, and Div) and double-operator (AddSub, AddMul, AddDiv, SubMul, and SubDiv) problems with equal accuracy both in the MW (single-operator, 93.0% ± 4.2%; double-operator, 92.5% ± 4.5%; $d = 0.11$) and ME (single-operator, 94.5% ± 3.1%; double-operator, 95.1% ± 2.5%; $d = 0.22$) conditions. These results indicate that subjects did not have specific difficulty in solving arithmetic problems in either format.

As expected, RT was higher in the MW condition than that in the ME condition (MW condition, 4.59 ± 0.84 s; ME condition, 1.56 ± 0.30 s; $d = 2.70$; **Figures S5B** and **S3D**, respectively). For the MW condition, RTs were significantly higher in the double-operator conditions than those in the single-operator conditions for all subjects (single-operator, 3.32 ± 0.63 s; double-operator, 5.60 ± 1.02 s; $d = 2.70$). For the ME condition, RTs were higher in the double-operator problems than those in the single-operator problems (single-operator, 1.06 ± 0.14 s; double-operator, 1.96 ± 0.45 s; $d = 2.69$).

For the control RW condition, there was no distinction among different operators, but there was a distinction between sentences used in the single-operator and double-operator problems. Overall accuracy in the RW condition was comparable to that in the MW condition (89.6% ± 4.5%; $d = 0.67$), whereas the accuracy was lower than that in the ME condition ($d = 1.37$) (**Figure S5E**, left). RT in the RW condition was comparable with that in the MW condition (4.16 ± 0.86 s; $d = 0.61$), but it was higher than that in the ME condition ($d = 4.02$) (**Figure S5E**, right). Additionally, RT was higher in the double-operator sentences (single-operator, 3.41 ± 0.62 s; double-operator, 4.76 ± 1.09 s; $d = 1.53$). In contrast, accuracy was higher in the single-operator sentences (single-operator, 92.5% ± 4.2%; double-operator, 87.3% ± 5.6%; $d = 1.06$). These results indicate that the RW and MW conditions had similar levels of difficulty for the subjects, but that the former was more demanding than the ME condition.

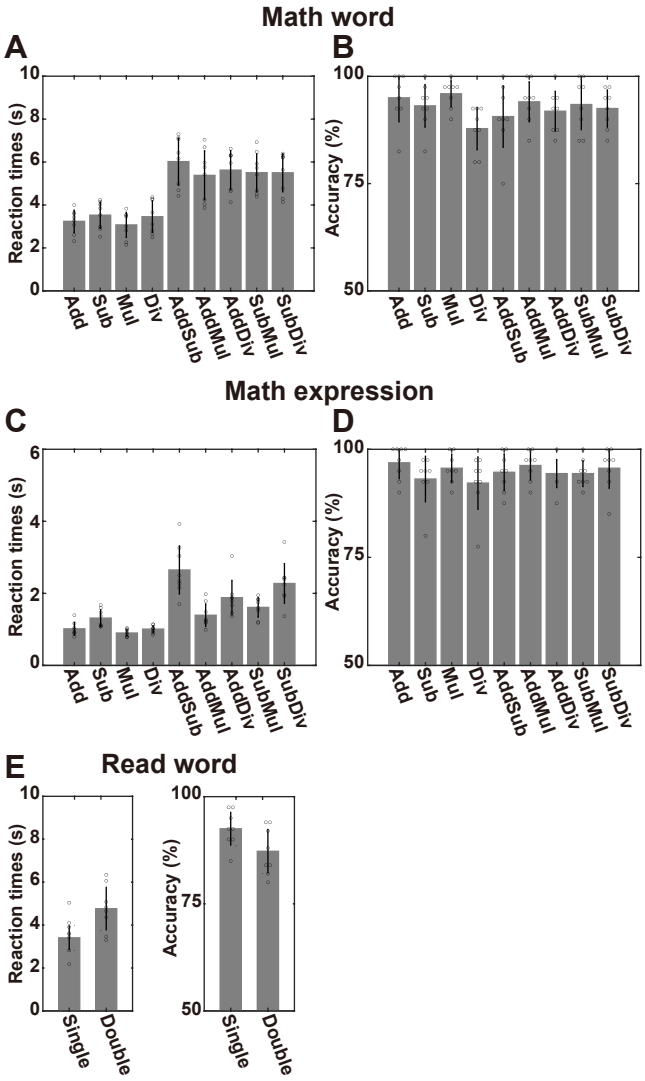

Figure S5. **Behavioral results.** Accuracy (**A** and **C**) and reaction time (**B** and **D**) of the math word and math expression conditions, respectively, are shown for nine operator types. (**E**) Accuracy and reaction time of the read word condition. Data for each subject are indicated by dots, and error bars represent standard deviation.